

# **Elucidating the events of transcription factor binding in human brain enhancers**



**By**

**Hizran Khatoun**

**National Centre for Bioinformatics**

**Faculty of Biological Sciences**

**Quaid-i-Azam University**

**Islamabad, Pakistan**

**2023**

# **Elucidating the events of transcription factor binding in human brain enhancers**



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

**Doctor of Philosophy in Bioinformatics**

By

**Hizran Khatoon**

**National Center for Bioinformatics**

**Faculty of Biological Sciences**

**Quaid-i-Azam, University**

**Islamabad, Pakistan**

**2023**

بِسْمِ اللَّهِ الرَّحْمَنِ الرَّحِيمِ

---

**“IN THE NAME OF ALLAH, THE MOST GRACIOUS,  
THE MOST MERCIFUL”**

---

### Author's Declaration

I, Hizran Khatoon, hereby state that my Ph.D. thesis titled as “**Elucidating the events of transcription factor binding in human brain enhancers**” is my own research efforts and has not been submitted previously by me for taking any degree from Quaid-i-Azam University Islamabad, Pakistan or anywhere in the country/world.

At any time if my statement is found to be incorrect, the University has the right to retract my Ph.D. degree.



Name: Hizran Khatoon

Date: \_\_\_\_\_

## Plagiarism undertaking

I hereby solemnly declare that the work "**Elucidating the events of transcription factor binding in human brain enhancers**" presented in the following thesis is my own research efforts, except where otherwise acknowledged and that the thesis is my own composition.

The thesis has neither published previously nor does it contain any material from the published resources that can be considered as the violation of the international copyright law. No part of the thesis has been previously presented for any other degree.

I also declare that I am aware of the term copyright and plagiarism I will be responsible for the consequences of violation to these rules (if any) found in this thesis. The thesis has been checked for plagiarism by turnitin software.

Date: \_\_\_\_\_


A handwritten signature in blue ink, appearing to read 'Hizr', with a large, stylized loop above it.

**Hizran Khatoon**

## Certificate of Approval

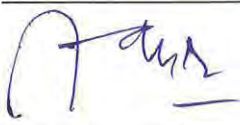
This is to certify that the research work presented in the thesis, entitled "Elucidating the Events of Transcription Factor Binding in Human Brain Enhancers" was conducted by **Miss Hizran Khatoon** under the supervision of **Prof. Dr. Amir Ali Abbasi**. No part of this thesis has been submitted anywhere else for any other degree. This thesis is submitted to the National Centre for Bioinformatics, Faculty of Biological Sciences, Quaid-i-Azam University, Islamabad, Pakistan, in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Field of Bioinformatics, National Centre for Bioinformatics, Quaid-i-Azam University, Islamabad, Pakistan.

Student Name: **Miss Hizran Khatoon**

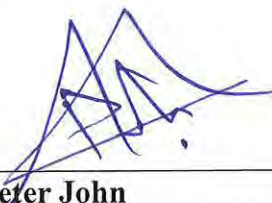
Signature: 

Examination Committee:

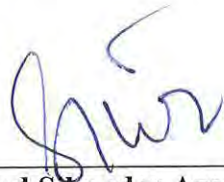
a) External Examiner 1:

  
**Prof. Dr. Mahmood A. Kayani**  
Biosciences Department, COMSATS,  
Institute of Information Technology,  
Islamabad.

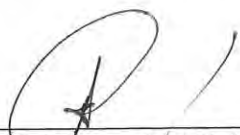
b) External Examiner 2:

  
**Prof. Dr. Peter John**  
Atta-ur-Rehman School of Applied Biosciences  
National University of Science and Technology  
(NUST) Islamabad

Internal Examiner:

  
**Prof. Dr. Syed Sikander Azam**  
Chairman  
National Centre for Bioinformatics.

Supervisor:

  
**Prof. Dr. Amir Ali Abbasi**  
Professor  
National Centre for Bioinformatics.

Dated: July 25, 2023

## **DEDICATION**

*I dedicate this effort to the*

*Holy Prophet **HAZRAT MUHAMMAD (PBUH)**, and his (PBUH) Beloved **Family***

*The greatest educationists of mankind*

***My Father (Late)***

***Imdad Hussain Imdad***

*Who would have been so happy and proud to see me become a doctor. May his memory forever be a comfort and a blessing. A father whose love and affection can never be replaced.*

*&*

***My Mother***

***Nasreen***

*Who has been a source of motivation and strength during moments of despair and discouragement. Her motherly care and support have been shown in incredible ways.*

## ACKNOWLEDGEMENT

*Words are bounded and knowledge is limited to praise **Almighty Allah**, The most beneficent and the most merciful. Whose countless blessing flourished my thoughts and bloomed my ambition to have the cherish fruit of my quite efforts in the form of His manuscripts. Without His will and guidance, I would be left stranded on an empty path. My handicapped words are insufficient to give thanks to the Lord of the worlds.*

*I cordially present my salutations and bouquet of greetings to the last Holy **Prophet Hazrat Muhammad (SAW)**, and his beloved **family**, the most perfect, ideal and exalted among people ever born on this planet, who (SAW) declared acquisition of knowledge an obligatory duty of every Muslim.*

*I express my deepest gratitude to my **teachers** and **mentors** from the very childhood to this day. Owe to their sincerity, I am able to achieve this goal. I am humbly grateful to my worthy supervisor **Professor Dr. Amir Ali Abbasi**, Quaid-i-Azam University, Islamabad, for his benevolence, invaluable ideas, suggestions, devotion and dedication which led this too long commitment to completion. I am cordially thankful to him for his kindness that is difficult to express in words. His skilful directions and dedicated advices have really guided me throughout the course of my research.*

*I want to pay my special thanks to Chairperson, Professor **Dr. Syed Sikander Azam**, Professor **Dr. Sajid Rasheed** and **Dr. Adnan**, National centre for Bioinformatics for being an outstanding leader of the NCB team and to the entire faculty and staff for being there for us in the time of need, especially **Mr. Naseer Raja** and **Mr. Talib Hussain**.*

*I gratefully acknowledge the **Education Monitoring Authority (EMA)** District Kurram for granting me relaxation as per need especially at the final stages of my PhD thesis. I must mention the kind cooperation of **Mr. Umar Ahmad Khan**, District Monitoring officer and **Mr. Muhammad Mateen** and all EMA staff.*

*I am grateful to my dear Lab fellows (too many to name), of whom I saw many left and many entered in my success journey at NCB, QAU. Each of them has left an impact, but those who stayed with me till the end have a special place in my heart. I would like to extend my gratitude to my senior lab fellows **Dr Rabail Zehra**, **Dr. Nashiman**, **Dr. Shahid**, **Dr. Irfan** for their valuable suggestions and help during the course of my work. Especially I would like to extend my special regard to my senior Lab fellow **Dr. Irfan Hussain** for giving extraordinary experiences during my research work and for urging me to think creatively and rationally. I am also grateful to my lab colleagues **Huma**, **Saba**, **Erum**, **Itfa**, **Shoaib**, **Alina**, for their support, concern and prayers. I would like to thank **Fatima batoool** and **Muhammad Abrar**, for being the most supportive fellows and also for their conducive discussion on various projects.*



*I feel proud that I have best friends like **Dr. Farwa, Dr. Yasmeen, Dr. Anisa,** and **Dr. Maria Ameen**, I have no word to define their purity and honesty. My prayers are always with them. **Dr. Maria Ameen & Dr. Mehreen**, I am really thankful to you both; you were always there for me whenever I need you. You always release my stress. I also wish to give my sincere thanks to my friends **Dr. Alia, Dr. Ghazala, Saima, Sameera, Sadaf, Qandil, Shama, Qabila, Rahila Bashir, Maryam, Lubna, Sidra Rahman, Ansa Andleeb and Rehana** for their moral support. They are a source of joy and happiness.*

*I profoundly appreciate my Grandfather **Professor Abis Hussain (Late)** for his inspiring guidance and valuable advices and **Sir Zeenat Bangash, Dr. Muneeb Hussain and Dr. Iqbal Hussain** whose indirect support and suggestions helped me in my study. I am greatly indebted to my Aunti **Miss Nargis** and mamu **Mutahar Hussain, Awais Hussain and Hassnain** for their guidance, love, and mellifluous affections, which hearten me to achieve success in every sphere of life. I submit my earnest thanks to my special relations, my uncles, Aunt and cousins. My handicap words can never express my sentiment, attachment and indebtedness to all of them.*

*No deeds can return and no words can express the affection, love, amiable sacrifices, attitude, unceasing prayers, advices, inspiration and support that received my Grandparents, parents (**Imdad Hussain and Nasreen**) and my siblings (**Miqdad Hussain, Amar-i-Yasir, Zeenat, Binty Hussain, Sitara and Nida**). I am blessed to have little younger brothers and sisters like angle (**Fakhir ali, Aatair ali, Samana, Nighat and Behisht**) who always bring smile on my face and I want to say a big thanks to **Sakina and Ulia** for their adoration.*

*I am thankful to my loving husband **Zahid Hussain** and all family for their prayers and good Wishes.*

*Last but not the least, I express my earnest love to my beloved parents **Imdad Hussain Imdad and Nasreen**, whose love and affection to their children will forever contribute to what we will ever achieve in our lives. I wish to say affable thanks to my father (late) for believing in his daughter just like a son and for paving my way to empowerment and independence through education. I feel an enormous admiration and immense obligation for my caring and loving mother who is just like a bank where we deposit all our hurts and worries. She always does extraordinary efforts for our easiness and comfort. No words can thank enough for their contribution in my life.*

*May Almighty Allah be their gracious guide and source of success in all their pursuits of life (Ameen)*

**Hizran Khatoon**

## Table of Contents

List of Figures .....	iv
List of Tables .....	v
List of Abbreviations .....	vi
Abstract .....	viii
1 Introduction.....	1
1.1 Eukaryotic Transcriptional Regulation .....	2
1.1.1 Cis-Regulatory Elements.....	2
1.1.2 Promoter, Core promoter and proximal promoter .....	3
1.1.3. Enhancer .....	4
1.1.4. Silencers.....	7
1.1.5. Insulators .....	7
1.1.6. Locus Control Regions .....	7
1.2. Cis-Regulatory Elements in Morphological complexity.....	7
1.3. Disease-Related Cis-Regulatory Elements in Human.....	8
1.4. Enhancer’s function in phenotypic evolution.....	9
1.5. Progression in the Human Genome.....	10
1.5.1. Human Accelerated Regions .....	10
1.5.2. Role of enhancer sequence acceleration in human cognition.....	11
1.6. Human Brain Expansion and Evolution of Modern human and Archaic hominin	13
1.6.1 Genetics bases between Modern human and Archaic hominins .....	14
1.6.2 Genetic differences between Human and Chimpanzees.....	15
1.7. Human brain development and gene regulation.....	16
1.8. Transcription Factors (TFs) and Transcription Factors Binding Sites (TFBSs) ....	17
1.9. The SOX gene family .....	19
1.9.1 SOX2 as a Transcription Factor .....	21
1.9.2. Role of SOX2 in brain developemnt: .....	22
1.10. Experimental approaches for Validation of Transcription Factor Binding Sites:	23
1.10.1. ChIP-chip.....	24
1.10.2. Chromatin Immunoprecipitation (ChIP-seq).....	24
1.10.3. DNase Hypersensitivity .....	24
1.10.4. EMSA .....	25
1.11. Molecular modelling theory and methods.....	26

1.11.1. Molecular Docking .....	26
1.11.2. ClusPro .....	26
1.11.3. PatchDock.....	27
1.11.4. FireDock .....	27
1.11.5. AutoDock.....	27
1.11.6. HADDOCK .....	28
1.12. Computer simulation .....	28
1.12.1. Classical molecular dynamics simulation.....	29
1.13. Aims and Objectives.....	31
2. Materials and Methods.....	32
2.1. DNA Protocols .....	33
2.1.1. Genomic DNA Extraction from Human Blood.....	33
2.2. PCR Protocols .....	33
2.2.1. Primer designing and dilution.....	33
2.2.2. PCR Amplification .....	34
2.2.3. Gel Electrophoresis.....	34
2.2.4. Purification of PCR Product.....	35
2.2.5. DNA Quantification .....	35
2.3. Cloning .....	35
2.3.1. Preparation of Media and Agar plates .....	35
2.3.2. Preparation of the Competent Cells.....	35
2.3.3. Restriction Digestion .....	36
2.3.4. Ligation.....	36
2.3.5. Transformation .....	36
2.3.6. Plasmid Extraction and purification .....	37
2.3.7. Restriction digestion .....	37
2.4. Protein Analysis .....	38
2.4.1. Protein expression.....	38
2.4.2. SDS PAGE Analysis .....	38
2.4.3. EMSA and Super-shift Assay.....	39
2.5. In Silico Analysis .....	39
2.5.1. Collection of sequences and comparative analysis.....	39
2.5.2. DNA and Protein modelling.....	40

2.6. DNA-protein docking and Complexes refinement.....	40
2.6.1. HADDOCK .....	40
2.6.2. UCSF Chimera .....	41
2.6.3. Interaction Analysis .....	41
2.7. Molecular dynamics simulation .....	42
2.7.1. Binding free energy calculations .....	42
3. Results.....	44
3.1. Comparative sequence and functional analysis of human accelerated enhancer hs1210 and domain organization of SOX2 .....	44
3.2. In vitro binding analysis of <i>SOX2</i> to target DNA containing ancestral and derived alleles.....	46
3.2.1. <i>SOX2</i> Amplification .....	46
3.2.2. Cloning and conformation of SOX2.....	46
3.2.3. Expression, purification and conformation of SOX2 Protein.....	47
3.2.4. EMSA and Super shift.....	48
3.3. Molecular docking characterization of the protein-DNA complex.....	49
3.3.1. Structure comparison between SOX2 DBDs bound to Human and Neanderthal DNA sequences.....	52
3.4. Evaluation of dynamical properties of the protein-DNA complex .....	55
4. Discussion.....	64
5. Conclusions.....	71
References.....	73
Appendices.....	93

## List of Figures

Figure 1. Diagrammatic view transcriptional regulation	3
Figure 2. Three promoters combine to regulate a gene	4
Figure 3. Schematic model of enhancer function	5
Figure 4. Models for enhancer's role in initiating transcription	6
Figure 5: Human and Chimpanzee structure	16
Figure 6. Existing models of enhancer activity	19
Figure 6. Schematic diagram of steps carried out in work design	32
Figure 7. Human accelerated enhancers with a transcription factor exclusive to Homo sapiens	45
Figure 8. Electropherogram of PCR products of SOX2	46
Figure 9. Electropherogram of circular and digested plasmid.	47
Figure 10. Expression of SOX2 protein	48
Figure 11. EMSA/Gel shift assay	49
Figure 12. Structural analysis of Ancestral <sup>T-allele</sup> DNA- SOX2 <sup>(HMG)</sup> complex	50
Figure 13. Structural analysis of Derived <sup>A-allele</sup> DNA- SOX2 <sup>(HMG)</sup> complex	51
Figure 14. (A-B) Structural analysis of DNA in complex with Ancestral <sup>T-allele</sup> SOX2 <sup>(HMG)</sup> and Derived <sup>A-allele</sup> SOX2 <sup>(HMG)</sup> complexes	53
Figure 15. Dynamic stability of DNA-protein complexes along the course of 100 ns Simulation	57
Figure 16. Residual flexibility of DNA-protein complexes along the course of 100 ns Simulation	58
Figure 17. Radius of gyration (RoG) Analysis	59
Figure 18. MD simulation based Inter- and Intramolecular hydrogen bonds Analysis.	62

## List of Tables

Table 1. Mammalian SOX2 factors and their subgroups.....	20
Table 2. Molecular docking based energetic profile evaluation through HADDOCK .....	51
Table 3. DNA binding residues of SOX2 (HMG) reported previously (experimentally determined) and in the present study.....	53
Table 4. Hydrogen bond interactions between DNA and HMG box of SOX2 protein determined through molecular docking experiments.....	54
Table 5. Hydrogen bonds between the SOX2 protein and DNA before and after MD simulation.....	60
Table 6. Free energy estimates based on Molecular Mechanics/Generalized Born Surface Area (MM/GBSA)..	62

## List of Abbreviations

3D. DART	3DNA-Driven DNA Analysis and Rebuilding Tool
BFE	Binding Free Energy.
Bp	Base Pair
ChIP	Chromatin Immunoprecipitation
CNE	Conserved Non-Coding Element
CNS	Central Nervous System
CRE	Cis Regulatory Elements
CTCF	CCCTC-Binding Factor
DNA	Deoxyribonucleic Acid
EMSA	Electrophoretic mobility shift Assay
GTF	General Transcription Factor
HARs	Human Accelerated Regions
HMG	High Mobility Group
Kb	Kilo Base
kDa	kilodalton
LB	Luria Broth
LCR	Locus Control Region
LRH	Long Rang Haplotype
Mb	Mega Base
mRNA	Messenger Ribonucleic Acid
MSA	Multiple Sequence Alignment
Mya	Million years ago
PDB	Protein Data Bank
PME	Particle Mesh Ewald
RNA	Ribonucleic Acid
SDS-PAGE	Sodium Dodecyl Sulphate Poly Acryl Amide Gel Electrophoresis
Shh	Sonic hedgehog
SNP	Single Nucleotide Polymorphism
SOX2	SRY (sex determining region Y)-box 2;
TBE	Tris borate EDTA

TBP	TATA Binding Protein
TF	Transcription Factor
TFBS	Transcription Factor Binding Site
TSS	Transcription Start Site
UCSC	University of California, Santa Cruz
ug	Micro gram
ul	Microliter



## **Abstract**

**Background:** The information responsible for animal diversity, complexity and embryonic patterning lies in non-coding regions of a genome. The non-coding portion of a human genome has become into a light now-a-days because of its regulatory aspects. Sequence acceleration in the human lineage has been found to harbour cis-regulatory elements among them enhancer constitute the most portion play a major role in gene regulation These sequences of the human genome compared with other vertebrates identified genomic regions that were highly conserved among vertebrates but fast-evolving in the human lineage called Human accelerated regions (HARs). HARs are short conserved genomic portions that have attained considerably more nucleotide variations than would be predicted in the human lineage following the divergence from chimpanzee. The rapid evolution of HARs is the reflection of their crucial role in the emergence of human-specific attributes. Human accelerated genomic regions were mostly non-coding in nature, and upon subjection to *in vivo* testing, confirmed the presence of cis-regulatory enhancers which regulate the expression of numerous developmental genes. Recent study has reported the positively selected single nucleotide variants (SNVs) within brain exclusive human accelerated enhancers (BE-HAEs) hs563 (hindbrain), hs304 (midbrain/forebrain) and hs1210 (forebrain). Furthermore, these SNVs were revealed to be modern human-specific by incorporating data from archaic hominins, since they were residing within the transcriptional factors binding sites (TFBSs) for RUNX1/3 (hs563), FOS/JUND (hs304) and SOX2 (hs1210). Although these results imply that these predicted alterations in TFBSs are crucial for determining current brain structure, much more research is necessary to confirm whether these changes actually correspond to functional modification. Here, in order to empirically test the binding events and binding affinity of SOX2 protein with the modern human (derived A allele) and archaic hominin (ancestral T allele) carrying DNA, we employed electrophoretic mobility shift assay (EMSA). In further analysis, molecular docking was explored to analyze how binding domain (DBD) of SOX2 TF interacts with the respective TFBSs. Moreover, exploiting the dynamic binding features for modern human and archaic hominin (Neanderthal) alleles based complexes, all-atoms biomolecular simulation was performed for the solvated systems using FF19SB force-field in AMBER20.

**Results:** We investigated the SOX2 SNV, with strongest results of positive selection in human population that was relevant for forebrain expression. We demonstrated that the

binding domain of SOX2 (HMG) binds in vitro with modern human-specific derived A-allele and ancestral T-allele carrying DNA sites. In further analysis, DNA-protein interaction studies indicated that HMG domain of SOX2 directly interacts with the minor groove of the derived A-allele as well with the ancestral T-allele type. Energetic profile evaluation through HADDOCK score calculations revealed higher affinity of SOX2<sup>(HMG)</sup> for the derived A-allele carrying target DNA site ( $-281.6 \pm 4.2$  kcal/mol), whereas relatively lower affinity was noticed for the ancestral T-allele carrying target DNA site ( $-270.3 \pm 4.0$  kcal/mol). The interface analysis also demonstrated a better interface in the derived A-allele-SOX2 complex as compared to ancestral T-allele, as evident by energetics and the number of hydrogen bonds. Furthermore Simulation analysis indicated the significant dynamic behavior and binding differences of THE DNA binding domain of HMG box with derived A-allele containing DNA target site when compared to site carrying ancestral T-allele. We speculate that such changes in TF affinity within BE-HAEs hs1210 and other enhancers could accumulate during recent history of human evolution to produce modifications in gene expression with functional consequences during forebrain formation.

**Conclusion:** Taken together with the combinations of in vitro and bioinformatics analysis to comparatively characterize the evolutionary significance of positively selected modern human specific substitution (T>A) within BE-HAEs hs1210. Our comparative molecular structural analysis showed that modern human-specific single nucleotide substitution has increased the affinity of SOX2 transcriptional factor for its target binding site within BE-HAEs hs1210. These findings suggest that this predicted enhanced affinity of SOX2 towards its target site could drive the target gene expression more robustly within forebrain of modern humans compared with the archaic humans or alternatively within novel territories in the forebrain of modern humans. These findings imply that adoptive changes in TF affinity within BE-HAE (hs1210) and other HAR enhancers may have altered gene expression patterns, which may have functional effects on the development and evolution of the forebrain.





## **1 Introduction**

No subject has created better intrigue or fury among the wonders that science has revealed about the universe in which we live, than evolution. Evolution describes changes that occur throughout time as species change and diverge to produce several descendant species. In place of the myths that had long since given us satisfaction, it provides the accurate history of our origins. *Homo sapiens* is substantially different from other non-human primates (NHPs) by its unique eccentricities such as physiological, morphological, anatomical and behavioral, including relative brain size, cranio facial attributes, bipedalism, vocal organs, small canine teeth, hair less skin, shorten finger, language, opposable elongated thumb and advance tool making capabilities (Carroll, 2003; Martin, Rayner, Gagneux, Barnwell, & Varki, 2005). These unique *Homo sapiens* specific phenotypic traits are emerged during the last 6 million years of evolution after its divergence from *Pan* (chimpanzee and baboon) lineage. The decoding of human and non-human primates' genome was providing an opportunity for evolutionary biologist to precisely understand how and what genetic underpinnings arose in the evolution of human unique oddities.

Moreover, the defining attribute of human evolution is structurally complex brain that differs from nonhuman primates in shape, size, organization and function. Prior investigation found that modern humans have somewhat smaller faces but larger brain case when examining and contrasting the endocasts of both modern and archaic humans. The brain and endocasts of modern humans are globular, with rounded, expanding cerebral portions, bulging parietal regions, and steep frontal regions. But there was an anterior-posterior elongation in our ancestors such as Neanderthals and other species (Neubauer, Hublin, & Gunz, 2018). The extent of the dynamics that led to the genesis and persistence of modern humans' facial, mandibular, cranial, and dental improvements is also demonstrated by evidence from craniodental data of our ancestors (Richter et al., 2017). Consequently, using this information to evaluate what happened differently during the evolution of the brain and its relevant attributes can reveal crucial details about the evolution of modern humans.

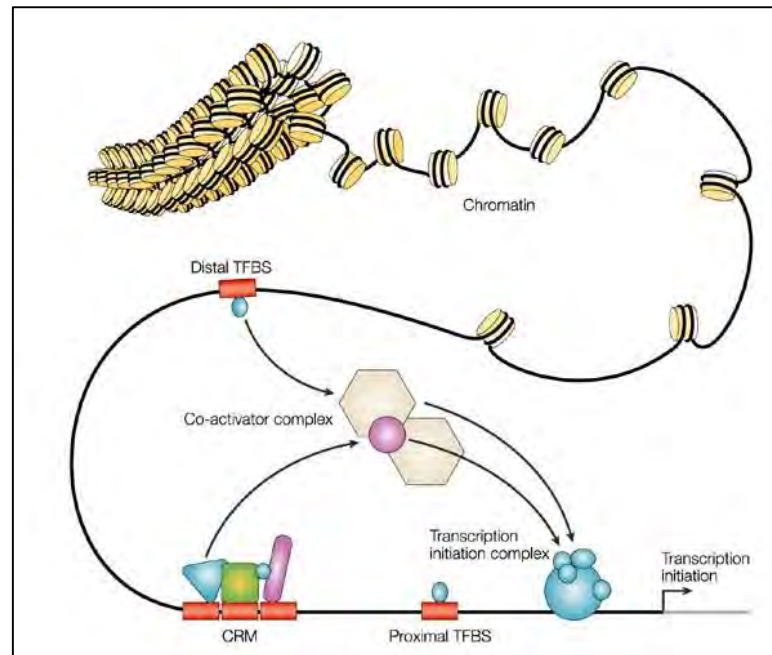
## 1.1 Eukaryotic Transcriptional Regulation

Numerous biological activities that occur within cells, include cell division, proliferation, and homeostasis, depend on the spatiotemporal regulation of gene expression. Eukaryotic transcription is more complicated than prokaryotic transcription and activity is regulated by integrated action of *cis*-regulatory (promoters, enhancers, silencers and insulators) and *trans*-regulatory (DNA binding proteins including transcription factors, activators and co-activators) elements (Levine, Cattoglio, & Tjian, 2014).

In eukaryotes, RNA polymerase II mediates the transcriptional regulation. Initially highly conserved General Transcription Factors (GTFs) binds sequentially to core promoter and assemble into Pre-Initiation Complex (PIC), compels RNA polymerase II to bind at Transcription Start Site (TSS) of the gene (Roeder 1996). After PIC assembly helicase unwind the double helix and set polymerase II to start ATP dependent transcription (Goodrich & Tjian, 1994). Most GTFs are produced when the nascent mRNA reaches 25–50 nucleotides and for termination phosphorylation of Ser2 terminal of Polymerase is required (Marshall, Peng, Xie, & Price, 1996; Marshall & Price, 1995). However, there are no stringent criteria for assembly of initiation complexes due to availability of diverse range of *cis*-regulatory elements which can modulate transcription per space and time (Müller & Tora, 2014). Brief notes on these elements are discussed below

### 1.1.1 Cis-Regulatory Elements

Based on position *cis*-regulatory elements (CRE) can be divided into two subsets: Proximally located promoters (core promoter and proximal promoter) and distally located *cis*-regulatory elements (enhancers, silencers and insulators) (**Figure 1**). All these *cis*-regulatory elements contain binding motifs for the *trans*-regulatory elements i.e. transcription factors, activators, or coactivators (Aziz Khan et al., 2018; G. A. Maston, S. K. Evans, & M. R. J. A. R. G. H. G. Green, 2006–b).



**Figure 1. Diagrammatic view transcriptional regulation**

*Diagrammatic illustration for overview of proximal and distal cis-regulatory elements (CRE): The promoter region consists of core promoter, encompassing the transcription start site (TSS), and proximal promoter. The distal regulatory region is composed of enhancer, silencer, and insulator. Transcription factors bind to these CREs by transcription factor binding site (TFBS) Adapted from (Aziz Khan et al., 2018).*

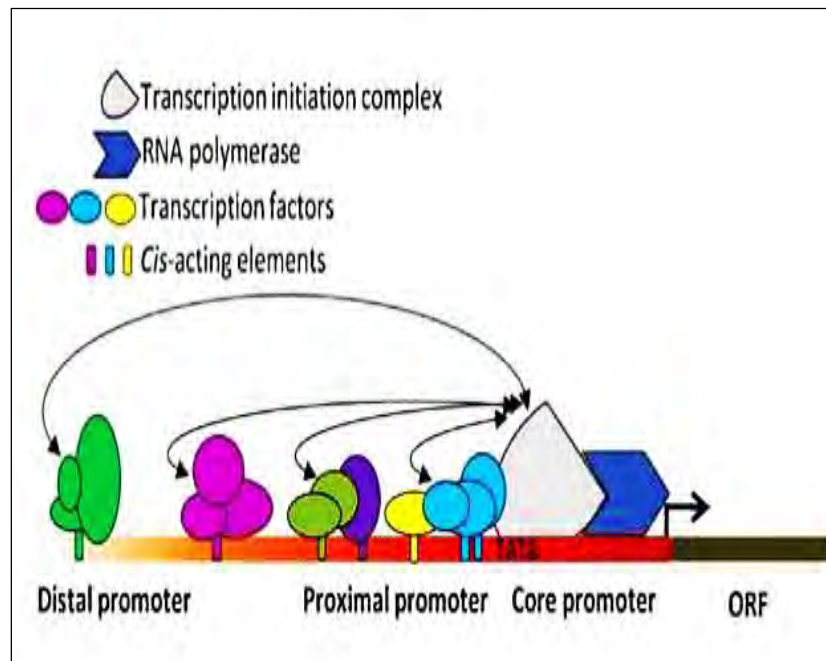
### 1.1.2 Promoter, Core promoter and proximal promoter

An essential group of sequences known as promoters is responsible for the transcriptional initiation of RNA and protein-coding genes (R. K. Umarov & V. V. Solovyev, 2017; R. K. Umarov & V. V. J. P. o. Solovyev, 2017). Activating transcription factors (TFs) could bind to the regulatory motifs in such 5' flanking regions to start the expression of genes. Promoters are further classified as core promoter (upstream of gene), proximal promoter (upstream of core promoter) and distal promoters (anywhere in the genome) according to their presence.

The **core promoter** is at immediate vicinity, about 30 bps upstream of TSS of a gene. It integrates total input from the transcriptional machinery anchored to proximal and distal elements and build a refined and regulated proportion to start transcription (Figure 2) (Aziz Khan et al., 2018). TATA-box was first reported and characterized

core promoter. Interestingly TATA-box and TBP (TATA-box Binding Protein) are conserved throughout ancient bacteria to humans.

The **proximal promoter** generally present at 200 to 250 bps upstream to TSS and contain multiple binding motifs for activator elements and capable of producing tethering effect to induce homotypic transcription factor binding (Calhoun, Stathopoulos, & Levine, 2002).



**Figure 2. Three promoters combine to regulate a gene**

*Three types of promoters: core promoter, proximal promoter and Distal promoter located upstream of a gene which together play role in tissue-specific expression of a gene. Adapted from (Davidson & Erwin, 2006).*

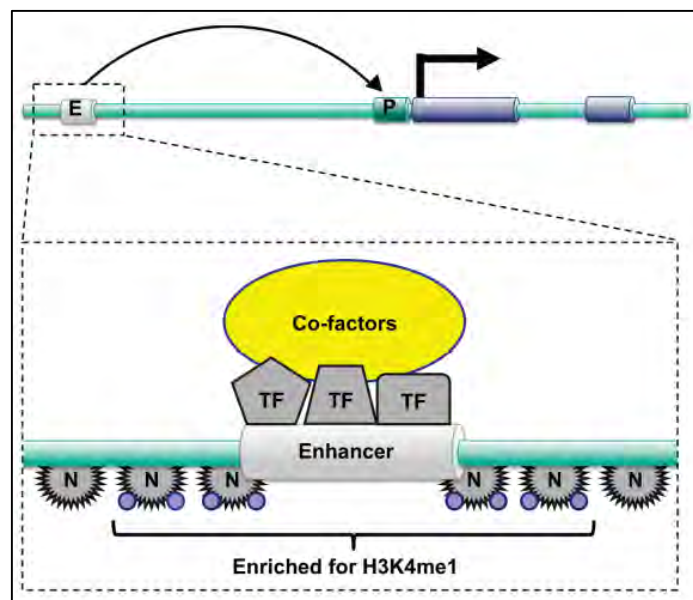
### 1.1.3. Enhancer

The effects of SV40 DNA on the ectopic expression of a rabbit cloned beta-globin gene were initially mentioned as the cause of the word "enhancer". Transcription was made active by the SV40 DNA elements from a distance and independent of their orientation in regard to the target gene (Banerji, Rusconi, & Schaffner, 1981). Small DNA regions termed as enhancers, typically a few hundred base pairs long, have been functionally characterized and are normally occupied various transcription factors through small and specific DNA sequence (motifs), to control transcription (Levine &



Tjian, 2003; Panne, 2008; Spitz & Furlong, 2012). They are the DNA elements that transcribe the gene from a distance no matter what the position of an enhancer on DNA is. Most of the transcriptional regulation of developmental stages of mammals is because of the diverse activity of enhancers that are bound by transcription factors and controls the specific gene expression patterns of multiple cell types (Bulger & Groudine, 2011; Hawrylycz et al., 2012; Maston et al., 2006b).

Enhancers are basically scattered through the 98% of human genome which is non-protein coding and hence producing a tremendous amount of search space in order to predict them. They tend to control their target genes in *cis* position. *Cis*-regulatory modules have highly variable locations from their target genes. They can be either found at upstream or downstream region, and even within the intronic region of the genes (L. A. Pennacchio, W. Bickmore, A. Dean, M. A. Nobrega, & G. J. N. R. G. Bejerano, 2013b). Enhancers hold multiple transcription factor binding sites that are required to control its activity. Activation of enhancers coincides with DNase I hypersensitivity of these regions and sometimes with specific histone modifications (Figure 3) (Lelli, Slattery, & Mann, 2012; Shlyueva, Stampfel, & Stark, 2014).

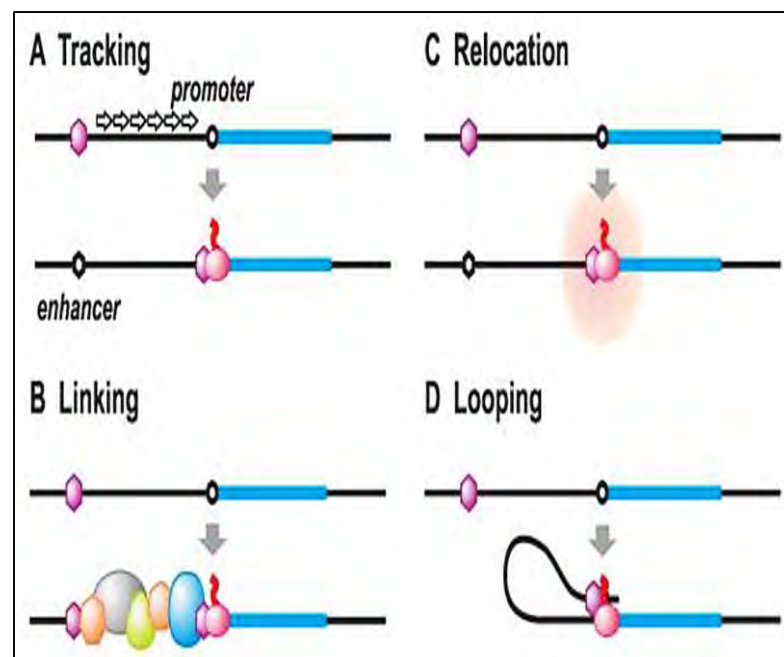


**Figure 3. Schematic model of enhancer function**

*The figure shows the different epigenetic marks and multiple binding sites for transcription factors on enhancers, in the right cellular, temporal and spatial conditions, they are bound by the appropriate transcription factors, which in turn recruit various cofactors, and the enhancers contacts its cognate promoter to drive expression. Nucleosomes flanking active*

enhancers are often enriched for monomethylation of histone H3, lysine 4 (H3K4me1). Shown in blue. Reproduced from (Noonan & McCallion, 2010).

Four different models have been described for the gene expression regulation by enhancers: The enhancer is bound by a pink-colored transcription factor (TF), which moves down the DNA toward the promoter position where it binds with the polymerase to initiate transcription. In the linking model, transcription is started by polymerizing more TFs in the direction of the promoter once a Transcription factor (TF) in pink color bind to the enhancer position. Relocation model, where the genes relocate to make enhancer-promoter interaction feasible, and Loops model, where a protein-protein interaction brings the enhancer and the appropriate promoter into close proximity. It causes transcriptional activation and loops out any interfering chromatin (Figure 4) Adopted from: (Kolovos, Knoch, Grosveld, Cook, Papantonis, et al., 2012).



**Figure 4. Models for enhancer's role in initiating transcription**

Enhancers models Adopted from: (Kolovos, Knoch, Grosveld, Cook, & Papantonis, 2012) describing the gene regulation by enhancer shown as ((A) TM (Tracking model), (B) LM (Linking model), (C) RM (Relocation model), (D) LM (DNA looping model)

#### 1.1.4. Silencers

Silencers were first discovered in yeast. They are DNA sequences, capable of switching off the promoters regardless of their orientation and position, thus having an inhibiting effect on the transcriptional activity of the corresponding gene (Brand, Breeden, Abraham, Sternglanz, & Nasmyth, 1985). There is a similarity between the features of enhancers and silencers (Ogbourne & Antalis, 1998). Silencers contain repressor binding sites and are normally present upstream of the transcription start site. Upon binding to the silencer, the repressors inhibit the initiation complex formation leading to the suppression of transcription (G. A. Maston, S. K. Evans, & M. R. Green, 2006a).

#### 1.1.5. Insulators

Insulators are short *Cis*-acting DNA sequences. Their length may vary from 0.5 to 3 kilobase pairs. They inhibit the interaction between enhancers and promoters by functioning position-dependently and orientation-independently, thus blocking genes by inhibiting or limiting the effects of transcription activity of the nearby genes. Insulators differentiate heterochromatin from euchromatin, partitioning and outlining the genome into contrasting spheres of expression (Maston et al., 2006b; Sun & Elgin, 1999). Insulators of the vertebrate genome have binding sites for cohesion complexes and CCCTC binding factor (CTCF) which are involved in looping interactions (Zuin et al., 2014).

#### 1.1.6. Locus Control Regions

Gene clusters are mainly regulated by transcription regulatory elements known as locus control regions. They consist of some *Cis*-acting elements like silencer and enhancers, to which binds trans-acting elements such as chromatin modifiers, repressors, co-activators and tissue-specific transcription factors, thus enabling the locus control region capable of gene expression regulation and increases the tissue specific expression of genes linked with each other (Li, Peterson, Fang, & Stamatoyannopoulos, 2002).

### 1.2. Cis-Regulatory Elements in Morphological complexity

A significant portion of the genetic material responsible for animal diversity is found in the large non-coding sequences of genome (Abbasi, 2011). The huge genetic

variability in non-coding regulatory elements has been linked to phenotypic variation, including illness, morphology, physiology, and behavior, according to recent studies (Tycko et al., 2019). Previous research demonstrates that the highly complex and the deeply conserved cis regulatory elements in the genome have contribution much more towards complexity in the organization of vertebrate body. A remarkable rise in transcription factors, an increase in the number of regulatory elements that work to cis-regulate gene expression, and morphological complexity are all associated with animal evolution (Levine & Tjian, 2003). Thus, functional diversity of genes modular CREs has been linked to both independent morphological similarity evolution and evolutionary novelty (Prud'Homme et al., 2006). The cumulative organization of facts from field of comparative genomics and evolutionary developmental biology, it is extensively accepted at the current duration that the morphological diversity is also related to the evolutionary development of the cis-regulatory DNA that regulates the spatiotemporal expression of the developmental genes (Levine & Tjian, 2003). Thus, functional diversification of modular CREs of genes contributed to evolutionary novelty and independent evolution of morphologic resemblances (Franchini & Pollard, 2015).

Further if sequence of the transcription factor binding sites change it will directly show effect on binding potency of the transcription factors so it is associated with phenotypic traits (Hornshøj et al., 2018). The cis-regulatory regions contain a large number of sequence variants that are acceptable and have little impact on their activity and expression. Such modifications brought about morphological diversity and complexity during the evolutionary processes. (Franchini & Pollard, 2015). Contrary to coding sequences, many sequence alterations inside cis-acting regulatory elements may thus exert acceptable effects on the activity and level of expression of the linked genes, serving as a catalyst for the evolution of morphological complexity and diversity (Carroll, 2005).

### **1.3. Disease-Related Cis-Regulatory Elements in Human**

Enhancer sequence regulate expressions of the developmentally important genes of a tissue in specific manners. The enhancer element typically has a length of 500 bp and has multiple binding sites for different transcription factors. Changes in cis-acting sequences cause morphological, physiological, and developmental changes as

well as more severe effects on the expression of linked genes (Anand et al., 2003; Shashikant, Bolanowsky, Anand, & Anderson, 2007). The modifications in the cis-acting regulatory repository may be related to the defect in human development. The disease importance of cis-acting mutations was overlooked in earlier decades due to a lack of computational and functional approaches for the detection and functional characterization of the huge non-coding genomic space's cis-acting gene regulatory elements. Human genetic disorders can be caused by 1459 mutations in regulatory components of more than 700 genes, according to data and statistics from the Human Gene Mutation Database in 2009 (Epstein, 2009). Preaxial polydactyly (PPD), which has mirror-image digit duplication and is similar to an ordinary human developmental aberration, is more closely related to the *Shh* (Sonic hedgehog gene). Preaxial polydactyly is caused by an enhancer element mutation in the *LMBR1* gene's intron 5 that prevents *Shh* from expressing polarized on the posterior limb side (Lettice et al., 2003).

One more analogy to elucidate subject would be that mutations in another gene for a protein were once thought to be one of the causes of limb deformities. While *in vivo* experiments using the enhancer demonstrate that the phenotypes of limb deformation are caused by particular germ line mutations in non-coding areas (Allou et al., 2021). Instead of mutations inside the exon of the *formin* gene, changes in the limb specific enhancer intron were the real causes of limb abnormalities (Dickmeis & Müller, 2005). Similarly, Hirschsprung disease has been linked to single nucleotide alterations in the enhancer element inside intron-1 of the *RET* gene (*HSCR*) (Arnold et al., 2009).

#### **1.4. Enhancer's function in phenotypic evolution**

The most often tested cis-regulatory elements fall within the category of enhancers. From their first discovery to their ongoing dissection to determine genetic variability even within the human population, several research have carefully explored these mechanisms. It is now clear that the genomes contain a wide range of metabolic alterations that offer information about how to classify regulatory genomic elements. For an enormous number of TFs and their co-factors in different virtual cellular environments, the chromatin structure, many histone modifications, and binding sites for different TFs have generally been determined. Also noteworthy is the fact that 10–

20% of the human genome, which might include enhancers, promoters, and other regulatory regions, controls the expression of genes (Pennacchio et al., 2013b). Enhancers are thought to make up the bulk of the regulatory repertoire, making them more likely to incorporate alterations that might aid in the formation of a phenotype specific to a species.

The crucial role that enhancers play in promoting evolution is also reflected in their modular manner of function. Notably, 80% of human GWAS-associated SNPs are non-coding, indicating that a higher proportion had to be present in such regulatory regions (Hindorff et al., 2009). The ability of a gene to express itself in a variety of tissues and cells makes it susceptible to mutation, which can be harmful. However, due to the modularity of enhancers, it is possible to distinguish between a completely different expression pattern seen in a different context where the enhancer or assisting regulatory regions may not be active and tissue-specific coordination between enhancer and other regulatory elements that can drive the expression of a gene in one cellular context (L. A. Pennacchio, W. Bickmore, A. Dean, M. A. Nobrega, & G. Bejerano, 2013a). Therefore, selection and mutation in enhancers can work together to make regions of choice into sources of adaptability and fitness. There are several instances where the contribution of enhancers to the gain or loss of a trait in a lineage-specific aspect is evident. For example, in *Drosophila*, such adaptations include the creation of larval trichomes and the coloration of the wing (Pennacchio et al., 2013a). Lactase persistence in the human population is an excellent illustration of a regulatory mutation that influences the phenotypic expression (Fang, Ahn, Wodziak, & Sibley, 2012).

## **1.5. Progression in the Human Genome**

Two significant aspects of human genome progression — Human accelerated regions (HARs) and species-specific genome level reorganizations such as segmental gene duplication, deletion, and insertion have come to notice over the past few years, taking into account that humans are the most distinct of all primates and the most advanced in terms of their physiologic and anatomical characteristics -(Hubisz & Pollard, 2014; Sassa, 2013).

### **1.5.1. Human Accelerated Regions**

Accelerated regions created as a result of single nucleotide substitutions are the most prevalent form of fine-tuning regulatory elements in creating species specific loss or gain of traits. Human accelerated DNA fragments or HARs are those bits of the genome that have experienced frequent sequential changes after the human-chimp split (Hubisz & Pollard, 2014). Not only are the substitutions comprising the human lineage specific acceleration important, their presence in a highly conserved, evolutionarily substantial patches of the genome make the pursuit of dissecting these regions mandatory (Levchenko, Kanapin, Samsonova, & Gainetdinov, 2017). This theory contends that in vivo analysis of these human accelerated non-coding areas appears to have been triggered by the existence of cis-regulatory transcriptional enhancers, which regulate the expression of numerous developmental genes (Prabhakar et al., 2008). In a meta analysis of five studies (Bird et al., 2007; Bush & Lahn, 2008; Pollard et al., 2006; Prabhakar, Noonan, Pääbo, & Rubin, 2006; Zuckerkandl & Pauling, 1965) that predicted 2649 non-coding HARs in the Human genome after the protein coding regions were excluded. The majority of these non-coding HARs were found in intronic and intergenic regions (Capra, Erwin, McKinsey, Rubenstein, & Pollard, 2013). Intriguingly, studies also asserted that the accelerated regions of human genome comprise significant portion of these accelerated portions of neural enhancers (Doan et al., 2016). According to the evolutionary study, the rate at which enhancer sequences evolved during vertebrate land adaptation in comparison to coding and non-coding/non-enhancer genomic sequences was similarly accelerated (Yousaf, et al., 2015).

### **1.5.2. Role of enhancer sequence acceleration in human cognition**

Numerous accelerated regions of genomes contain developmental enhancers, and genetic modifications in these regions can result in significant changes to the function of brain (Burbano et al., 2012; Hubisz & Pollard, 2014; Prabhakar et al., 2008). Additionally, evolutionary investigations have supported the acceleration of enhancer portions during vertebrate land adaptation relative to coding and non-coding/non-enhancer genomic chunks (Yousaf, Raza, et al., 2015). Recent research has shown that human-specific mutations in enhancers can significantly alter gene regulation processes and ultimately result in disparities in brain size (J. Lomax Boyd et al., 2015). For instance, a recent study has supported this theory by showing significant variations in brain size were caused by human-specific alterations in a neuro-developmental

enhancer of the FZD8 gene (Franchini & Pollard, 2015). Sequential alterations that quickly accumulated in human brain enhancers should be assessed in order to determine whether they are necessary enhancers and what role they play in primarily regulating the spatiotemporal expression of the genes (Franchini & Pollard, 2015). Evaluation of the sequential modifications that quickly accumulated in human brain enhancers is required to determine the significance of enhancers and their function in primarily directing the spatiotemporal expression of the genes (Maston et al., 2006a). Recent study reported the SNPs catalog encompasses 27 SNPs linked to Alzheimer diseases, and 5 of these SNPs occur in the super-enhancers of brain tissue. Thus this study revealed that ~19% (5/27) of all of the Alzheimer diseases SNPs occur in the 1.4% of the genome comprised by brain tissue super-enhancers (Hnisz et al., 2013). Moreover, 67 SNPs in the non-coding sequence are found to be associated with type 1 diabetes, among them 13 SNPs were occurring in the super-enhancer regions of genes with prominent roles in T-helper cells biology (Hnisz et al., 2013). Similar to this, a recent study examined the consequences of deleting the Gli3 enhancers that are specific to the limbs (hs1586/mm1179), which significantly reduced Gli3 expression in the embryonic hand plate and revealed forelimb-specific polydactyly (Osterwalder et al., 2018).

Another Studies have also shown a significant association between the gene types of mental illness or those who are believed to play a significant function in accelerated regions. In the introns of the HAR-associated gene autism susceptibility candidate 2 (AUTS2), three human-specific variations have been discovered, which is rumored to harbour structural variants that contribute to a variety of neurological abnormalities (Pollard et al., 2006; Prabhakar et al., 2006). Another example is the cut-like homeobox 1 (CUX1) gene, which is a transcriptional repressor. According to Prabhakar and colleagues, this gene is associated to a HAR-containing enhancer that gains an additional transcription factor binding site as a result of a G>A mutation (Prabhakar et al., 2006). The overexpression of the gene and this enhancer substitution together cause the onset of autism and other intellectual impairments (Doan et al., 2016). A comprehensive study linking SNPs linked to diseases such as schizophrenia through accelerated sections has been published, demonstrating the evidence that gene variants or SNPs, which linked to the schizophrenia-specific disorder in Homo sapiens also have a linked to accelerated portion which may also serve as a regulatory genomic



segments (Britten & Davidson, 1969; Levchenko et al., 2017). Another interesting study discovered a very high concentration of HARs in the introns of the Neuronal PAS domain-containing protein 3 (NPAS3) (Kamm, Pisciotto, Kliger, & Franchini, 2013). In the earlier reported research, NPAS3's function in brain development and neuro-signaling has been well established (Brunskill et al., 2005). Polypyrimidine tract binding protein 2 (PTBP2) and glypican 4 (GPC4) are two other genes that have reportedly been controlled by the accelerated regulatory areas and have a role in brain development (Bird et al., 2007).

### **1.6. Human Brain Expansion and Evolution of Modern human and Archaic hominin**

Evolution is an ongoing process, today operating at a faster rate than in times past in this human dominated world. One species on our planet developed the intelligence to wonder about the past few billion years after life first appeared on Earth and half a billion years after the evolution of the first animals (Pearce, Stringer, & Dunbar, 2013). The human tale has elicited more powerful feelings than any other aspect of the evolution story. Human genome is the home of millions of nucleotides encapsulating the mysteries of human development. Understanding of this complex catalog of protein coding and non-coding strings of DNA is a phenomenon of continuous research in the world of genomics. The scientific community holds the belief that the human brain is so highly advanced in comparison to the brains of other primates that it must be in a class by itself. Many biologists hypothesized that over a million-year transition period from ancestral primates to anatomically modern humans, the size and complexity of brains gradually increased (Pearce, Stringer, & Dunbar, 2013; S. A. Williams, Middleton, Villamil, & Shattuck, 2016). This increase contains numerous bursts and stability periods on a smaller timescale, although appearing progressive over a long length of evolutionary time. During the first few million years of hominid evolution, the most pronounced increase in brain size was observed (McHenry, 1994). Given the history of increasing brain enlargement throughout the lineage to humans, species that diverged from this lineage more recently, such as apes, likely to have larger and more complex brains than species that diverged earlier, such as Prosimians (Gilbert, Dobyans, & Lahn, 2005).

In last few decades, various efforts have been made to decipher the complex genomic architecture of human (Karolchik et al., 2003). Completion of human genome project has opened avenues to several other domains in order to understand the mechanism of DNA. Comprehensive understanding of human genome requires robust experimental and computational techniques to elucidate gene expression (Burley et al., 1999; Clark et al., 2001). It has been a long way to scientifically probe the structure and function of a developed vertebrate brain which has three major divisions: forebrain, midbrain and hindbrain out of which forebrain is the major area of cognitive abilities (Charvet & Striedter, 2011). These three divisions can be easily distinguished during development by both their gross appearance and extent of cellular differentiation. The more evolved a brain region is, the more differentiated its cells become during the development process. The adult vertebrate forebrain embraced the products of embryonic telencephalon and diencephalon, contains numerous other adult structures is complex in both structure and function. This section of CNS is the most evolved and complex of brain divisions, known for the regulation of several purposes which include thought and emotion, vision and olfaction, behavior and homeostatic mechanisms such as hunger and circadian rhythms and voluntary movements. The midbrain or mesencephalon is the most rostral portion of the brainstem lying in the middle of three primary cerebral vesicles of vertebrate brain. The primary role of the midbrain is to support movement as well as the processing of auditory and visual information. The developing vertebrate brain's hindbrain, also known as the rhombencephalon, is made up of the pons, medulla oblongata, and cerebellum. The hindbrain performs activities such as breathing, motor activity, rhythm, wakefulness, and sleep that are essential for survival. It is a corridor between forebrain and midbrain on one side and the spinal cord on the other.

### **1.6.1 Genetics bases between Modern human and Archaic hominins**

The advent of comparative genomics and the success of Human Genome Project has greatly escalate our ability to uncover the genetic characteristics that differentiate humans from other primates and has helped us to begin to apprehend the genetic bases of human phenotypic specializations (Hacia, 2001; Johnson et al., 2008). Human brain is very large as compare to other vertebrates, weighing about 1400 g, which is roughly three times larger than those of other great apes. Gene control is a crucial factor in

optimizing the circuits of brain that differentiate extremely intellectual human brain activity from comparably less developed adaptive non-human primates brain activity (Cáceres et al., 2003a). Necessitating enhancer's role in regulating the spatial and temporal gene expressions, sequential changes that rapidly assembled in human brain enhancers were uncovered (Maston et al., 2006a). However, the recently reported brain exclusive enhancers showing signatures of positive selection in human lineage were validate in vitro (Zehra & Abbasi, 2018).

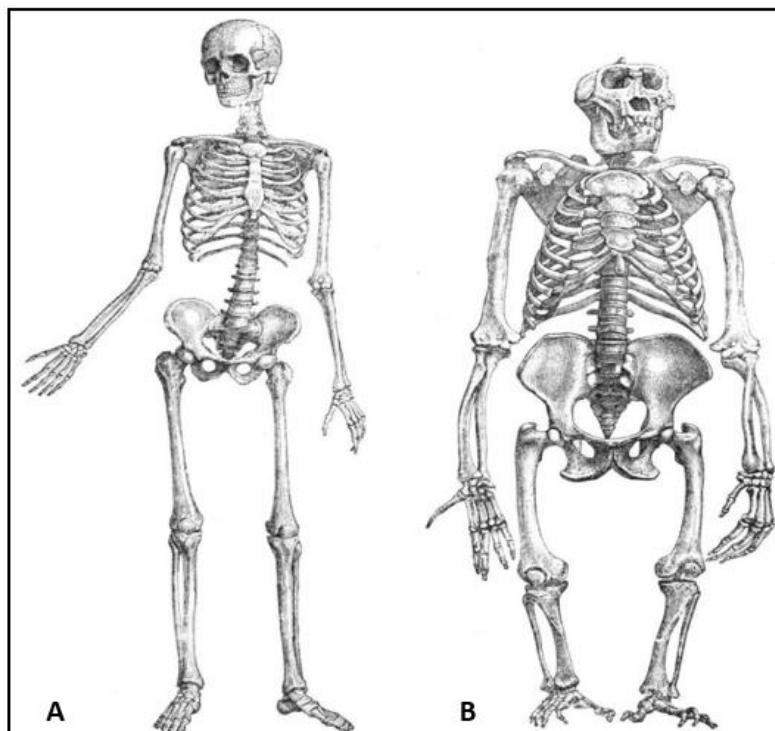
### **1.6.2 Genetic differences between Human and Chimpanzees**

Humans and chimpanzees have evolved progressively from their common ancestor of *Pan* and *Homo* (M. L. Wilson, 2021). Interestingly about 6.5–7.5 years back; identifying the genetic components that encode traits of human physiological and mental identity are still of great interest. Chimpanzees are genetically very similar to humans about 98.6% of their DNA is actually shared by them. Which has transformed over time. Based on protein content, 29% of genes encode similar amino acid sequences, interestingly, since chimpanzees and humans are closely related, their ancestry possessed advanced intelligence not found in many other mammals (M. L. Wilson, 2021).

After divergence of their ancestor lineages, human and chimpanzee genomes underwent multiple changes, including single nucleotide substitutions, deletions and duplications of DNA fragments of different sizes, insertion of transposable components, and rearranging of chromosomes (Suntsova & Buzdin, 2020). Humans have 46 chromosomes, and chromosome 2 was created by the joining of two ancestral chromosomes, hence there are chromosomal differences in the molecular genomic differences between apes (chimpanzees) and humans (Suntsova & Buzdin, 2020). Humans have large pericentric inversions on chr1 and chr18, and chimpanzees have bulky pericentric inversions on Chr 4, Chr 5, Chr 9, Chr 12, Chr 15-Chr 17. In humans, 134 genes increased the copy number and 6 genes decreased the copy number. In chimpanzees, 37 genes had increased copy number and 15 decreased (Suntsova & Buzdin, 2020). Human Accelerated Regions (HAR) show that human substitution rates have accelerated significantly since divergence. Genes associated with transcriptional regulation and neurodevelopment were found to be significantly enriched alongside

HAR. Therefore, rapid changes in human HAR1 may be related to the evolution of the human brain (Kehrer-Sawatzki & Cooper, 2007).

From structure point of view, Chimpanzees have a cranium capacity between 400 and 600 cm<sup>3</sup>, but humans have a cranium capacity between 1400 and 1500 cm<sup>3</sup>. Humans typically have a high forehead and an elevated nose, compared to chimpanzees' slanted forehead and flat nose. Chimps have longer arms as compared to their legs which can reach below knees, while humans in comparison to chimps have shorter arms. Chimps have C shaped spine while humans have S shaped spine (Kehrer-Sawatzki & Cooper, 2007).



**Figure 5: Human and Chimpanzee structure**

(A) Represents the Human skeleton. (B) Represents the Chimpanzee skeleton.

### 1.7. Human brain development and gene regulation

Gene regulation facilitates the fine-tuning of brain circuits that distinguish profoundly different cognitive function from that of the protein through gene regulation (Cáceres et al., 2003b). Neocortex, the frontal lobe, and the overall brain size of primate evolution exhibit disproportionate enlargement, which are features that support their intelligence (Dunbar & Shultz, 2007). The human brain is three times as large as a great apes, and it is better suited to performing extremely complex assessments using

language and cognitive abilities (Geschwind & Rakic, 2013). Additionally, evidence points to the fact that the human neocortex is larger and has distinct cell-cycle characteristics that promote enhanced corticogenesis (J Lomax Boyd et al., 2015). The behavioral traits that distinguish chimpanzees, human closest living relative, into two distinct cognitive strata, are thought to be influenced by changes in gene sequences, but little evidence linking these two ideas has been uncovered. However, it has been accepted that gene regulation and the spatiotemporal expression of genes play a crucial part in determining the current structure of the remarkably adapted brain of modern humans (Cáceres et al., 2003a; W. Enard et al., 2002; W. J. C. B. Enard, 2015; J. Gu & X. Gu, 2003; J. Gu & X. J. T. i. G. Gu, 2003). It is hypothesized that the cerebral cortex of chimpanzees and humans depends on particular gene expression patterns. This study discovered that 169 genes express differently in humans and chimps. Using macaques as an out-group, 91 of these genes indicated that they were expressed differentially only in the human lineage (Cáceres et al., 2003a). Around 90% of the genes in the human lineage that revealed differential expression belonged to the brain, while nearly equal numbers of genes showed up- and down-regulation between humans and chimpanzees in the liver and heart (26a) (Cáceres et al., 2003a). Another studies reported the results of 54 pre-frontal cortex (PFC) genes having lineage-specific up regulation in the human PFC gene after divergence from the other hominoids (Geschwind & Rakic, 2013).

### **1.8. Transcription Factors (TFs) and Transcription Factors Binding Sites (TFBSs)**

Although the human genome has the complete information on every gene, not every gene is expressed in every cell. Different genes express differently in various tissues and stages of development. To demonstrate that the right genes are regulated at the right time, transcription factors are essential. In the broadest sense, "transcription factor" refers to any protein that has the ability to regulate gene transcription in cells. Typically, transcription factors change chromatin structure, recruit cofactors to the target genes, or directly bind to specific DNA sequences known as "regulatory elements" to control gene transcription (Lee et al., 2000).

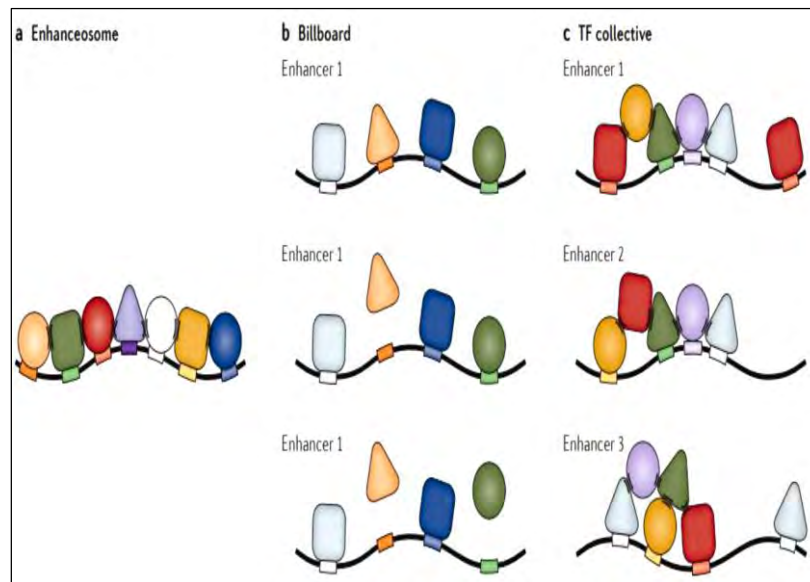
Transcription factors usually recognize small degenerated DNA sequences of 6-12 bps in length. This short sequence specificity shows that more intricate principles are present than just affinity of distinct transcription factors that are responsible for

controlling the enhancer occupancy as well as their functional outcome. Normally there are four to eight dissimilar TFs that can be bound within an enhancer and every factor can usually bind to those enhancer regions that have clusters of multiple transcription factor binding sites. Transcription factors can frequently occupy different sets of enhancers depending upon the condition. Combined occupancy of TFs can produce different types of transcriptional outputs, depending upon the interactions of TFs with each other. In some cases, *cis*-regulatory modules triggering is directly proportionate to concentration of the single transcription factor. In addition, cooperative binding of transcription factors has a non-linear connection between the intensity of concentration and how much each site is occupied by a particular enhancer. The occupancy of TF is affected by the nucleosome location at enhancer, where histones marks and transcription factors can make every effort for the contact of the DNA. Specific histone modifications such as: Histone H3 mono-methylation on lysine 4 (H3K4me1) and histone H3 acetylation of lysine 27 (H3K27ac) are highly associated with *cis*-regulatory elements, where H3K27ac is strongly related to the activity of enhancer and the expression of gene present in the nearest proximity (Spitz & Furlong, 2012).

There are two major properties of an enhancer: motif composition and motif positioning. Motif configuration is the presence of binding sites for particular TFs within an enhancer, which essential for regulating the expression of genes in a certain cell type. Motif positioning is frequently specified as as motif grammar and is basically the respective order, orientation and spacing of transcription factor binding motifs inside an enhancer. Motif position makes sure that that protein-protein interactions are aided by the proper positioning of the transcription factors and in that way promotes cooperative binding along with the recruitment co-factors and transcriptional machinery. This has led to three models of enhancers includes: enhanceosome, billboard and TF collective as shown in Figure 5.

In the “enhanceosome” model for enhancer activity, extremely ordered protein interface is formed by recruited TFs that have a need for a strict and specific TFBSs positioning relative to one another in the DNA. In such an ordered structure, recruitment of cooperative TF cause a potent and switch-like activation (Senger et al., 2004). A billboard model of enhancer activity enables more flexible location of TFBSs. In this model, TFs cooperate, however there are some restrictions on where their binding motifs can be positioned, as enhancers works by means of information display elements

(Arnosti & Kulkarni, 2005). The third model TF collective proposes that a limit for a determinant of enhancers that generates the overall output is extended considering interactions between proteins in addition to a linear sequence-based coding (Spitz & Furlong, 2012).



**Figure 6. Existing models of enhancer activity**

*(a) In the model of enhancesosome, all transcription factors that binds within the enhancer that are important for cooperative occupancy and enhancer activation. The motif configuration and positioning may act as a scaffold for the recruitment of all TFs cooperatively and as a result, highly ordered protein interface is produced that controls gene expression. (b) Billboard or type 1 enhancer model that allows a flexible motif positioning, , organizational restraints or a loose distance. The enhancer's binding sites can only be partially active at any given time (c) Third model is a TF collective that embodies a state in which a similar set of TFs binds to multiple enhancers and can occupy each and every enhancer in a different way. In certain circumstances, only a small subset of transcription factors has specific binding motifs, but protein-protein interactions kept the remaining TFs connected to the enhancers. By using variable motif conformation and adaptable motif grammar, collective binding of TFs onto binding motifs occurs (Junion et al., 2012).*

### 1.9. The SOX gene family

The family of SOX gene underwent breakthrough research following the discovery of the mammalian testis-determining factor (Sry) (Gubbay et al., 1990; Sinclair et al., 1990). The binding domain of SOX2 high-mobility group (HMG) domain that Sry has is distinctive and binds DNA in a sequence-specific manner. SOX

proteins are proteins with HMG domains that are 50–70% identical to the HMG domain of Sry (Sry-related HMG box). Twenty distinct SOX genes have so far been identified in mice and humans (Schepers, Teasdale, & Koopman, 2002). Two SOX-like genes have been found in the unicellular choanoflagellate *Monosiga brevicollis*, indicating that the Sox proteins were first produced before multicellularity or that they may have served as a transitional marker between unicellular and multicellular organisms (Guth, Wegner, & sciences, 2008; King et al., 2008). The SOX factors are classified into several groups known as A through H with a sequence identity of more than 80%(Table 1)

S.NO	Group	SOX Member
1	SOX A	Sry
2	SOX B1	SOX1, SOX2 SOX3
3	SOX B2	SOX14, SOX21
4	SOX C	SOX4, SOX11, SOX12
5	SOX D	SOX5, SOX6, SOX13
6	SOX E	SOX8, SOX9, SOX10
7	SOX F	SOX7, SOX17, SOX18
8	SOX G	SOX15
9	SOX H	SOX30

**Table 1. Mammalian SOX2 factors and their subgroups**

HMG domain of SOX2 is composed of 79 amino acids and shows a preference for binding to the variant linear DNA sequence (A/T A/T CAAA/TG) in major groove (Laudet, Stehelin, & Clevers, 1993). The binding of L-shaped HMG domain of SOX2 forces the DNA to bend significantly (Lefebvre et al., 2007). The proteins of the SOX family contain a noncanonical HMG domain, evolved from the canonical HMG domain found in SRY (the sex determining gene on the Y chromosome). Although the identity



of the HMG domain of the SOX family and SRY can be as low as 50%, the ability to alter DNA conformation is conserved (Murphy et al., 1999). Within the proteins of SOX family, sequences are quite variable except within the HMG domain.

Many tissues and developmental processes have been shown to employ SOX proteins. For instance, SOX9 has a function in sex determination during the development of embryos by being expressed in the gonads of embryo (Kent, Wheatley, Andrews, Sinclair, & Koopman, 1996). In addition, SOX2, together with Oct4, was demonstrated to regulate FGF4 and osteopontin which play roles in early development (Botquin et al., 1998; Yuan, Corbi, Basilico, Dailey, & development, 1995). As for neural development, SOX2 has been defined as one of the earliest pan-neural markers and is known to maintain the multipotency of neural stem cell (Bylund, Andersson, Novitch, & Muhr, 2003; Uwanogho et al., 1995). On the other hand, SOX10 is expressed in the neural crest and contributes to peripheral nervous system development (Kuhlbrodt, Herbarth, Sock, Hermans-Borgmeyer, & Wegner, 1998). However, the neural crest expresses SOX10, which plays a role in the development of the peripheral nervous system (5). All members of SOXB1 subgroup (SOX 1-3) were shown to stimulate  $\delta$ -crystallin through binding to the DC5 enhancer (Kamachi, Uchikawa, Tanouchi, Sekido, & Kondoh, 2001). The differentiation of optic cup progenitors is also regulated by SOX2 and Pax6 (Matsushima, Heavner, & Pevny, 2011). In other tissues, SOX9 mutations also cause defects of skeletal structure in human (Südbeck, Schmitz, Baeuerle, & Scherer, 1996). Finally, the B-cells of SOX4 knockout mice are blocked in a pro-B-cell stage (Schilham et al., 1996b). The precise function of SOX2 is critical in early embryonic development. Mutations in the SOX2 gene cause anophthalmia, microphthalmia and anomalies in brain, pituitary, genitourinary problems and gastresophageal (Reis, Tyler, Schneider, Bardakjian, & Semina, 2010). Recently, SOX2 has been found to be related to several cancers as an oncogene, such as lung squamous cell carcinomas, glioblastoma, gastric carcinomas and breast cancer (Cui et al., 2018; Gangemi et al., 2009; Schaefer, Steiner, & Lengerke, 2020). These examples demonstrate the importance and functions of the proteins of SOX family in vertebrates.

### 1.9.1 SOX2 as a Transcription Factor

Hallmark of protein of SOX family is the occurrence of a high-mobility group (HMG) box which facilitates DNA binding and thereby allows them to act as transcription factors (Uchikawa, Kamachi, & Kondoh, 1999). One of the first regulatory target genes for transcription factors involved in developmental processes was *SOX2*. In 1995, significant discoveries were made. In their investigation of the activation of fibroblast growth factor 4 (*Fgf4*) in teratocarcinoma (and later embryonic stem (ES)) cell lines, Lisa Dailey and colleagues discovered that *SOX2* and OCT3, a synonym of OCT4 that the Mouse Genome Informatics Consortium renamed as POU5F1, collaborate to activate the *Fgf4* enhancer bearing their juxtaposed binding sites. (Yuan et al., 1995). It is clear that *SOX2* is involved in lens formation because Kamachi identified *SOX2* as the primary regulator of the  $\delta$ - and  $\gamma$ -crystallin genes that are uniquely expressed in the lens (Kamachi et al., 2001).

A specific DNA binding domain is bound by HMG box of *SOX2* in order to activate the transcription of the target genes, which is how SOX proteins serve as transcription factors (M. Wilson, Koopman, & development, 2002). Studies have made various attempts to map the transcriptional control of the family SOXB1. According to Kamachi et al., transcriptional activation activity in chicken *SOX2* is located at the C-terminus (Kamachi, Cheah, Kondoh, & Biology, 1999).

### **1.9.2. Role of *SOX2* in brain development:**

*SOX2* plays a critical function in the development of the central nervous system (CNS) and peripheral nervous system (PNS) by controlling the proliferation and differentiation of foetal progenitor cells (Pevny, Nicolis, & biology, 2010). In the CNS, the *SOX2* expression copies the actions of the other SOXB1 components, namely *SOX1* and *SOX3* (Bylund et al., 2003; Graham, Khudyakov, Ellis, & Pevny, 2003; Wood & Episkopou, 1999). The proliferation of CNS progenitor cells is generally boosted by the overexpression of any of these SOXB1 factors, whereas the onset of differentiation is triggered by their reduction (Bylund et al., 2003; Cavallaro et al., 2008; Ferri et al., 2004; Graham et al., 2003; Kishi et al., 2000).

Comparing an allelic series of *SOX2* hypomorphic mice with conditional null mice further showed that *SOX2* influence on retinal progenitor cells (RPCs) is, like that on ESCs, extremely dosage-dependent; RPCs without *SOX2* expression lose the competence to proliferate and differentiate, whereas reductions in *SOX2* levels induce

fluctuating microphthalmia. Surprisingly, *SOX2* expression has been shown to be critical for the differentiation of specific subsets of neurons, proving that the protein's role is not necessarily limited to the maintenance of progenitors and stem cells. For instance, adult olfactory bulb and neonatal cortical GABAergic interneurons in *SOX2* hypomorphic or knockout mice are attenuated (Cavallaro et al., 2008). Generally, beta-tubulin-positive, poorly arborized neuronal-like cells that lack markers for mature neurons and GABAergic neurons are produced by *SOX2* mutant Neuronal stem/progenitor cell (NPC) cultures (Cavallaro et al., 2008; Ferri et al., 2004). *SOX2* has been demonstrated to facilitate the development of migrating neural crest progenitor cells into sensory ganglia in an independent in vitro differentiation paradigm. (Cimadamore et al., 2012).

Furthermore, *MEIS1* gene and *SPRED2* gene were designated as the target genes of VISTA enhancer hs1210 due to the syntenic gene conservation in the enhancer region. Recent research has linked the proliferation of cells at the site of damage for neural healing to the downregulation of sproutly related protein 2 (*SPRED2*) in adult zebrafish brain. The *MEIS1* gene, along with other genes like the *TALE* genes, perform diverse functions in the differentiation of cells and the organogenesis process in the forebrain, are actively transcribed during the forebrain's developmental stages. Thus, it is conceivable that *SOX2* controls *MEIS1* and *SPRED2* expression in the developing and adult central nervous system.

## **1.10. Experimental approaches for Validation of Transcription Factor**

### **Binding Sites:**

Binding of TFs to DNA sequence has been identified by numerous experimental approaches. In the beginning different experiments were geared to validate proximal and core promoter elements by random cloning of respected region and the deletion mapping by the cell based reporter assays. The identification of DNA sequences having binding sites for transcription factors have been performed through EMSA (electrophoretic mobility shift assay) and DNase1 hypersensitivity method (Crawford et al., 2004). For the validation of binding sites of transcription factors, to which proteins bind, chromatin immune precipitation method is reliable. Another powerful and accurate method to find TFBSs (transcription factor binding sites) is chip-chip

approach (Visel et al., 2009). Below some experimental approaches are explained briefly.

### 1.10.1. ChIP-chip

Chromatin immune precipitations ChIP-chip, a powerful experimental approach, combining ChIP with DNA microarray. ChIP-chip method is used for the *in-vivo* determination of proteins and DNA interactions. This method in particular permits the identification of binding sites for proteins interacting with DNA in the whole genome. The primary objective of this method is to locate the binding sites for various proteins, which may than be helpful in locating the functional regions in other related genomes. ChIP-chip technique is subdivided into three steps. The 1<sup>st</sup> step is the designing of proper array and probe type. The 2<sup>nd</sup> steps comprise of wet-lab experimentation. The third step focuses on *in-silico* analysis of the obtained data. The major limiting factors are the size of DNA fragments and antibodies used along with the cost of DNA microarray (Buck & Lieb, 2004).

### 1.10.2. Chromatin Immunoprecipitation (ChIP-seq)

Chromatin immune-precipitation (ChIP), one of the main approaches to validate the genome wide transcriptome and protein-DNA interactions (Pepke, Wold, & Mortazavi, 2009). This method was testified in 2007, to *in-vivo* identify the binding sites for various proteins interacting with DNA like nucleosomes, chaperones, transcription factors, histones and DNA-binding enzymes (Bailey et al., 2013). The whole machinery of biological processes like cell cycle, DNA replication, cell differentiation, genes expression and chromosomes stability are dependent upon the various interactions of proteins with DNA. To allow exact genomic functional attempt ChIP and DNA sequencing technology are being used in combinatorial manner (Mundade, Ozer, Wei, Prabhu, & Lu, 2014). ChIP (Chromatin immuno precipitation) followed by DNA sequencing has one of the applications of Next Generation Sequencing (NGS) to differentiate protein-DNA interaction processes from the chemical modification of histone proteins. In this experiment, the proteins and DNA are cross linked. Then the cross-linked DNA is fragmented. Then to isolate it from the protein, a protein-specific antibody is used. The DNA sequence is then purified for sequencing after reversing the cross-links (Furey, 2012).

### 1.10.3. DNase Hypersensitivity

DNase hypersensitive (HS) sites, are specific regions in the nuclear chromatin, being highly prone to cleavage in the presence of degrading enzymes like DNase-1. Remodeling of chromatin structures is required to allow transcription of some silenced genes, sometimes needed in various cellular processes like cell differentiation. In these regions chromatin attains the less condensed state, making DNA accessible for transcription. Modifications in the chromatin topology while responding to transcriptional efficiency could be calculated as hyper-sensitivity of concerned fragment of DNA to digested with the enzyme DNase-I. The vulnerability is also known as the DNase-1 hypersensitivity (John et al., 2013). The identification of DNase-1 hypersensitive sites in the whole genome is a powerful technique to locate cis-regulatory elements or regions including enhancer, promoter, silencer, locus-control region and insulators. Therefore, these regions are used as markers for DNA regulatory regions (John et al., 2013).

#### **1.10.4. EMSA**

Electrophoretic Mobility Shift Assay is a widely used rapid approach to find interactions between protein and DNA. EMSA can be used for the characterization of various interactions including RNA/protein interaction and determination of stoichiometry and binding affinities. The basis of this technique is on the fact that the mobility of DNA-protein complexes is less than that of free oligonucleotides of nucleic acid sequences. The mixture of protein and nucleic acid is subjected to PAGE or agarose gel under required conditions and the gel is observed after electrophoresis. EMSA super shift can further be used by adding an antibody against the protein. It is a robust technique with broader binding conditions and compatibility for large range of sizes and structures of nucleic acids. The sensitivity of assay can be modified either high or low according to the requirement needed by using variants like chemiluminescence, radioisotope and fluorescence (L. M. Hellman & M. G. J. N. p. Fried, 2007).

#### **1.10.5. CUT & RUN**

The CUT & RUN (Cleavage Under Target & Release Using Nuclease) procedure is one of several used to map nucleosomes and chromatin accessibility for transcription regulatory factors (along with CHIP, ATAC, FAIRE etc.). This new approach has the

benefit of working in situ, on whole cells or nuclei, avoiding the crosslinking and fragmentation procedures that result in DNA pieces outside or far from the DNA-protein interaction sites, causing a lot of background sequencing noise. CUT & RUN enables targeting of histone modifications, transcriptional factors, and co-factors based on the specificity of the antibody (Sken and S Henikoff, 2017).

## **1.11. Molecular modelling theory and methods**

### **1.11.1. Molecular Docking**

The atomic-level interaction between a small molecule and a protein can be modelled using a computer method known as molecular docking. This technology enables us to characterize how tiny compounds react at the binding sites of target proteins and to comprehend fundamental biological processes (McConkey, Sobolev, & Edelman, 2002). Predicting the ligand structure, as well as its location and orientation inside these sites (known as pose), and determining the binding affinity are the two key components of the docking technique. Although each docking program operates slightly differently, they all share the same ligand and receptor, sampling, and scoring methods. Sampling involves positioning the ligand within the confines of receptor-site binding in terms of conformation and orientation. In order to rank the ligands, a scoring function chooses the optimum poses for ligand conformation, orientation and translation. Both the ligand structure (pose prediction) and its binding propensity must be correctly predicted for a docking exercise to be successful (affinity prediction). The main areas of variation among the docking program are the ligand placement in the "combining" site, the exploration of conformational space, and the score or binding estimate. Both the orientation of the side chains in the binding site and the fold of the protein backbone in that area are necessary for the interaction with the ligand. One of the main drawbacks of docking is that it is often conducted while the protein surface remains rigid, preventing evaluation of the impact of induced-fit inside the binding pocket.

### **1.11.2. ClusPro**

ClusPro (<https://cluspro.org>) is an internet docking server that allows two interacting proteins to dock directly. ClusPro first appeared in 2004 (Comeau, Gatchell, Vajda, & Camacho, 2004). However, it has been significantly updated and expanded

since then (Comeau et al., 2007; Kozakov et al., 2013). The server performs the following three computing steps: (i) rigid-body docking using billions of conformations as a sample; (ii) RMSD-based grouping of the 1,000 lowest-energy structures to identify the biggest clusters that represent the complex most likely models; and (iii) energy minimization refinement of selected structures.

### 1.11.3. PatchDock

PatchDock is a geometry-based molecular docking technique. Finding docking modifications that result in favorable complementarity in molecule shape is its main objective. When these changes are applied, they cause both significant steric collisions and sizable interface areas. A broad interface has the matched local features with complementary qualities of the coupled molecules (Duhovny, Nussinov, & Wolfson, 2002) The PatchDock method partitions the Connolly dot surface depiction of the molecules into concave, convex, and flat patches. In order to produce candidate modifications, complementary patches are matched next (Connolly, 1983). A scoring system that takes into account atomic desolvation energy and geometric fit is used to further assess each possible change (Zhang, Vasmatzis, Cornette, & DeLisi, 1997). After that, duplicated solutions are discarded by applying an RMSD (root mean square deviation) clustering to the candidate solutions.

### 1.11.4. FireDock

The first online server with an aside-chain optimization capability for refining protein-protein docking solutions is called Firedock (<http://bioinfo3d.cs.tau.ac.il/FireDock/>). It allows for the quick optimization of up to 1000 potential solutions. The approach tackles the flexibility issue while simultaneously evaluating the results of rapid rigid-body docking techniques. A list of refined complexes organized by binding energy function and a 3D visualization for viewing and contrasting refined complexes are included in the outcome (Andrusier, Nussinov, Wolfson, & Bioinformatics, 2007).

### 1.11.5. AutoDock

AutoDock is the first docking tool that model the ligand with full conformational freedom. The package comprises of the programmes AutoGrid and AutoDock, which are run in succession. In the beginning, AutoGrid is used to

determine the noncovalent energy of contact between the stiff portion of the receptor and a probe atom that is positioned at various grid positions of the matrix. Additionally, AutoGrid creates a desolvation map and an electrostatic potential grid map. AutoDock uses the flexible portion of the receptor and the whole set of grid maps to direct the docking of the chosen ligands (Morris et al., 2009).

#### 1.11.6. HADDOCK

HADDOCK, a web server, provides an integrative framework for modelling biomolecular complexes (<http://haddock.chem.uu.nl/Haddock>). It supports a large variety of input data and can deal multi-component assemblies of proteins, peptide, small molecules and nucleic acids. It accepts a wide range of input data and is capable of handling multi-component assemblies of proteins, peptides, tiny molecules, and nucleic acids. It can handle a wide range of experimental data (Van Zundert et al., 2016), supports nucleic acids and small compounds, and offers enhanced docking methods in the 2.0 edition of HADDOCK. Numerous issues, including as protein-protein, protein-nucleic acid, protein-oligonucleotides, and protein-small molecule complexes, have been addressed using HADDOCK. HADDOCK, in contrast to many other docking tools, permits conformational changes in both the side chains and the backbone of the molecules during complex formation. The docking of NMR structures and other Protein Data Bank (PDB) structures with different models is additionally directly supported by HADDOCK.

#### 1.12. Computer simulation

In addition to experimentation, computer simulation is a new technique for tackling scientific problems. One of the goals of computer simulation is to mimic experiments to light up the invisible microscopic details and thus explain the results. In parallel, simulations can be a useful tool to predict experimental results. Monte Carlo and molecular dynamics simulation are two frequently used techniques for simulating molecular systems. Based on stochastic methods that depend on probabilities, the Monte Carlo method is an easy-to-use methodology (Marcelli & Sadus, 1999). This technique generates numerous microstates or topologies of equilibrated systems moving from one microstate to the next in a specific statistical ensemble. Finally, the quantities are averaged over all produced microstates. Each arrangement, as well as its orientations and conformations, is subjected to random changes. There are multiple



advantages to using Monte Carlo simulation, but three stand out: simplicity, sampling flexibility, and the capacity to model various ensembles (Allen & Tildesley, 2012).

In order to determine the next configuration, molecular dynamics simulations compute equations of motion based on the force between atoms in an initial configuration (Allen & Tildesley, 2012). MD calculates atom migration by taking into consideration new locations, velocities, and orientations with regard to time. MD generates a set of configurations based on the starting setup and velocities. Several numerical integration algorithms can be used to calculate the equations of motion. There are two categories of MD simulations: one for non-equilibrium and the other one for equilibrium systems. Most of systems are simulated in the equilibrium state which is defined as an isolated system with a constant volume ( $V$ ) and fixed number of particles ( $N$ ). Since the system is isolated, the total energy  $E$  is constant. Therefore, by knowing  $E$ ,  $V$ , and  $N$  values of an isolated system, we can easily define its thermodynamic properties (Holian, 1995). The advantage of employing molecular dynamics over Monte Carlo simulation is that molecular dynamics, by calculating the ensemble average, analyses different qualities and values that the Monte Carlo technique cannot generally achieve (Leonhard & Deiters, 2002). As a result, the entire phase space is investigated. The molecular dynamics simulation has become essential in chemical, biological, and biophysical research because it can calculate various aspects of biomolecular systems that cannot be examined by experimentation. The programme Amber20 and classical molecular dynamics simulation were used in this study (Abbas Khan, Khan, et al., 2021). The FF14SB force field was used to mimic biomolecular systems (Suleman et al., 2021).

### 1.12.1. Classical molecular dynamics simulation

Newton's second law is the foundation of the molecular dynamics simulation approach. This technique assumes that every particle in the system operates like a Newtonian particle and entirely ignores quantum phenomena. This means that electronic motions are ignored, and electrons are supposed to stay in their ground state and quickly modify their dynamics when atomic locations change (the Born-Oppenheimer approximation). In fact, the motion of the particles is solely described by classical mechanics. Therefore, the equation of motion  $F = ma$ , where  $F$  is the force,  $m$

represents mass, and  $a$  is the acceleration, applies to the particles. The state of the system may be predicted and new locations and velocities can be determined once the positions and velocities of each atom are known. It is possible to carry out the process repeatedly until an atomic motion trajectory is produced.

### 1.13. Aims and Objectives

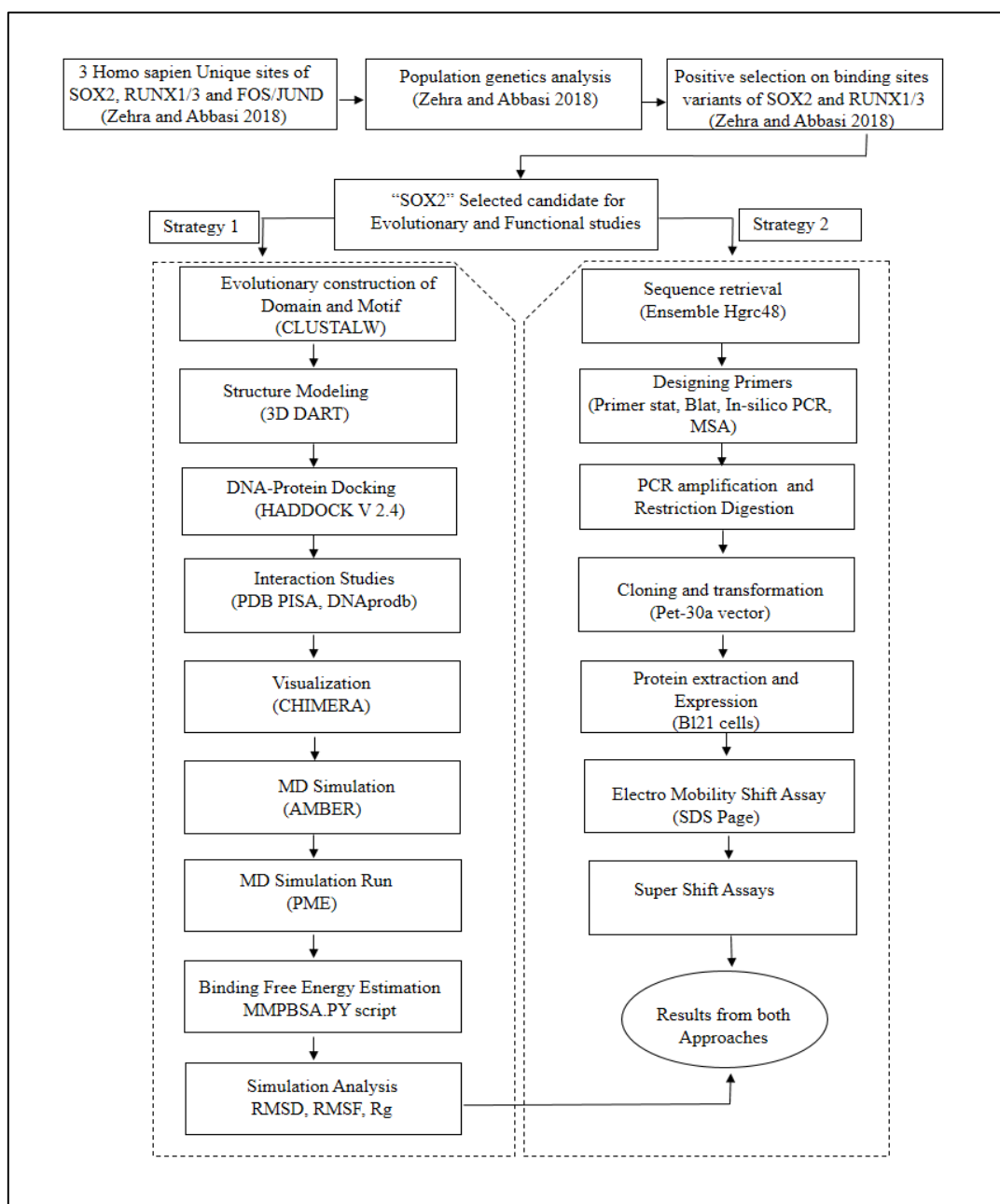
The aim of this research is to have a better comprehension of evolutionary changes controlling the expression of genes differences between humans (Modern human) and archaic hominins (Neanderthals and denisovans). More specifically, we aim to:

- Characterize the allelic variants of *SOX2* binding site within HAEs hs1210.
- Examine how structural analyses highlights significant conformational modifications and important residual contributions upon DNA-Protein binding at the atomic level.

**The study's objectives, which can be emphasized for this purpose, are**

1. Sequence acquisition of *SOX2* Protein and DNA sequences of human, archaic human, primates and non-primate mammals from Ensembl Genome Browser
2. Comparative analysis of *SOX2* Protein and DNA sequences of human, archaic human, primates and non-primate mammals through MSA.
3. Amplification of a gene encoding *SOX2* by utilizing gene specific primers.
4. Molecular cloning and overexpression of identified *SOX2* protein in bacterial BL21(DE3) E.coli expression system.
5. Characterization of identified *SOX2* protein by binding studies with DNA probes (oligonucleotides, carrying the ancestral T-allele (Neanderthal/Denisovan) and derived A-allele (fully modern Human)) using Electrophoretic Mobility Shift Assay and Supper-shift Assay.
6. In silico analysis for Protein-DNA interaction studies by Molecular Docking and Molecular dynamic simulation.

## 2. Materials and Methods



**Figure 6. Schematic diagram of steps carried out in work design**

*SOX2 TF was selected from Zehra and Abbasi 2018 for evolutionary and functional studies. In first strategy, evolutionary domain and motif was constructed then DNA-Protein docking was performed. Molecular dynamic simulation was performed to evaluate the dynamic properties of the protein-DNA complex. Second strategy includes Designing primers, PCR amplification of SOX2 by PCR and cloning into PET-30a (+) vector. Protein was expressed in E. coli BL21 cells. For in vitro analysis of enhancer hs1210 with SOX2 interaction, EMSA was performed and then conducted the super shift assay. Finally, results from both strategies were overlapped, analyzed and concluded.*

## 2.1. DNA Protocols

### 2.1.1. Genomic DNA Extraction from Human Blood

Extraction of Human genomic DNA (gDNA) was carried out using phenol chloroform method (Sanbrook, 1989). EDTA (ethylene-diamine-tetraacetic acid) tube used to collect human blood sample. Homogenization was achieved by inverting 5-8 times. Then same volumes (0.5 ml each) of Solution A (**Appendix 1**) and whole blood were added to labelled Eppendorf tube. To properly homogenize, the mixture was inverted thoroughly and kept for 15 minutes at the room temperature. Centrifugation of the mixture was carried out at 13000 revolutions per minute (rpm) for 01 minute. Supernatant discarded carefully. 0.5 ml of the Solution A was re-added to pellet and same centrifugation was carried out again. After discarding the supernatant, 0.4 ml of Solution B (**Appendix1**), 10 $\mu$ l of chilled Proteinase K and 10-12 $\mu$ l of 20% of SDS (sodium dodecyl sulphate) added to the eppendorf tube. Then proper vortexing, the tube was kept overnight in Redline oven (Binder, Germany) for incubation at 37 °C. After approximately 24 hours, 0.5 ml from fresh mixtures, having equal volumes of solution C and D (**Appendix 1**), were added to the tube. Proper mixing by inverting and centrifugation for 10 minutes at 13000 rpm. The supernatant was carefully collected in another Eppendorf tube and was added 0.5 ml solution D. Same centrifugation was done for 10 minutes. The supernatant containing DNA was carefully collected in a new labelled tube. DNA precipitation was achieved by adding 55 $\mu$ l of 3M Sodium Acetate and chilled 0.5 ml of absolute ethanol (100%). After inverting 4-5 times, the tube was centrifuged to pellet down DNA. After discarding the supernatant, washing of the DNA pellet with 70% ethanol was carried out to remove impurities. The DNA pellet was left open to dry after centrifuging at 13000 rpm and carefully removing any alcohol residue. To dissolve the DNA pellet, PCR water was added to the tube. DNA was quantified by using Scientific NanoDrop 1000 spectrophotometer at 260 nm wavelength and confirmed on 2% agarose gel.

## 2.2. PCR Protocols

### 2.2.1. Primer designing and dilution

*SOX2* gene, located on chromosome 3, has one transcript with sequence length of 2513 base pairs. Region from the start and end of coding sequences of *SOX2* protein (954 nucleotides, 317 amino acids) was retrieved from Ensembl genome browser for

designing primers. Specific primers for *SOX2* were designed manually by sequence analysis and checked by use of PCR primer stat ([https://www.bioinformatics.org/sms2/pcr\\_primer\\_stats.html](https://www.bioinformatics.org/sms2/pcr_primer_stats.html)) and multiple primer analyzer (Thermofisher Scientific). Specificity of the primers was confirmed by UCSC *in-silico* PCR, BLAST/BLAT at UCSC ([UCSC Genome Browser Home](#)) and NCBI ([National Center for Biotechnology Information \(nih.gov\)](#)) browser. Lipolised primers were diluted in Molecular grade water. Primers are given as follows.

*SOX2*-Forward primer: CATGATGGAGACGGAGCTG

*SOX2*-Reverse Primer: TGTGTGAGAGGGGCAGTGT

### 2.2.2. PCR Amplification

Standard procedure was followed for amplifying *SOX2* from human genomic DNA. PCR reaction of 25 $\mu$ l was prepared containing 2.5 $\mu$ l 10X PCR Buffer, 2 $\mu$ l 25 mM MgCl<sub>2</sub>, 2 $\mu$ l 10mM dNTPs, 1 $\mu$ l 10 $\mu$ M forward primer, 1 $\mu$ l 10 $\mu$ M reverse primer, 0.15 $\mu$ l 10units/ $\mu$ l Taq Polymerases, 1 $\mu$ l 90ng/ $\mu$ l human genomic DNA and finally the volume was raised to 25 $\mu$ l by the addition of PCR grade water in 0.2 ml tubes. For thoroughly mixing the reaction tube were centrifuge at 8000rpm for 30 second. PTC-200 DNA Engine cycler (Bio Rad, USA) was used for amplification. Cycling conditions of thermo cycler were programmed as initial denaturing for 5 minutes at the temperature 95°C then 37 cycles of amplification, each including three steps, denaturation for the 30 second at 95°C, while annealing of primers at temperature 60°C for 30 seconds, extension at the 72°C for 1 minute and a final extension at 72°C for the 20 minutes. The amplified PCR products were electrophoresed on 2% agarose gel and further purified using Kit (Pure-Link PCR Purification kit, Life Technologies). Purified PCR products of *SOX2* were quantified by the NanoDrop 1000 Spectrophotometer (Thermo Scientific) at 260nm wavelength and verified on agarose gel (2%).

### 2.2.3. Gel Electrophoresis

Gel agarose 0.6g was taken in the flask (250ml) containing 6ml 10X TBE (Tris/Borate/Ethylene-di-amine-tetra-acetic-acid) (**Appendix 2**) and volume was raised by addition of 54ml distilled water. To dissolve the components heated, it for 1 minute and then allowed it to cool down at the room temperature (RT), 6 $\mu$ l (5 mg/ml) of the substance ethidium bromide was added to the mixture after it had been heated for one minute and kept to cool at room temperature (RT) in order to dissolve the components.

The mixture was poured in tray and allowed to solidify. 3µl of 6X loading dye (Thermo Scientific) was mixed with PCR amplified product (5µl) and loaded in the well of agarose gel with micropipette. 2µl of 100 bp ladder (100ng/µl) (Thermo Scientific GeneRuler™ 100bp DNA ladder) was run in parallel to measure the size. The gel run was performed on 90 volts for 30 min in 1X buffer (TBE buffer). The gel was viewed under Ultra Violet Transilluminator (UV Trans-illuminator). Gel images taken by Dolphin gel documentation system (Wheeltech, USA).

#### **2.2.4. Purification of PCR Product**

PureLink PCR Purification Kit (Invitrogen, Germany) was used to purify the PCR product. 1µl of the purified PCR product run on 2% agarose gel electrophoresis and quantification done by the Thermo Scientific NanoDrop1000 Spectrophotometer at 260 nm wavelength.

#### **2.2.5. DNA Quantification**

PCR product was quantified by NanoDrop at 260 nm wavelength. The PCR product concentration was 346 ng/µl and 260/280 ratio was 1.82. The absorbance ratio at 280 nm is used for measuring the purity of DNA (Teare et al., 1997).

### **2.3. Cloning**

#### **2.3.1. Preparation of Media and Agar plates**

Luria-Bertani (LB) broth prepared by adding 5g of the yeast extract, 10g of bacto-tryptone and NaCl (10g) to 1 Litre of ddH<sub>2</sub>O and mixing until homogenized. 2ml of 2N-NaCl was added to adjust the pH to 7.5 Appendix 3). Prepared LB plates by adding 7.5g of agar to 500ml LB media and autoclaved (121 °C for 15 min). Media stored at 4 °C temperature.

#### **2.3.2. Preparation of the Competent Cells**

Cells of the DH5 strain of E. coli were cultured overnight at temperature 37°C on a non-antibiotic LB on agar plate. Colony inoculated into 5 ml of LB broth, shaken at 37 °C in shaking incubator for 18 hours at 250 rpm. The culture was added the following day to 100 ml of LB media in a 500 ml flask with 200 µl culture while being continuously shaken at 37 °C for 4 hours. After being placed into 50 ml falcon tubes,

the turbid media was incubated on ice for 30 minutes. It underwent a 10-minute centrifugation at 4 °C and 4000 rpm following the incubation period. After discarding the supernatant, 5ml of 0.1M CaCl<sub>2</sub> were added, gently mixed, and then chilled for an hour. It was then centrifuged at 4°C for 10 minutes at a speed of 4,000 rpm. 1.5 ml of 0.1M CaCl<sub>2</sub> was added, the supernatant was discarded, and centrifugation was performed as before. Gently mix two ml of 4:1, 0.1 mM CaCl<sub>2</sub>: 100% Glycerol after adding. Aliquots of 30 µl were prepared and kept at -80°C. They were streaked on LB agar plates containing 50mg/µl ampicillin antibiotic for the purpose of confirming the formation of competent cells, and colonies were not seen.

### 2.3.3. Restriction Digestion

Vector pET-30a (+) was used for cloning and expression in DH5α and BL21 cells. NdeI and XhoI enzymes were used to cut the vector. Reaction mixture of total 15µl for digestion was prepared by adding reagents as 7µl PET-30a, 2 µl Buffer, 0.5µl NdeI, 0.5µl XhoI and 5 µl water. It was incubated at temperature 37°C for 3 hrs. The PCR product of *SOX2* gene were digested by NdeI and XhoI enzymes. The resultant digested mixture of vector and DNA (PCR products) run on agarose gel (2%) at 80 volts for 40 min. Ultra Violet Transilluminator (UV Transilluminator) was used to view the gel. Gel images taken using Dolphin gel documentation system (Wheeltech, USA). The corresponding bands were excised and purified using gel purification kit and the concentration was measured by NanoDrop.

### 2.3.4. Ligation

Digested DNA (PCR product of *SOX2*) and vector were ligated in different ratios (1:1, 1:2, 1:3). Total Reaction mixture of 10µl was prepared by adding DNA (1 µl), Vector (1 µl), Buffer O (1 µl), T4 DNA ligase Enzyme (0.5 µl) and water (6.5 µl). After that, the mixture was incubated for an hour at 22°C.

### 2.3.5. Transformation

In chemically competent cells (*E. coli* DH5 for propagation and *E. coli* BL21 for expression), the vector pET-30a (+) and DNA ligation mixture were transformed. Competent cells were chilled on ice, 3 µl of ligate added into competent cells and was chilled on ice for 30 minutes for incubation. After 30 min, Cells were given heat shock for transformation at 42°C for 30 second in the water bath and then returned to the ice



for 2 minutes. 250 µl of media (**Appendix 3**) was added in cells and incubated at the 37°C for one hour at 195 rpm in shaker (Innova 43). 70µl of transformation mixture was streak in agar plate of LB containing 100µg/ml kanamycin and incubated overnight at 37°C in red Line oven (Binder, Germany). After 24 hours incubation, one colony picked from the plate and added in a falcon tube containing LB media with 100µg/ml kanamycin. The culture was incubated overnight at temperature 37°C with a continuous shaking of 190rpm in shaking incubator. After the completion of the incubation period, 1.5ml culture from every culture tube was proceeded for isolation of the bacterial plasmid.

### **2.3.6. Plasmid Extraction and purification**

The Plasmid was extracted manually. The cultured colonies were centrifuged at the 8000 rpm for 1 minute and the pellet was obtained. A pipette was used to entirely remove the LB medium. Furthermore, the pellet was re-suspended in 300 µl of P-1 Buffer (Appendix 4). It was then given 300 µl of newly prepared P-2 Buffer (Appendix 4), gently mixed, and incubated for 5 minutes at room temperature. 300 µl of P-3 buffer (Appendix 4) were pipetted out before 5 minutes had passed, and the mixture was gently agitated. In the meanwhile, 1.5ml Eppendorf tubes containing 800 µl of isopropanol were placed at -20°C. After centrifuging the bacterial lysate mixture at maximum speed for 10 minutes, proteins and cell debris were collected as a pellet. The supernatant was then transferred to cold isopropanol and spun at maximum speed for 15 minutes. To completely eliminate all traces of isopropanol, the pellets were washed with 500 µl of 70% ethanol. After centrifugation, pellets were collected and dried. The elution was made in 30 µl of pre-warmed TE Buffer. The purified plasmid was confirmed on agarose gel and then quantified by use of Thermo Scientific NanoDrop 1000 Spectrophotometer (USA).

### **2.3.7. Restriction digestion**

Plasmid was digested by Xho1 and Nde1 for the conformation of cloning by preparing a total mixture of 15µl containing plasmid 2µl, Xho1 0.5µl, Nde1 0.5µl, Buffer 1.5µl and water 10.5µl then was incubated at temperature 37°C for three hrs. Mixture was run on agarose gel at 90 volts for 35 minutes. Gel was observed in Ultra

Violet Trans-illuminator (UV trans-illuminator) and pictures were captured using Dolphin gel documentation (Wheeltech, USA).

## 2.4. Protein Analysis

### 2.4.1. Protein expression

After cloning the gene of *SOX2* into DH5 $\alpha$  cell, protein expression was done. The recombinant proteins were expressed in chemically competent *E. coli* BL21 cell. The cultures were prepared by adding a single colony of BL21 cells in LB-medium containing antibiotic kanamycin and incubated at 20°C for 6 hours. 100mM IPTG (**Appendix 5**) was added in to the culture after 6 hours and incubation done overnight at 20°C on 190 rpm. After 24 hours, the culture was transferred to eppendorf tube and centrifuge for six min at 6000 rpm. Supernatant discarded and protein pellet was extracted. 200 $\mu$ l of Protein loading dye (**Appendix 6**) added to the pellet and dissolved by incubating at 95°C for 10 min.

### 2.4.2. SDS PAGE Analysis

Denaturing Sodium dodecyl-sulfate polyacrylamide gel electrophoresis (SDS-PAGE) was carried out according to the Invitrogen NuPAGE® specifications (**Appendix 7**). In brief, 32ml of Separating gel was made by adding water 6.7ml, 30% acrylamide 12.8ml, 1M tris (PH 8.8) 12ml, 20% SDS 0.16ml, 10% APS 0.32ml and TEMED 0.032ml. It was poured in a gel tray and allowed to solidify. After that 10ml of stacking gel was prepared by adding water 6.59ml, 30% acrylamide 2ml, 1M tris (PH 6.8) 1.25ml, 20% SDS 0.05ml, 10% APS 0.1ml and TEMED 0.01ml. It was poured over solidified separating gel and the comb was placed over it to make wells. After solidification the wells were filled with buffer and the samples were loaded along with control sample. Protein ladder was loaded parallel to samples. The gel was run for three hours at 120 volts. After that the gel was removed and fixed in fixative solution for 15 minutes. It was then placed in Coomassie dye solution, heated for 40 seconds and left for 15 minutes. It was washed with water and left overnight in water to let the bands appear.

### 2.4.3. EMSA and Super-shift Assay

Following the oligonucleotides and their reverse complimentary strands synthesized, labeled with biotin, annealed, and used as probes for EMSA analysis: oligonucleotides (derived A-allele containing probe: 5'-GCTTAGACAACAATGGATAAAGAG-3' and 5'-CGAATCTGTTGTTACCTATTTCTC-3'; ancestral T-allele containing probe: 5'-TAGCTTAGACTAACAATGGATAAAG -3' and 5'-ATCGAATCTGATGTTACCTATTTTC-3'), carrying the substitutions in enhancer region of human and neanderthal respectively (L. M. Hellman & M. G. Fried, 2007). EMSA carried out with Gel Shift Kit (Viagene). Briefly, the purified SOX2 was incubated with the double-stranded probes (20 fmol) for 20 minutes at room temperature in a reaction mixture contained 100 mM Tris (pH 7.5), 500 mM KCl, and 10 mM DTT. The same conditions were used for the competitive binding assay, but 100 times more unlabeled double-stranded oligonucleotides were added. To perform the supershift experiment, 1.0 mg of anti-SOX2 antibody was added and incubated for 30 minutes on ice. DNA-protein complexes were resolved on 6.0% (wt/vol) native polyacrylamide gels (Bio-Rad), transferred to Biodyne nylon membranes (Pierce), viewed under a UV transilluminator, and their images were documented using Dolphin gel documentation (Wheeltech, USA).

## 2.5. In Silico Analysis

### 2.5.1. Collection of sequences and comparative analysis

SOX2 Protein and DNA sequences of human (*Homo sapiens*), primates (*Pan troglodytes*, *Gorilla beringei*, *Pongo abelii*, *Mus musculus*, *Rattus norvegicus*), non-primate mammals (*Loxodonta Africana*, *Oryctolagus cuniculus*, *Equus caballus*, *Monodelphis domestica*, *Callithrix jacchus*, *Lama pacos*, *Canis lupus familiaris*, *Felis catus*, *Ornithorhynchus anatinus*, *Galago*) obtained from Ensemble genome browser and the orthologous sequences of archaic human extracted from the Neanderthal Ensembl Genome Browser (<http://neandertal.ensemblgenomes.org/index.html>) were subjected to multiple sequence alignment (MSA) through ClustalW (Fernández &

Birney, 2010; Larkin et al., 2007). The resultant MSA was analyzed to determine the conserved segments.

### 2.5.2. DNA and Protein modelling

3D structural model of DNA was generated from enhancer sequence carrying the derived A-allele (5'TTAGACA\*ACAATGGATA 3') and ancestral T-allele (5'TTAGACT\*ACAATGGATA 3') by use of 3D-DART (3DNA-Driven DNA Analysis and Rebuilding Tool) provided by High Ambiguity Driven protein-protein DOCKing (HADDOCK) webserver (<http://haddock.science.uu.nl/services/3DDART/>) followed by energy minimization. In order to facilitate the structural analysis of DNA in association with proteins, it gave the desired sequence's ideal B-DNA structure (van Dijk & Bonvin, 2009). The crystal structures of the human HMG domain (39-121 AA) of SOX2 (PDB ID: 1O4X) were retrieved through Protein data bank PDB (D. C. Williams, Cai, & Clore, 2004). 3D structures were examined using UCSF Chimera (Version 1.11.2) extendable molecular modeling system package (Version 1.11.2) (Pettersen et al., 2004)

## 2.6. DNA-protein docking and Complexes refinement

### 2.6.1. HADDOCK

HADDOCK version 2.2, employs an extensible docking approach based on the biophysical and biochemical association statistics from already predicated protein interface. This data is used in form of Ambiguous Interaction Restraints (AIRs) that are ambiguous distance between all the residues that are involved in interactions (Dominguez, Boelens, & Bonvin, 2003). Before running the HADDOCK, the surface residues were identified for CPROT analysis. CPROT is an algorithm for predicating the surface residues of protein (de Vries & Bonvin, 2011). The CPROT predicated active and passive residues were used in the HADDOCK using the easy or predication interface. Generally, 1000 structures are docked using HADDOCK's rigid body minimization (it0) mode, and the top 200 are refined using semiflexible refinement in torsion angle space (it1) before being refined using an explicit solvent (most favorable cluster is listed first) (Van Zundert et al., 2016).

HADDOCK scores each model using Equation-1, where  $E_{AIR}$ ,  $E_{elec}$ ,  $E_{vdw}$  and  $E_{desolv}$  are the AIR restraints, electrostatic, van der waals and desolvation energies, respectively. BSA is buried surface area and  $E_{data}$  encompasses the energy of other restricted data.

$$E = 0.01E_{vdw} + 1.0E_{elec} + 1.0E_{desolv} + 0.01E_{air} - 0.01 BSA + 0.1 E_{data} \quad (1)$$

The selected models are subjected to Simi flexible refinement followed by the water refinement step in the torsion angle space and explicit water shell, respectively and scored by the Equation 2 and 3, respectively. RMSD values and cluster ranks are rendered to the average score of the top 4 structures for each cluster.

$$E = 0.1 E_{vdw} + 1.0 E_{elec} + 1.0 E_{desolv} + 0.1 E_{air} - 0.01 BSA + 0.1 E_{data} \quad (2)$$

$$E = 0.02 E_{elec} + 0.1 E_{vdw} + 1.0 E_{desolv} + 0.01 E_{air} + 0.1 E_{data} \quad (3)$$

To study interaction in between the amino acids in SOX2, Human and Neanderthal DNA directly, DNA binding domain of SOX2 was docked onto either A allele or T allele. The initial complex was further refined by the maximum likelihood method in REFMAC5 (Murshudov et al., 2011).

### 2.6.2. UCSF Chimera

For visualization of docked complexes UCSF Chimera 1.11 was used (<http://www.cgl.uscf.edu/chimera/>). It generates high resolution images for comparative analysis. UCSF Chimera is an extremely extensible program and offer 3D visualization of molecular structure and related data including density maps, super molecular assemblies, trajectories, energy minimization and conformational assemblies (Pettersen et al., 2004)

### 2.6.3. Interaction Analysis

For interaction (hydrogen bonding and hydrophobic interactions) analysis of SOX2 with derived DNA and SOX2 with ancestral DNA structures, PDBsum, DNAproDB and PDBe-PISA were used.

#### 2.6.3.1. PDBsum

**PDBsum** is a web-based server providing all the information of the structure that is deposited in PDB. It includes the images of the structure, annotated plots of each

protein and schematic diagram of the protein-protein, protein-ligand, and the protein-DNA interaction. PDBsum is reorganized whenever any new structure is out by the PDB. In the larger protein-DNA complexes the LigPlus is unable to show the interaction between protein and DNA, and then PDBsum is used. PDBsum homepage include a generate link that is used for the analysis of user provided newly modeled structure (Laskowski, Jabłońska, Pravda, Vařeková, & Thornton, 2018).

### 2.6.3.2. DNAproDB

DNAproDB is a web-based visualization tool that makes structural analysis of DNA-protein complexes easy (Sagendorf, Berman, & Rohs, 2017). Herein, we used DNAproDB to visualize our docked complexes and understand the interaction pattern. Hydrogen bond were analyse by PDBe-PISA via taking default criteria (Krissinel, 2010).

## 2.7. Molecular dynamics simulation

MD simulation is remarkably advantageous, though computationally expensive tool for bimolecular and chemical systems analysis. The dynamic behavior of both the complexes i.e. derived A-allele containing DNA (fully modern human)-SOX2 complex and the ancestral T-allele containing DNA (Neanderthal/Denisovans)-SOX2 checked by MD simulation performed on Amber20 using OL15 force field. (Salomon-Ferrer, Case, & Walker, 2013). In a TIP3P water box, system solvation was conducted, and the system was neutralized by the addition of counter ions. (Price & Brooks III, 2004). The bad clashes were removed from the system using an energy-minimization process. Steepest descent algorithm (Meza, 2010), and conjugate gradient algorithm used for 6000 and 3000 cycles (Watowich, Meyer, Hagstrom, & Josephs, 1988). After heating to 300 K, the system cooled to equilibrium at 1 atm of constant pressure, with a weak restraint, and without any restraint. The production phase ran for 100 ns. The long-range electrostatic interaction treated with Particle Mesh Ewald (PME) algorithm (Salomon-Ferrer et al., 2013), with cutoff distance of 10.0 Å. The SHAKE algorithm used to treat covalent bond (Kräutler, Van Gunsteren, & Hünenberger, 2001). Finally, production step of MD simulation executed on the PMEMD.CUDA and processing of trajectories using the Amber20 CPPTRAJ package(Roe & Cheatham III, 2013).

### 2.7.1. Binding free energy calculations

The MMGBSA approach was used to estimate the real binding energy calculations for both complexes. Same method is the best methodology used by different studies to estimate real binding energy of the different biological complexes, like protein–protein, Spike protein–ligand and protein–DNA/RNA (Abbas Khan et al., 2021; M. T. Khan et al., 2020; ul Qamar et al., 2019). MMGBSA.py (Hou, Wang, Li, & Wang, 2011), script used to estimate total binding free energy of the top ligand complexes. Energy term like electrostatic, vdW, GB, and SA calculated as part of total binding energy. For the free energy calculation, following equation used:

$$\Delta G (\text{bind}) = \Delta G (\text{complex}) - [\Delta G (\text{receptor}) + \Delta G (\text{ligand})]$$

**Every component of total free energy estimated by using following equation:**

$$G = G_{\text{bond}} + G_{\text{ele}} + G_{\text{vdW}} + G_{\text{pol}} + G_{\text{npol}} - TS$$

Where  $G_{\text{vdW}}$ ,  $G_{\text{bond}}$  and  $G_{\text{ele}}$ , indicate van der Waals interactions, bonded, electrostatic respectively. The  $G_{\text{pol}}$  and  $G_{\text{npol}}$  are polar and nonpolar solvated free energies.  $G_{\text{pol}}$  and  $G_{\text{npol}}$  are calculated by generalized born (GB) implicit solvent method with solvent accessible surface area SASA term.

### 3. Results

#### 3.1. Comparative sequence and functional analysis of human accelerated enhancer hs1210 and domain organization of *SOX2*

Over the years, accelerated regions in the human genome have been known to harbor hundreds of cis-regulatory elements. These cis-regulatory element, as enhancers, have proven to perform a significant role in spatio-temporal expression of many developmental genes. In our previous study, Zahra and Abbasi identified selection signatures on transcription factor binding site (TFBS) modifying derived A-allele of *SOX2* TF in one such accelerated human brain enhancer (hs1210). The TFBS modifying derived A-allele was also found to be *Homo sapien*-specific when compared to non-human primates and archaic human (Neanderthals and Denisovans) data that carried the ancestral T-allele in the TFBS of *SOX2* (Figure 7A). Furthermore, Vesil et al., in 2007 reported the hs1210 expressed the reporter gene exclusively in the forebrain of transgenic mice, more specifically in lateral ganglionic eminence (LGE), a transient structure in the developing telencephalon (Figure 7C).

Domain annotation was performed for an insight into comparative domain organization of *SOX2*. Annotation disseminates the distinguished architecture of *SOX2* gene which is comprised of High Mobility Group domain (HMG) comprises 79 (amino acid) residues and contains of 3  $\alpha$  helices and N-terminal  $\beta$  strand that were arranged in twisted L-shape. Homeobox domain (DNA binding domain) bind DNA via highly conserved helix-turn-helix (HTH) motif structure. Motif incorporate 2 alpha helices, that make close contact with the DNA and joint by short turn. First helix of the motif helps to stabilize the structure and second helix bind to the DNA by a number of hydrogen bonds and the hydrophobic interactions, which occurs between the specific side chain and the exposed bases and thymine methyl groups within major groove of DNA. Comparative sequence analysis showed that the DNA binding HMG box of *SOX2* (*SOX2*<sup>(HMG)</sup>) is highly conserved among human, archaic human, non-human primates and mammals (Figure 7B)

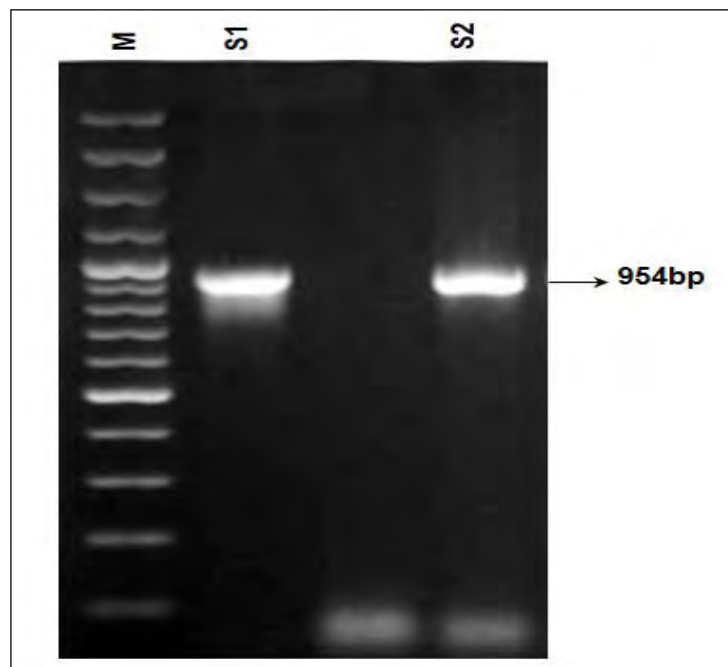




### 3.2. In vitro binding analysis of *SOX2* to target DNA containing ancestral and derived alleles

#### 3.2.1. *SOX2* Amplification

Genomic DNA isolated from fresh blood samples by phenol chloroform method and concentration were determined with spectrophotometer readings. Concentration was found to be 496 ng/ul and to check the integrity of DNA, Gel electrophoresis was performed which showed intact high molecular DNA band. PCR amplification was performed with genomic DNA. Results showed the presence of 954 nucleotides bands of *SOX2* along with 100bp DNA ladder. The PCR product were purified using PureLink PCR Purification Kit (Invitrogen, Germany). Purified PCR product was confirmed and analysed on agarose gel (2%). Figure 8.



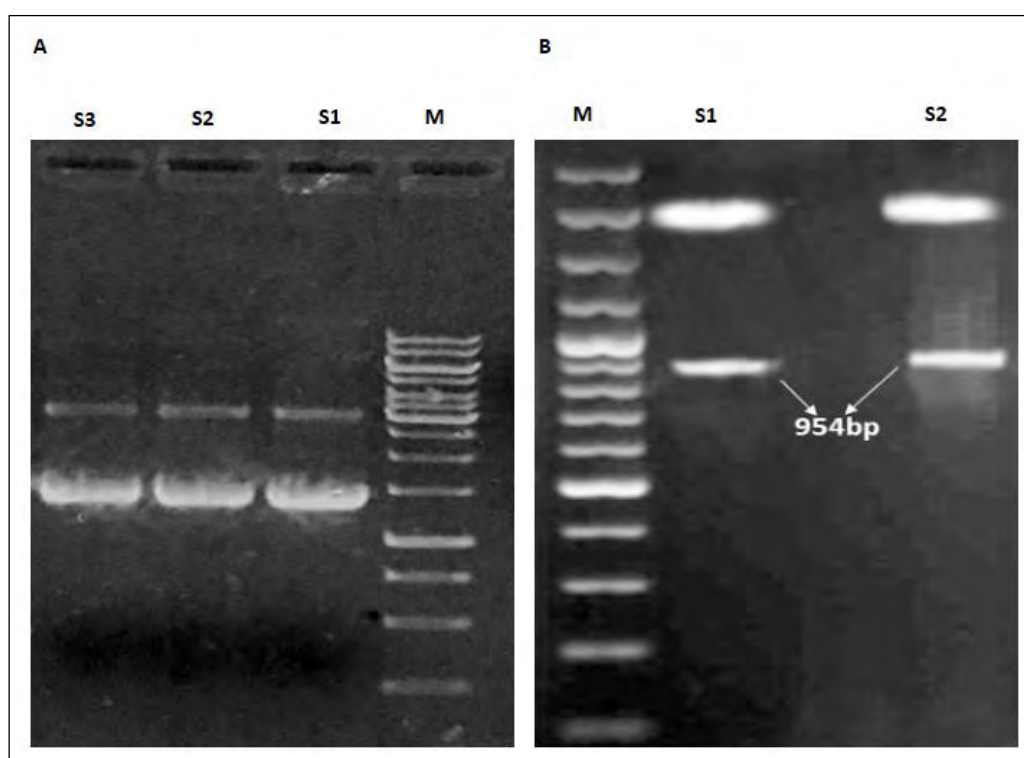
**Figure 8. Electropherogram of PCR products of *SOX2***

2% agarose gel stained by ethidium bromide shows PCR products of *SOX2*. M represents 100bp Molecular marker and S represents sample.

#### 3.2.2. Cloning and conformation of *SOX2*

To clone the PCR amplified *SOX2* DNA, gel bands were purified and also ligated into predigested Pet-30a (+) vector for propagation followed by the

transformation into the DH5 $\alpha$  cells. Clone was screened for presence of the insert by colony PCR and the positive clone selected. Plasmids were isolated from positive clones by Miniprep plasmid extraction kit and subjected to restriction digestion by Nde1 and Xho1 enzymes. Double bands shown in figure 9A, given below represents successful isolation of plasmids. On restriction digestion linearized vector with PCR fragments of *SOX2* was obtained shown in figure 9B.



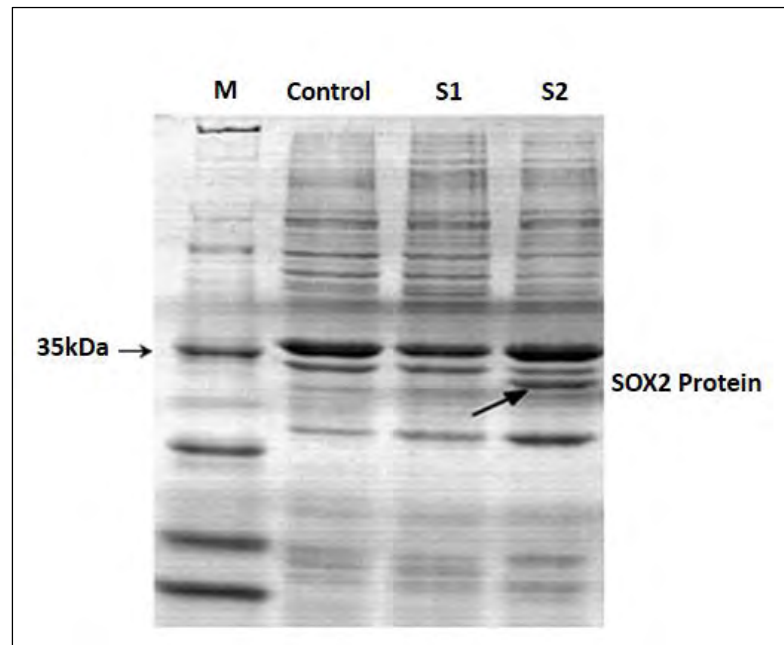
**Figure 9. Electropherogram of circular and digested plasmid.**

(A) Purified circular plasmids of *SOX2* cloned in a vector. (B) PCR products of *SOX2* obtained by restriction digestion in S1 and S2, S: sample; M: 100bp Molecular marker.

### 3.2.3. Expression, purification and conformation of *SOX2* Protein

After successful cloning, proteins were expressed in B121 cells. After IPTG induction in LB media, the verified clone was incubated, and cultures were sampled both before and after induction. The samples were prepared, and in order to assess the cellular proteins, SDS-PAGE was performed. In comparison to the uninduced sample, it was observed that the recombinant protein of the predicted size was overexpressed. After subjected to the SDS-PAGE, corresponding respective bands of the *SOX2* protein

was visualized along with marker at 35 kDa as shown in figure 10. Different treatment parameters were employed to increase the SOX2 protein's presence in the soluble fraction, including varying the temperature, IPTG concentration, and ethanol concentration. The ideal conditions for this protein expression in the soluble fraction were found to be 18 °C, 0.4 mM IPTG, and 3% ethanol.



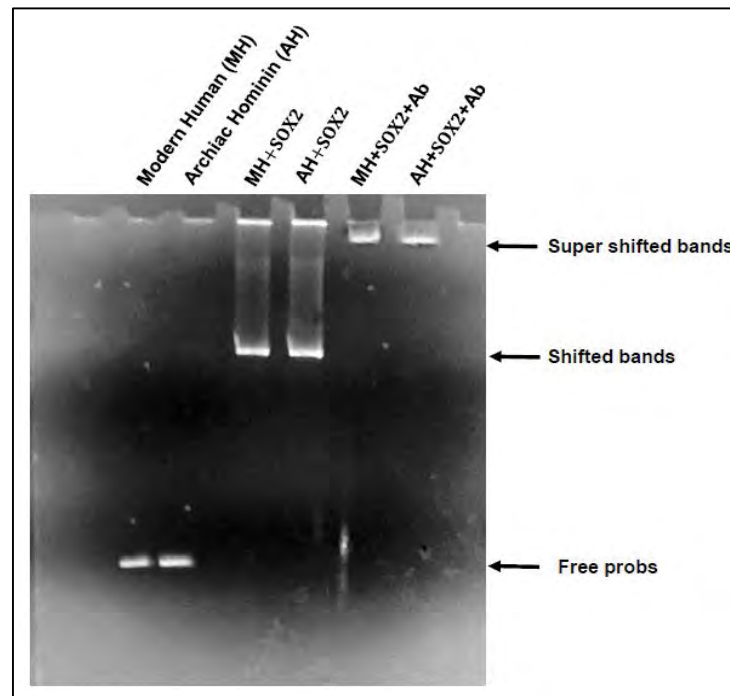
**Figure 10. Expression of SOX2 protein**

*SDS-PAGE of SOX2 protein. M: protein marker (kDa); control: B121 cells; S: samples.*

#### **3.2.4. EMSA and Super shift**

EMSA were performed using purified SOX2 protein and two de novo pairs of complementary oligonucleotides, carrying the ancestral T-allele (Neanderthal/Denisovan) and derived A-allele (fully modern Human). An EMSA binding buffer was co-incubated with purified proteins and DNA probes in order to promote the binding of purified proteins with DNA probes. The DNA-protein complexes are loaded and run on a non-denaturing polyacrylamide gel. We found that both ancestral and derived alleles containing oligonucleotides were capable of binding to the purified SOX2 protein and as a result, it moves through a polyacrylamide gel more slowly than the corresponding free, unbound DNA. Addition of antibodies directed against SOX2 caused further retardation (EMSA supershift protocol) within the gel and thus confirmed that the bound protein in these complexes is SOX2 (Figure 11). Therefore, based on EMSA protocol it can be suggested that SOX2 is capable of

binding with the derived (TAGACA\*ACAATGGAT) as well as the ancestral (TAGACT\*ACAATGGAT) versions of its target DNA sites.



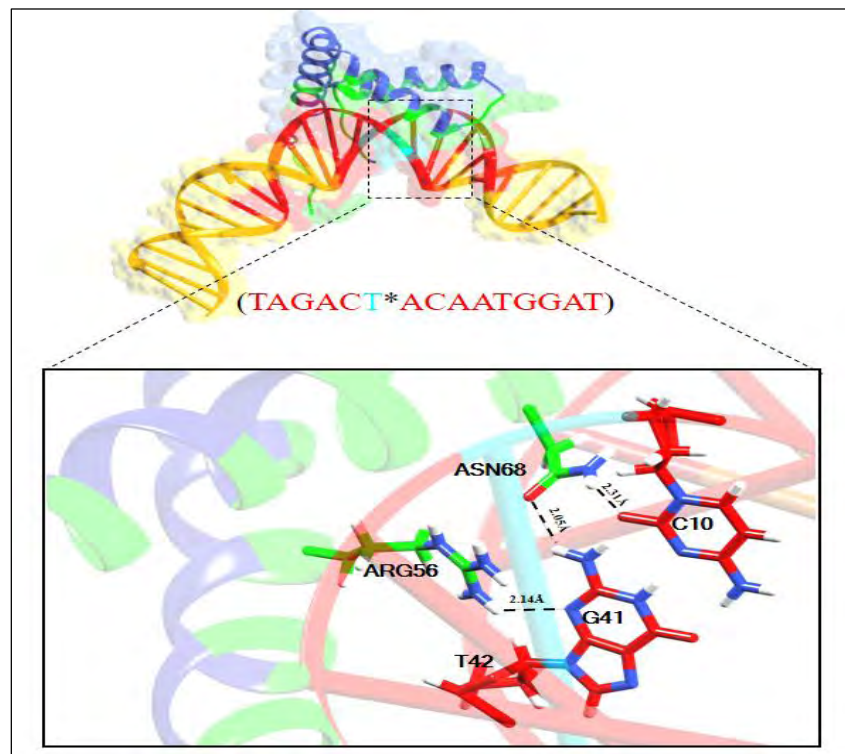
**Figure 11. EMSA/Gel shift assay**

Electrophoretic mobility shift assay shows shift in the mobility of SOX2 protein-DNA complexes as compared to the free probes (Modern Human and Archaic Hominin). Binding of SOX2 protein hinders the mobility of DNA probe (shifted bands) and addition of antibody to SOX2 bound DNA probe further reduced the mobility of complex in the gel (super shifted bands).

### 3.3. Molecular docking characterization of the protein-DNA complex

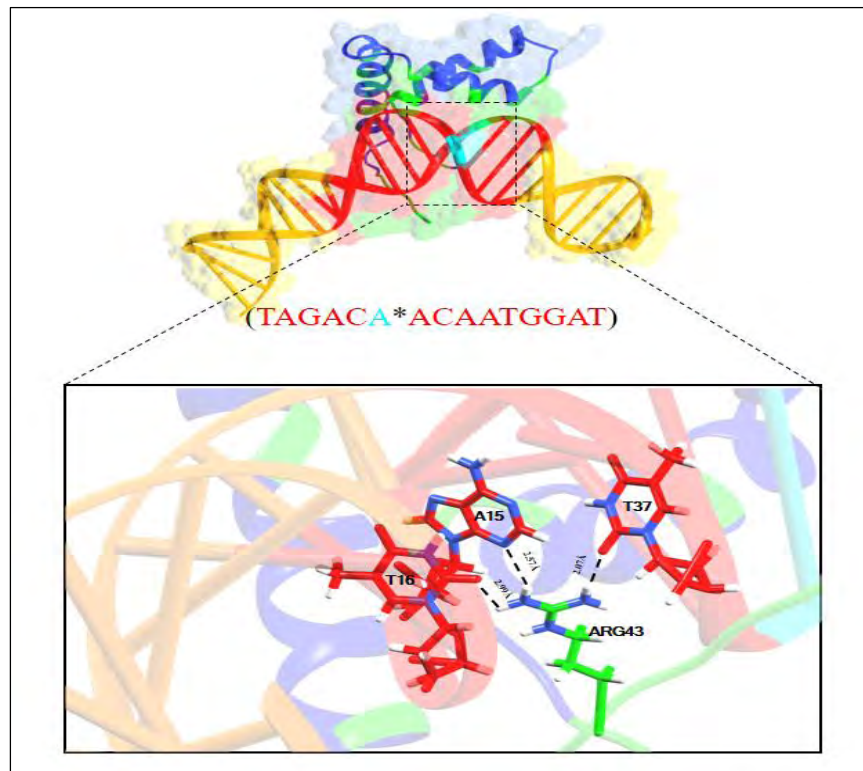
Molecular docking was performed to obtain an atomic level understanding of conformational alterations in protein-DNA complex upon the binding of HMG box to ancestral (Neanderthal/Denisovans) and derived (fully modern human) target sites. The molecular docking results corroborates well with *in vitro* data and revealed the binding of SOX2<sup>(HMG)</sup> with both target sites carrying ancestral T-allele or derived A-allele. For instance, in corroboration with previously reported experimental data our molecular docking results indicate that the HMG box of SOX2 protein (SOX2<sup>(HMG)</sup>) grips directly the *major* groove of the double helix DNA structure for both ancestral T-allele and derived A-allele containing target sites (Figures 12 & 13) (Reményi et al., 2003; Scaffidi & Bianchi, 2001). ARG56 and ASN68 of SOX2<sup>(HMG)</sup> formed HBs with C10

and G41 of the ancestral T-allele (Figure 12), while ARG43 of SOX2<sup>(HMG)</sup> made contacts with A15 and T37 of derived A-allele (Figure 13). Thus, our docking results showed that SOX2<sup>(HMG)</sup> forms a reliable and energetically more favored contact with the derived A-allele containing DNA than it does with the ancestral T-allele containing DNA.



**Figure 12. Structural analysis of Ancestral T-allele DNA- SOX2<sup>(HMG)</sup> complex**

*Topological model for SOX2<sup>(HMG)</sup> binding to ancestral T-allele DNA shown as semi-transparent surface and ribbons. The zoomed image illustrates the interface between the residues of SOX2<sup>(HMG)</sup> and corresponding nucleotides. Black dotted lines with calculated distances in angstroms (Å) represent hydrogen bonding.*



**Figure 13. Structural analysis of Derived <sup>A-allele</sup> DNA- SOX2 <sup>(HMG)</sup> complex**  
*Topological model for SOX2 <sup>(HMG)</sup> binding to derived <sup>A-allele</sup> DNA shown as semi transparent surface and ribbons. The zoomed image illustrates the interface between the residues of SOX2 <sup>(HMG)</sup> and corresponding nucleotides. Black dotted lines with calculated distances in angstroms (Å) represent hydrogen bonding.*

HADDOCK score is measured by the integration of four terms: Van der Waals energy (weight 0.1), electrostatic energy (weight 0.2), restraint violation energy (weight 0.1) and desolvation energy (weight 1.0) (De Vries, Van Dijk, & Bonvin, 2010). The more negative HADDOCK and Z-scores indicates a reliable interaction. Z-score is the quantitative measures of cluster standard from the average score. However, careful analysis has revealed the notable conformational and energetic differences between the two complexes Table 2.

**Table 2. Molecular docking based energetic profile evaluation through HADDOCK**

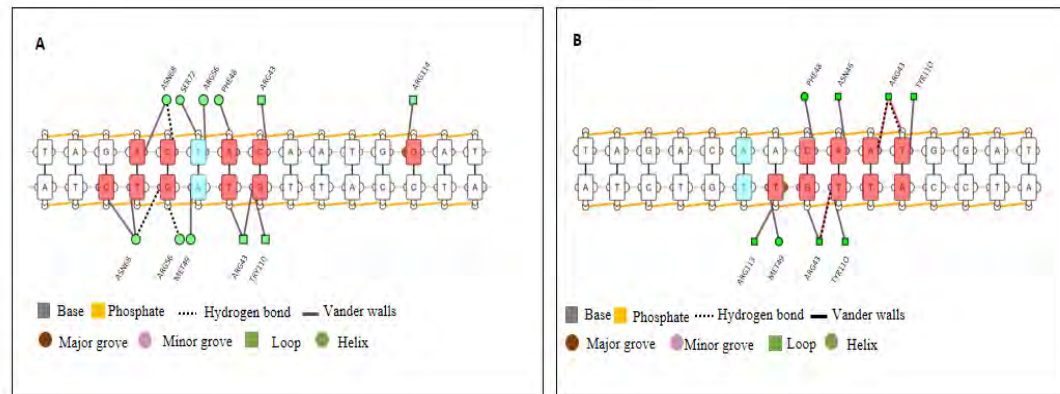
Docked complex	HADDOCK score	Z-score	Van der Waals energy	Electrostatic energy	Desolvation energy	Restraints violation energy	Buried surface area in Å <sup>2</sup>
Derived A-allele DNA-SOX2	-281.6 +/- 4.2	0	-97.7 +/- 2.5	-1061.0 +/- 36.3	28.0 +/- 3.2	2.3 +/- 0.23	2622.1 +/- 39.6
Ancestral T-allele DNA-SOX2	-270.3 +/- 4.2	0	-110.6 +/- 2.7	-892.6 +/- 21.5	18.7 +/- 3.7	1.7 +/- 0.31	2868.9 +/- 27.4

*All the energies are calculated as kcal/mol. The Haddock score is lower for derived <sup>A-allele</sup> DNA-SOX2 complex (-281.6 +/- 4.2kcal/mol) as compared to ancestral <sup>T-allele</sup> DNA-SOX2 complex (-270.3 +/- 4.0kcal/mol).*

### 3.3.1. Structure comparison between SOX2 DBDs bound to Human and Neanderthal DNA sequences

Despite the high sequence homology of SOX2, their molecular structures were distinct when they were bound to similar target DNA sites with a single nucleotide change. Complexes' protein-DNA connections studied by DNAProDB revealed a complicated pattern of interactions. However, nucleotides with which the SOX2<sup>(HMG)</sup> interacts directly are slightly different for ancestral T-allele containing target DNA site (5'-GACT\*AC-3') and derived A-allele (fully modern human-specific) containing target DNA site (5'-ACAAT-3') (Figures. 14 A and B). Intriguingly, the ancestral T-allele was involved in direct interaction with SOX2<sup>(HMG)</sup>, however its mutant version in fully modern humans, i.e., derived A-allele did not interact directly with SOX2<sup>(HMG)</sup>.





**Figure 14. (A-B) Structural analysis of DNA in complex with Ancestral <sup>T-allele</sup> SOX2 (HMG) and Derived <sup>A-allele</sup> SOX2 (HMG) complexes**

*Schematic diagram shows interactions between the Ancestral <sup>T-allele</sup> SOX2 (HMG) and Derived <sup>A-allele</sup> SOX2 (HMG) DNA backbone as a nucleotide-residue interaction map for (A) Neanderthal and (B) Human. Backbone contacts display considerable differences between both complexes. Nucleotides (red) are interacting with amino acid residues (green) of SOX2 (HMG) through hydrogen bonds (dotted lines) and Vander walls (solid lines). The DNA strands are displayed as orange. Cyan color shows the position of nucleotide variant in SOX2 binding site. Images were obtained from DNAProDB.*

Furthermore, noticeably different types of amino acid residues and secondary structural elements (SSEs) of HMG box were involved in interactions with ancestral and derived target sites (Table 3). It appears that fully modern humans-specific nucleotide substitution has changed the binding conformation and location of target DNA site for HMG box.

**Table 3. DNA binding residues of SOX2 (HMG) reported previously (experimentally determined) and in the present study**

Sr. No	Remenyi et.al. 2003 DNA-HMG box		Present study A-allele DNA- HMG box		Present study T-allele DNA- HMG box	
	Interacting residues	SSEs	Interacting residues	SSEs	Interacting residues	SSEs
1	ARG43	L1	ARG43	L1	ARG43	L1
2	ASN46	L2	ASN46	L1	-	-
3	ARG53	H1	-	L1	-	-
4	ASN68	H2	ASN68	H2	ASN68	H2
5	SER69	H2	-	-	SER69	H2
6	SER72	H2	-	-	-	-
7	LYS73	H2	LYS73	H2	-	-
8	TYR110	L3	-	-	TYR110	L3
9	ARG111	L3	-	-	-	-
10	ARG113	L3	ARG113	L3	-	-
11	ARG114	L3	ARG114	L3	ARG114	L3
12	LYS115	L3	LYS115	L3	-	-
13	-	-	ALA47	H1	-	-
14	-	-	ARG56	H1	ARG56	H1
15	-	-	-	-	PHE48	H1
16	-	-	-	-	MET49	H1

*The table highlights the differential binding of HMG box of SOX2 with A-allele containing DNA and T-allele containing DNA in terms of amino acid residues and SSEs involved in docked complex. Experimentally determined amino acid residues and SSEs are also given for the reference purpose (Remenyi et.al. 2003). SSEs, Secondary Structure Elements; L, Loop; H, Helix*

Hydrogen bonds (HBs) are known to determine the strength of intermolecular interactions (D. C. Williams et al., 2004). Therefore, docked complexes of SOX2<sup>(HMG)</sup> with ancestral and derived alleles carrying DNA target sites were analyzed for HBs pattern by employing PDBe-PISA (Krissinel, 2010). We empirically observed twelve HBs when SOX2<sup>(HMG)</sup> bind to the derived A-allele carrying DNA (Table 4). In contrast, the ancestral T-allele DNA-SOX2<sup>(HMG)</sup> complex involves only 7 HBs (Table 4). Thus, our docking results showed that SOX2<sup>(HMG)</sup> forms a reliable and energetically more favored contact with the derived A-allele containing DNA than it does with the ancestral T-allele containing DNA.

**Table 4. Hydrogen bond interactions between DNA and HMG box of SOX2 protein determined through molecular docking experiments**

S.No.	Derived A-allele DNA				Ancestral T-allele DNA			
	Interacting residues of SOX2	D...A Distance Å	D-H...A Distance Å	Interacting nucleotides	Interacting residues of SOX2	D...A Distance Å	D-H...A Distance Å	Interacting nucleotides
1	ARG43NH1	3.02	2.64	DT37O2	ARG43NH1	3.68	2.02	DT39O2
2	ARG43NH1	2.95	2.09	DG38O4	ARG43NH1	3.5	2.18	DT39O4
3	ARG43NH2	3.07	1.94	DT16O2	ARG43NH2	2.97	2.14	DA14O3
4	ARG43NH2	2.95	2.32	DT16O4	ARG56NH1	3.75	2.31	DT42O4
5	ASN46ND2	3.8	2.13	DA15O4	ASN68ND2	3	2.21	DC10O2
6	ALA47N	3.68	2.34	DA14O3	ASN68ND2	3.74	2.31	DC10O4
7	ARG56NH1	2.88	2.07	DT40O3	ASN68OD1	2.43	2.05	DG41H21
8	ASN68ND2	2.97	2.47	DG41O3	-	-	-	-
9	LYS73HZ1	2.15	2.22	DA11O3	-	-	-	-
10	ARG113N	3.01	2.09	DT37O3	-	-	-	-
11	ARG114NH2	3	1.85	DG18O3	-	-	-	-
12	LYS115HZ2	2.26	1.76	DA9O5	-	-	-	-

*Amino acid numbering is based on position of HMG box domain (39-115) within SOX2 protein. The sequence of bases in one strand of DNA (chain b) are numbered from 1-25 and in the other strand (chain a) are numbered as 26-50. D (A, C, T, G) denotes Deoxyribonucleotides, D...A, denotes distances between donor atom and acceptor atom while D-H...A illustrates distance between the hydrogen bonded to donor atom and acceptor atom.*

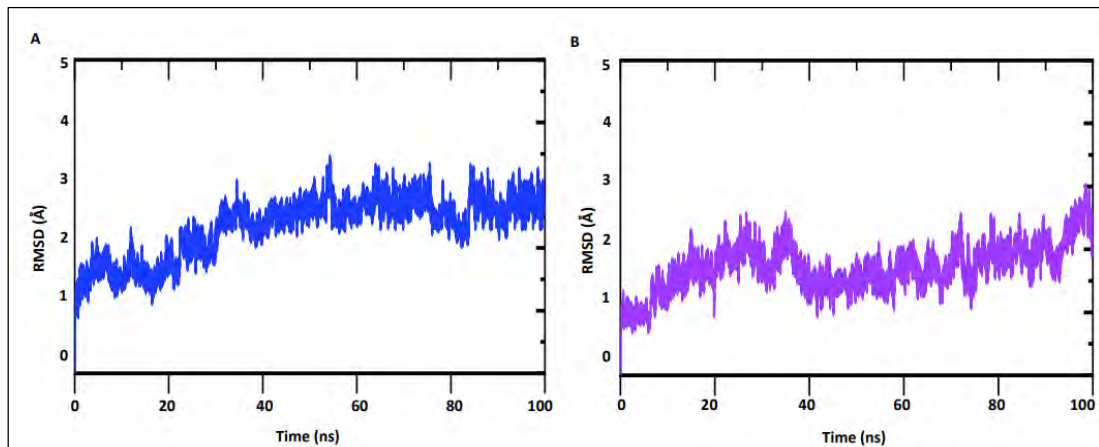
### 3.4. Evaluation of dynamical properties of the protein-DNA complex

To get a deeper insight into the dynamic behavior and binding differences induced by the fully modern human-specific evolutionary substitution (T>A), structural stability and flexibility of DNA-protein complexes and the corresponding binding free energies were measured at 100 ns trajectories through molecular dynamics (MD) simulation methods. After the completion of simulation, PDB files at different intervals were generated to observe the conformational changes in DNA upon binding to SOX2 HMG. Resulting trajectories were analyzed intensely to determine stability, convergence and structural changes during simulation.

The atom positional root-mean-square deviation (RMSD) is a standard method for comparing the similarity/difference of two molecular structures. It is primarily practiced for quantifying the variance between the backbone of a protein from static structure to living protein over the simulation time (0, t) (Cohen & Sternberg, 1980).

Smaller deviation means the structure is more stable and may not exhibit the significant deviation from original static structure. Herein, thermodynamics based conformational stability of each complex was evaluated via RMSD calculations. Trajectories from simulation were used to calculate the structural-dynamics features. Overall it can be seen that both the complexes i.e. derived A-allele containing DNA (fully modern human)-SOX2 complex and the ancestral T-allele containing DNA (Neanderthal/Denisovans)-SOX2 complex exhibit almost similar average RMSD values.

However, compared to the ancestral T-allele containing DNA (Neanderthal/Denisovans)-SOX2 complex, derived A-allele containing DNA (fully modern human)-SOX2 complex possesses more dynamic stability. For instance, the average RMSD for the derived A-allele containing SOX2 complex was 1.8Å and the structure did not deviate significantly over the simulation time, but a minor level of conformational deviations from the original static complex structure were observed between 70-80 ns (ns: nanosecond, that is one billionth of a second). The average RMSD for the ancestral T-allele containing DNA (Neanderthal/Denisovans)-SOX2 complex was comparable with the derived A-allele containing DNA (fully modern human)-SOX2 complex. However, over the simulation time of 0.00 ns to 100.00 ns the ancestral T-allele containing DNA (Neanderthal/Denisovans)-SOX2 complex revealed abrupt conformational fluctuations in RMSD values which is suggestive of extreme structural perturbation and relatively weaker and unstable intermolecular interactions (Figure 15). These data suggest that fully modern human-specific single nucleotide substitution within brain exclusive human accelerated enhancer (BE-HAE) hs1210 might have evolved conformationally more stable and efficient interaction between SOX2<sup>(HMG)</sup> and target DNA binding site.

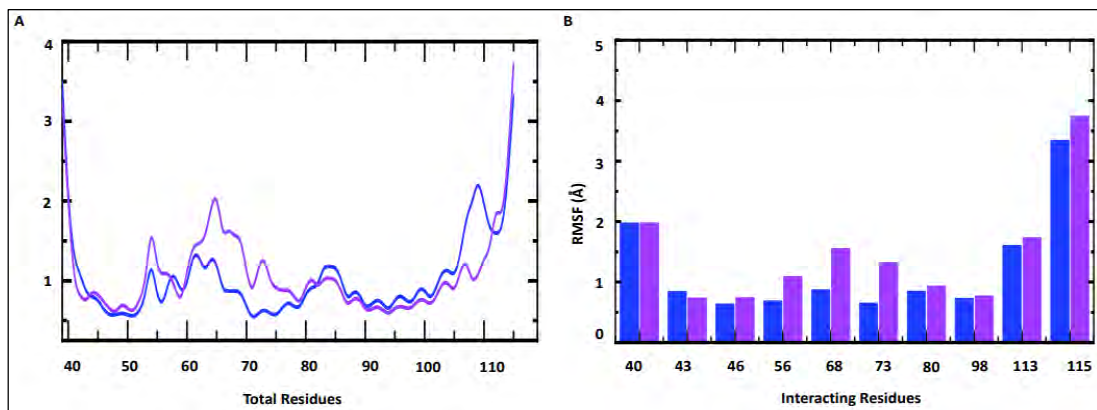


**Figure 15. Dynamic stability of DNA-protein complexes along the course of 100 ns**

### Simulation

(A-B) The RMSD graphs of both complexes throughout the simulation explaining their stability and equilibration nature. In comparison with Ancestral T-allele DNA-SOX2 (violet) complex (B), Derived A-allele DNA-SOX2 (blue) complex (A) appears to be well stabilized. The x-axis shows time in nanoseconds while y-axis show RMSD in Å. Violet color depict ancestral T-allele DNA-SOX2 complex whereas blue color depict derived A-allele DNA-SOX2 complex.

To obtain information on local flexibility and thermal stability of the protein, root mean-square fluctuations (RMSF) are often calculated from molecular dynamics simulations (Cooper, 1976). In context of the macromolecular interactions, the higher RMSF value indicates a more flexible and thus unstable interactions (Joshi, Joshi, Sharma, Chandra, & Pande, 2021). In contrast, the smaller RMSF values correspond to minimal atomic movements about their average positions during the simulation and hence depict the stable macromolecular interactions. RMSF plot in figure 16 exhibits almost a similar trend of residue fluctuation profile for both ancestral and derived alleles based DNA-protein complexes with an average RMSF of 2.5 Å. However, closer inspection revealed that the amino acid residues 65-95 of SOX2<sup>(HMG)</sup> are more stable in complex with fully modern human DNA target site (carrying derived A-allele) when compared to SOX2 complex with DNA target site carrying ancestral T-allele (Figure 16A). Here, we also measured the RMSF with respect to C $\alpha$  atom of each interacting residue of HMG box and a plot of RMSF was employed to depict the fluctuations for both ancestral and derived alleles carrying protein-DNA complexes. Figure 16B shows that each interacting amino acid residue of HMG box in derived A-allele containing complex is more stabilized.

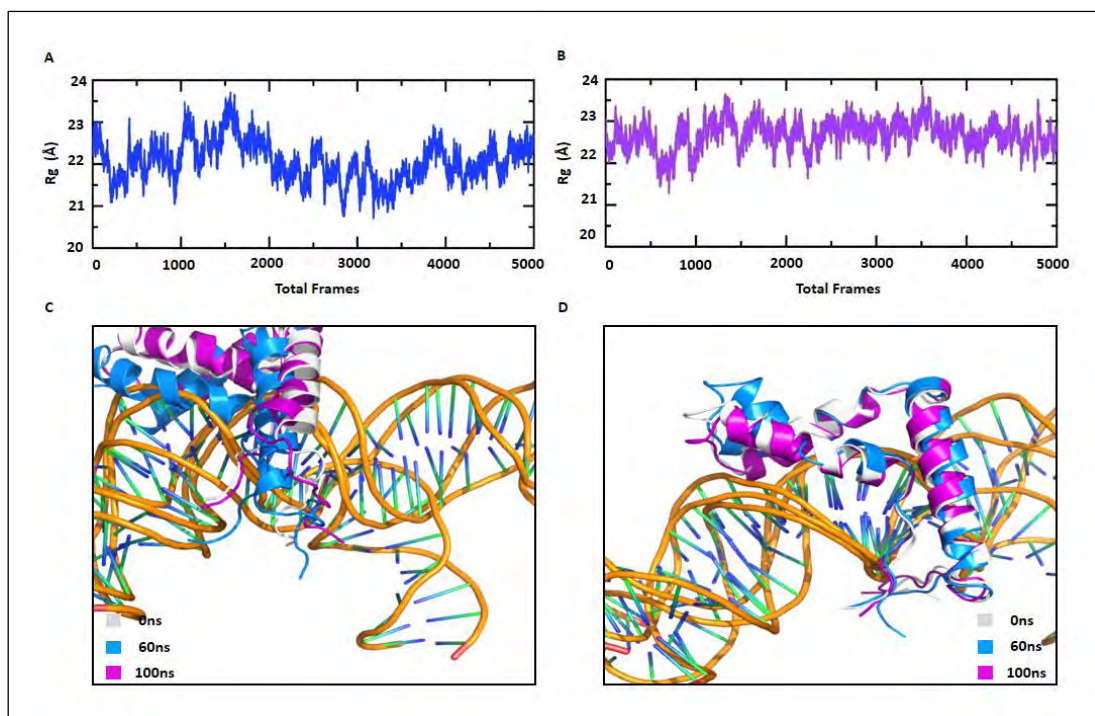


**Figure 16. Residual flexibility of DNA-protein complexes along the course of 100 ns Simulation**

(A) RMSF plots for each trajectory file. The x-axis shows total number of residues while y-axis show RMSF in Å. (B) RMSF plots for interacting residues of both complexes. x-axis shows the interacting residues while y-axis show RMSF in Å. Violet color depict ancestral T-allele DNA-SOX2 complex whereas blue color depict derived A-allele DNA-SOX2 complex.

The molecular spatial packing of amino acid residues is an important determinant of protein stability. From the analysis of different protein types, it has been shown that variations in protein compactness is determined by an intricate combination of the size, secondary structure of proteins, and relative composition of interacting macromolecules (Lobanov, Bogatyreva, & Galzitskaya, 2008; Tsai, Taylor, Chothia, & Gerstein, 1999). A compact packing of amino acid residues is known to affect the stability of macromolecular assemblies (Seeliger & De Groot, 2010). Therefore, we have used the MD simulations to calculate the Radius of gyration ( $R_g$ ) as function of simulation time, which is a measure to estimate the protein structure compactness (Lobanov et al., 2008). Ancestral T-allele containing DNA (Neanderthal/Denisovans)-SOX2 and derived A-allele containing DNA (fully modern human)-SOX2 complexes possess substantial differences in the pattern of  $R_g$  (Figure 17A and Figure 17B). For instance, the average  $R_g$  value for derived A-allele based complex was 22.0Å, whereas the average  $R_g$  value for ancestral T-allele based complex was observed to be 23.0Å. The lower  $R_g$  value suggests the tightest and most stable packing of derived A-allele based DNA-SOX2 complex compared to ancestral T-allele based DNA-protein complex (Figure 17A and Figure 17B).

Temporal aspects of structural stability of protein-DNA complex were investigated by alignment of PDB structures at different time points. These data revealed that at simulation time of 60ns the HMG box physically moved inside the major groove and thereby favored tightly packed binding with derived A-allele containing DNA target site, whereas at the simulation time of 0 ns and 100 ns the loose interaction was observed between HMG box and derived DNA (Figure 17C). However, at all set time points (0 ns, 60 ns and 100 ns), HMG box failed to tightly intercalate into the major groove of ancestral T-allele containing DNA (Figure 17D).



**Figure 17. Radius of gyration (RoG) Analysis**

*A-B RoG plot calculated for derived A-allele DNA-SOX2 (blue) and ancestral T-allele DNA-SOX2 (violet) complexes during the 100 ns simulation. x-axis shows total number of the frames while y-axis show Rg in Å. C-D Structural superposition of derived A-allele DNA-SOX2 and ancestral T-allele DNA-SOX2 PDBs at different time points. Grey, blue and magenta colors represent 0ns, 60ns and 100 ns.*

To provide further insights into the binding affinities of HMG box for ancestral and derived alleles containing target sites, hydrogen bonding (HB) differences between the two complexes were evaluated through CPPTRAJ module in Amber simulation package by using 5000 structural frames obtained from MD simulation during the 0.00

ns to 100.00 ns. An average PDB from each trajectory was obtained and checked for total number of HBs in each complex, that include both inter- and intramolecular HBs. During the simulation time, significant rearrangements in intermolecular HBs were observed between the two complexes (Table 5). In case of derived A-allele containing DNA, 12 hydrogen bonds were observed between HMG box and target DNA site before the MD simulation (0.00 ns) whereas during the MD simulation the HB network readjustments were seen with formation of three extra bonds with the DNA molecule through Arg40, Lys42 and Arg98 residues of HMG box. It can be seen that in terms of intermolecular HB networks Arginine residues of HMG box are major contributors in making links with derived A-allele containing DNA. Only 7 intermolecular HBs were observed between HMG box and ancestral T-allele containing DNA (Neanderthal/Denisovans) before the MD simulation while during the MD simulation hydrogen bonds number was increased to 12, with extra contributions from Arg98 and Arg113 residues of the HMG box (Table 5). These differences in intermolecular HB patterns clearly shows that HMG box of SOX2 interacts more robustly with derived A-allele containing DNA complex than to ancestral T-allele containing DNA.



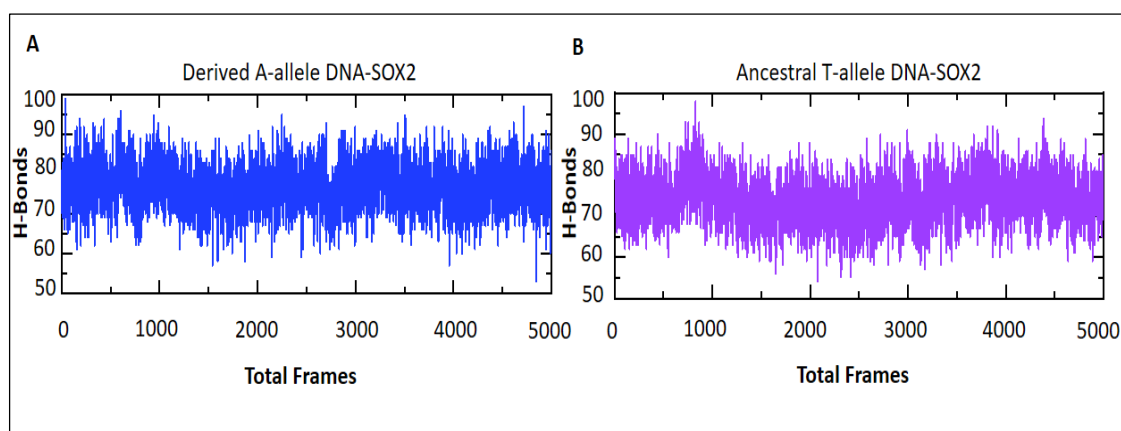
**Table 5. Hydrogen bonds between the SOX2 protein and DNA before and after MD simulation.**

Complex name	Before simulation				After simulation		
	Index	SOX2	Dist. [Å]	Human DNA	SOX2	Dist. [Å]	Human DNA
<b>Derived A-allele DNA-SOX2</b>	1	ARG43	3.02	DT	ARG40	1.81	DG
	2	ARG43	2.95	DG	LYS42	1.81	DT
	3	ARG43	2.95	DT	ARG43	1.93	DT
	4	ARG43	3.07	DT	ARG43	1.76	DT
	5	ASN46	3.8	DA	ARG43	2.03	DG
	6	ALA47	3.68	DA	ASN46	2.98	DA
	7	ARG56	2.88	DT	ARG56	1.73	DG
	8	ASN68	2.97	DG	ASN68	2.2	DT
	9	LYS73	2.15	DA	LYS73	1.83	DC
	10	ARG13	3.01	DT	LYS80	1.75	DC
	11	ARG14	3	DG	ARG98	2.11	DT
	12	LYS115	2.26	DA	ARG98	1.69	DT
	13	-	-	-	ARG113	2.28	DT
	14	-	-	-	ARG113	2.32	DG
	15	-	-	-	ARG113	1.76	DT
<b>Ancestral T-allele DNA-SOX2</b>	Before simulation				After simulation		
	Index	SOX2	Dist. [Å]	Neanderthal DNA	SOX2	Dist. [Å]	Neanderthal DNA
	1	ARG43	3.5	DT	ARG40	1.93	DT
	2	ARG43	3.68	DT	LYS42	1.9	DG
	3	ARG43	2.97	DA	ARG43	2.06	DT
	4	ARG56	3.75	DT	ASN68	1.73	DT
	5	ASN68	3.74	DC	LYS73	2.3	DC
	6	ASN68	3	DC	LYS73	1.93	DT
	7	ASN68	2.43	DG	LYS87	1.83	DC
	8	-	-	-	ARG98	1.75	DA
	9	-	-	-	ARG98	1.63	DA
	10	-	-	-	ARG113	2.26	DG
	11	-	-	-	ARG113	2.09	DT
12	-	-	-	ARG113	1.82	DT	

*The table depicts inter-molecular hydrogen bonding differences between the two complexes based on 5000 structural frames obtained from molecular dynamics (MD) simulation during the 0.00 ns to 100.00 ns. Dist. Å (angstrom) illustrates distance between the hydrogen bonded donor atom and acceptor atom or length of hydrogen bond. D (A, C, T, G) denotes Deoxyribonucleotides.*

To further evaluate the overall strength of protein-DNA interaction, the total number of HBs (both inter- and intramolecular hydrogen bonds) within each complex were evaluated during the 0.00 ns to 100.00 ns simulation (Figure 18A and Figure 18B). In total 85 HBs were detected for derived A-allele containing DNA-HMG complex, whereas 75 hydrogen bonds were detected for ancestral T-allele containing DNA-HMG complex (Figure 18A and Figure 18B). These results further validate the enhanced

interaction of HMG box of SOX2 with fully modern human-specific substitution carrying target DNA site through conformational changes in protein-DNA complex.



**Figure 18. MD simulation based Inter- and Intramolecular hydrogen bonds Analysis.**

*A-B Hydrogen bond plots for derived A-allele DNA-SOX2 (blue) and ancestral T-allele DNA-SOX2 (violet) complexes during the simulation (100 ns). X-axis shows total number of the frames while Y-axis show H-bonds count.*

For each complex (ancestral and derived) binding free energy was evaluated by using 5000 structural frames obtained from MD simulation. For this purpose Molecular Mechanics/Generalized Born Surface Area (MM/GBSA) method was employed that combines molecular mechanics calculations and continuum solvation models (Hou et al., 2011). Detailed comparisons of the energetic profiles of two complexes is given in Table 6. Taken together, the total free energy of binding for the derived A-allele containing DNA (fully modern human)-SOX2<sup>(HMG)</sup> complex is more favorable (-140.20 kcal/mol) than the total free energy for the ancestral T-allele containing DNA (Neanderthal/Denisovans)-SOX2 complex (-100.38 kcal/mol) (Table 6). These energetic profiles based on Molecular dynamics simulation analysis are consistent with the protein-DNA docking results which also suggest the tighter binding of SOX2 with the fully modern human specific derived A-allele containing DNA (Table 2).

**Table 6. Free energy estimates based on Molecular Mechanics/Generalized Born Surface Area (MM/GBSA).**

Complex Name	VdW	Electrostatics energy	GB	ESURF	Total binding energy
<b>Derived A-allele DNA-SOX2</b>	-143.98	-9817.95	9840.11	-18.36	<b>-140.20</b>
<b>Ancestral T-allele DNA-SOX2</b>	-94.68	-9022.54	9030.37	-13.52	<b>-100.38</b>

*All the energies are calculated as kcal/mol. vdW; van der Waals, GB; Generalized Born model, ESURF; non-polar contribution to solvation energy using SASA (solvent accessible surface area) for GB.*

#### 4. Discussion

An evolutionary analysis of gene regulation is vital to understanding the molecular basis of organismal phenotypic diversity (Abbasi, 2011; Koshikawa, 2020). Transcription of bilaterian protein-coding gene bodies is governed by non-coding DNA sequences known as enhancers (also referred as cis-regulatory elements) that are typically located upstream or downstream of the coding sequence, but may be within an intron or at distal position (Anwar et al., 2015). Transcription factors (TFs) bind with enhancers to set out the timing, level, and spatial expression of genes (Ali et al., 2021). Mutations in enhancers regions have been shown to produce an altered pattern of gene transcription (Kvon, Waymack, Gad, & Wunderlich, 2021). Therefore, divergence in enhancer sequence and functionality is considered to be an important determinant of inter- and intra-species phenotypic diversity (Long, Prescott, & Wysocka, 2016). For example, reporter assays in transgenic model animals have been used to validate the roles of human-specific mutations in altered enhancer activity and expression pattern differences between chimpanzee versus human (Glinsky & Barakat, 2019). Furthermore, polymorphisms within non-coding putative enhancer regions have been associated with phenotypic variability among living human populations (Doan et al., 2016). A key challenge in the post-genomic era has been to determine the contribution of single-nucleotide variations in the enhancer regions to unique aspects of human biology (Yousaf, Sohail Raza, & Ali Abbasi, 2015). This is a difficult problem, as non-coding portion of our genome is vast and contains millions of human-specific sequence variations, making it hard to prioritize the mutations for follow-up functional studies in model organisms and cell lines (Franchini & Pollard, 2017).

Human accelerated regions (HARs) represent short, evolutionarily conserved DNA sequences in mammals/vertebrates with elevated substitution rate than expected in the modern human lineage (Marais, 2003). Accelerated evolutionary rate of HARs could reflect potential mechanisms such as positive selection, gene-conversion events and neutral substitutions (Doan et al., 2016). It has been shown that many HARs (~50%) are evolving under adaptive evolution (Kostka, Hubisz, Siepel, & Pollard, 2012). Population genetics tests further suggest that while many sequence changes in HARs predate diversification of modern human populations, some HARs do show evidence for recent population-specific selection trends (Kanginakudru, Metta, Jakati,

& Nagaraju, 2008; Pollard et al., 2006). Thereby, single-nucleotide variations in HARs have been implicated in human trait evolution, such as cognitive status and other physiological adaptabilities (Kanginakudru et al., 2008; Prabhakar et al., 2006).

Recently reported data has shown accelerated sequence evolution of empirically confirmed brain-specific human enhancers on comparison with closer non-human primate (Zehra & Abbasi, 2018). These enhancers were termed as brain exclusive human accelerated enhancers (BE-HAEs). These assorted set of BE-HAEs possessed single nucleotide variants (SNVs) that made them unique to fully modern humans in comparisons with Neanderthals and Denisovans (archaic humans) orthologous sequences. These SNVs were shown to have modified the binding sites of transcriptional factors *SOX2*, *RUNX*, and *FOS/JUND* and were further substantiated as single nucleotide polymorphisms (SNPs) in present-day human populations (Zehra & Abbasi, 2018). Long range haplotype (LRH) analysis has revealed that derived allele in BE-HAEs hs1210, inhabiting modern human specific binding motifs of *SOX2* is under positive selection in modern human population (Zehra & Abbasi, 2018).

From mammals to birds to insects SOX transcription factor plays an active role in the embryonic nervous system from the earliest stages of development, largely in the proliferating, undifferentiated neural cell precursors (Pevny & Nicolis, 2010; Schilham et al., 1996a). Functions of SOX2 protein in CNS development are dose-dependent and cellular-context dependent and are known to have diverged among closely related species such as mice and humans (Fantes et al., 2003; Sisodiya et al., 2006). These observations point to evolutionary conserved and divergent roles of the *SOX2* in CNS development. These previously known critical functions of *SOX2* in developing nervous system of bilaterians points to evolutionary developmental relevance of its putative binding motif that carries the fully modern human specific SNV within BE-HAEs hs1210 (Figure 7). Therefore, understanding how a nucleotide difference between archaic humans/non-human primates/non-primate mammals (TAGACT\*ACAATGGAT) and fully modern humans (TAGACA\*ACAATGGAT) modifies the interaction of SOX2 protein with its predicted DNA binding site may provide an important insight into the potential gene regulatory molecular mechanisms by which human-specific neurodevelopment patterns have originated.

The present study evaluates the putative functional impact of positively selected modern human-specific nucleotide variation within functionally confirmed brain exclusive human accelerated enhancer hs1210 (BE-HAEs hs1210) (Zehra & Abbasi, 2018). In this respect, the DNA binding properties of the SOX2 protein to ancestral (TAGACT\*ACAATGGAT) and modern human specific derived (TAGACA\*ACAATGGAT) alleles within BE-HAEs hs1210 were inspected (Figure 8 and Figure 9).

SOX2 is SRY-related HMG (high mobility group) box transcriptional factor protein. HMG box of SOX2 recognizes the sequence A/TAACAAA/T, it binds to the major groove and dramatically alter the DNA geometry and thereby regulate cell fate in developmental stage and tissue specific manner (Pevny & Lovell-Badge, 1997). In the present study we performed an electrophoretic mobility shift assay (EMSA) to determine if substitutions at SOX2 binding site within BE-HAEs hs1210 have any effect on protein binding. We found that despite of single nucleotide variation among respective binding sites, SOX2 protein binds to both the human and archaic hominins (Neanderthal/Denisovan) based DNA probes (Figure 11).

However, rigid docking and molecular dynamics simulations based bioinformatics data suggests that there are significant geometric and energetic differences in the binding of HMG box of SOX2 with its cognate DNA target sites differed by single residue between archaic hominins and modern human versions of BE-HAEs hs1210 (Figure 7). For instance, rigid molecular docking revealed that SOX2 HMG box grips major groove and makes specific contact with cognate target site (5'-TACAATG-3') containing ancestral T-allele in Neanderthal/Denisovan DNA (Figure 15A and Table 3). In contrast, intriguingly, SOX2 HMG box binds to slightly different residues within the target site (5'ACAAT 3') carrying derived A-allele in modern human DNA sequence involving major groove (Figure 15B and Table 3).

Evaluation of DNA-protein interactions based on rigid molecular docking and MD simulation revealed that the derived A-allele carrying DNA-SOX2 complex involved an increased number of hydrogen bonds than the ancestral T-allele carrying DNA-SOX2 complex (Table 4 and Table 5). Prediction of protein-DNA binding free energy interactions using MD simulations revealed an energetically favorable

interaction between derived A-allele DNA-SOX2 complex than that for the ancestral T-allele carrying DNA-SOX2 complex (Table 6).

Taken together, EMSA-based *in vitro* experiments, molecular docking, molecular dynamics simulations and free energy calculations suggest that the modern human specific SNV (T>A) within DNA target site of SOX2 HMG box did not abolish the DNA-Protein interactions but instead significantly enhanced the affinity of SOX2 protein for its target DNA site by altering three-dimensional geometry and energetics of nucleoprotein complexes. These results fit well with the previously published work showing that SOX2 HMG box can alter gene transcription patterns based on its ability to manipulate DNA geometry (Scaffidi & Bianchi, 2001).

Several lines of evidence suggest that the nucleotide substitution (at position Chr 2; 66535938) within *SOX2* binding site of functionally confirmed brain exclusive enhancer, which differs between fully modern humans and Neanderthals/Denisovans, alters the regulatory activity of BE-HAEs hs1210. First, it is located in the non-coding region that has already been characterized as forebrain exclusive enhancer via transgenic mice assay (Visel et al., 2013). Second, the DNA region where the substitution occurs is conserved over mammalian evolution and has previously been characterized as human accelerated enhancers (HAEs) that are likely to regulate human-specific traits (Zehra & Abbasi, 2018). Third, it falls at the 6th position within the 15-bp sequence motif identified as a target site for the transcription factor *SOX2* (TAGACA\*ACAATGGAT), known to play a vital role in neural development. Fourth, using extant human population genetic data it has been shown that the derived allele inhabiting modern human-specific binding motifs of *SOX2* is under positive selection in modern human populations, implicating further the role of this region in evolution of modern human-specific traits (Zehra & Abbasi, 2018). Fifth, in the present study we showed that the DNA sequence where the substitution occurs binds SOX2 *in vitro*. Sixth, comparison of the DNA-binding energetic reveals that there is a difference in the mechanism of binding of *SOX2* with ancestral and derived alleles, potentially creating transcriptional activity differences between variant alleles in forebrain tissues. Previous studies have shown that single-nucleotide substitutions within DNA binding site of *SOX2* results in a different level of transcription through changes in correct three-dimensional geometry of nucleoprotein complexes (Scaffidi & Bianchi, 2001).

Furthermore, higher affinity of SOX2 with its target binding sites has been associated with long-lived binding that contributes to more pluripotent progeny in developing mouse embryo whereas the weaker DNA-SOX2 interaction is known to decrease long-lived binding, *SOX2* target genes expression, and pluripotent cell numbers (White et al., 2016). According to the results of these studies, the SOX2-DNA binding affinity governs the fate of mammalian cells as early as the four-cell stage. (White et al., 2016). Therefore, it is conceivable to suggest that evolution of higher affinity of SOX2 for its target binding site within BE-HAEs *hs1210* might have resulted in more robust transactivation of respective target genes in the forebrain tissues of fully modern humans after they diverged from archaic humans (Neanderthals and Denisovans) some 450,000 years ago (Vernot & Akey, 2015).

Conceivably, such small-scale changes in TFBSs of master developmental regulators such as *SOX2* might have been instrumental in evolving differences in brain physiology and anatomy between anatomically modern humans and archaic hominins (Neanderthals and Denisovans) and between hominins and great apes. However, a fuller understanding of how different binding affinity of *SOX2* at the ancestral site (Neanderthals /Denisovans) and at the derived position on the human Chr 2; 66535938 (GRCh38/hg38) affects transcription awaits further studies in model systems.

Gains and losses of TFBSs are widespread and are known to have profound effects on organismal development. Early work documenting this principle include the changes in the regulation of *Shh* in the loss of limbs in snakes (Padhi, Mehinovic, Sams, Ng, & Turner, 2021), changes in *Tbx5* regulation in the evolution of fish fins (Adachi, Robinson, Goolsbee, & Shubin, 2016), changes in *PAX3/PAX7* regulation in craniofacial evolution in humans (Prescott et al., 2015), changes in *GDF6* regulation in the evolution of the human foot (Indjeian et al., 2016), and changes in *GADD45G* and *FZD8* regulation and evolution of mammalian brain size (Boyd et al., 2015a; McLean et al., 2011). It is noticeable that evolutionary rewiring of transcription circuitry does not require only gains and losses of TFBSs but can also entails differences in binding affinities of existing TFs with their target sites through changes in three-dimensional geometry of nucleoproteins complexes and binding energetics. In one such example,



positions in the human genome sequence that are different from the orthologous positions in archaic hominins (Neanderthals and Denisovans) have been associated with differential TF binding affinity and consequently the evolution of traits unique to modern humans such as modern language (Maricic et al., 2013).

The present study demonstrates that positively selected modern human-specific nucleotide variant within the non-coding intergenic regulatory HAR hs1210 has increased the DNA binding affinity of *SOX2* through changes in the three-dimensional geometry and binding energetics of the nucleoproteins complex. Because the hs1210 expressed the reporter gene exclusively in the forebrain of transgenic mice, more specifically in lateral ganglionic eminence (LGE), a transient structures in the developing telencephalon, it is tempting to speculate that the substitution at position Chr 2; 66535938 (GRCh38/hg38) in intergenic region of chromosome 2 might have involve in the evolution of forebrain (Figure 7C). It is noteworthy that this derived nucleotide variant in modern humans is not present in Neanderthals and Denisovans (Figure 7A). Thus, it is possible that this increase in affinity of *SOX2* for its target DNA site might have altered the forebrain specific expression of concerned gene bodies in modern humans, after their split from archaic hominins about 550,000–750,000 years ago (Pervaiz, Kang, Bao, & Abbasi, 2021; Prüfer et al., 2014), These findings are in line with previous studies, which demonstrate that the certain sub-anatomical regions of the forebrain evolved after split of the Neanderthals and anatomically modern human, most conspicuously parieto-temporal lobe of neocortex has increased and orbitofrontal cortex is wider in modern human as compared to Neanderthal (Bastir et al., 2011; Florio, Borrell, & Huttner, 2017).

One of the crucial issues concerning the human evolutionary biology is related to genetic mechanisms through which anatomical, morphological and behavioral differences arose between closely diverged archaic and modern humans. Several recent reports have now associated cis-acting regulatory variants with the divergence of various morphological traits between populations or closely diverged species (Enattah et al., 2002; Jeong et al., 2008; Takahashi, Takahashi, Ueda, & Takano-Shimizu, 2007). Given these examples from other species, it has been proposed that non-coding gene regulatory regions should be considered serious candidates through which to investigate

divergence between modern humans and their closest evolutionary relatives, the Neanderthals and Denisovans (Yan & McCoy, 2020). However, because of the degradation of ancient RNA, archaic gene expression cannot be measured through experimental studies. Therefore, technical and conceptual advancements are required in the field of ancient genomics and functional genomics to investigate the hominin gene expression evolution by indirect means (Yan & McCoy, 2020).

Here we used the combination of *in vitro* and computational approaches to show precisely how the SOX2 protein binds more efficiently to its putative binding site containing a positively selected nucleotide position derived in modern humans than does the ancestral allele in Neanderthals and Denisovans. Based on these results, one may speculate that these non-coding single nucleotide changes in regulatory regions that are unique to modern humans could be involved in evolution of gene expression differences between archaic and modern humans. In this case, evolution of modern human-specific traits might not entail major sacrifices in regulatory architecture in terms of gain and loss of TFBSs. Instead, major differences in gene expression patterns and consequently the trait differences between archaic and modern humans might have involved the evolution of the affinity differences of TFs for their target DNA sites. Thus, our data offers general insights into how the functional diversification of cis-regulatory regions through changes in TFs binding affinities contributes to evolutionary novelty and the origin of differences between the two sibling species such as archaic and modern humans

## 5. Conclusions

We now know that the large majority of all genomic changes that happened since the divergence of human– chimpanzee ancestor or human-archaic hominin ancestor are in non-coding regions. The challenge in the post-genomic era has been to identify modern human- specific non-coding substitutions that are responsible for unique aspects of our biology such as language, civilization and human specific neurodevelopmental mechanisms (Almécija et al., 2021). This is a hard task, because the non-coding genome is vast, requiring approaches to prioritize the non-coding mutations that are relevant to evolution (Scacheri & Scacheri, 2015). To this aim comparative genomics-based approaches have been employed to identify the human accelerated regions (HARs) and other human-specific genome sequences. HARs are mostly non-coding regions, conserved across vertebrates but carry human-specific substitutions and thus considered as promising candidates to hunt for variations explaining some of the modern human-specific traits (Marais, 2003). Indeed, recently reported experimental data has implicated specific HARs for the essential cis-regulatory functions in the human brain that are potentially important targets of recent human brain evolution (Doan et al., 2016). In this work we investigated the evolutionary significance of previously identified non-coding regulatory HAR, i.e. BE-HAEs hs1210 (Zehra & Abbasi, 2018). This particular HAR was previously shown to act as forebrain exclusive enhancer and is known to contain positively selected human-specific nucleotide change that has arisen after the split between modern and archaic human lineages (Visel et al., 2013). Here we used the combinations of in vitro and bioinformatics analysis to comparatively characterize the evolutionary significance of positively selected modern human specific substitution (T>A) within BE-HAEs hs1210. Our comparative molecular structural analysis showed that modern human-specific single nucleotide substitution has increased the affinity of *SOX2* transcriptional factor for its target binding site within BE-HAEs hs1210. These findings suggest that this predicted enhanced affinity of *SOX2* towards its target site could drive the target gene expression more robustly within forebrain of modern humans compared with the archaic humans or alternatively within novel territories in the forebrain of modern humans. However, further experimental studies in model systems will be necessary to confirm whether in

fact these changes in transcriptional factor binding affinity translate into functional modifications of gene expression.

In a more general view, results presented in this study are relevant to general understandings of genetic architecture underlying in the human evolution, Furthermore, molecular mechanisms driving the evolution of the gene regulation, and role of cis-regulatory variants in origin of species modifications.

---

**References**

- Abbasi, A. A. (2011). Evolution of vertebrate appendicular structures: Insight from genetic and palaeontological data. *Developmental Dynamics*, 240(5), 1005-1016.
- Adachi, N., Robinson, M., Goolsbee, A., & Shubin, N. H. (2016). Regulatory evolution of Tbx5 and the origin of paired appendages. *Proceedings of the National Academy of Sciences*, 113(36), 10115-10120. doi:<https://doi.org/10.1073/pnas.1609997113>
- Ali, S., Arif, I., Iqbal, A., Hussain, I., Abrar, M., Khan, M. R., . . . Abbasi, A. A. (2021). Comparative genomic analysis of human GLI2 locus using slowly evolving fish revealed the ancestral gnathostome set of early developmental enhancers. *Developmental Dynamics*, 250(5), 669-683. doi:<https://doi.org/10.1002/dvdy.291>
- Allen, M. P., & Tildesley, D. J. (2012). *Computer simulation in chemical physics* (Vol. 397): Springer Science & Business Media.
- Almécija, S., Hammond, A. S., Thompson, N. E., Pugh, K. D., Moyà-Solà, S., & Alba, D. M. (2021). Fossil apes and human evolution. *Science*, 372(6542). doi:<https://doi.org/10.1126/science.abb4363>
- Allou, L., Balzano, S., Magg, A., Quinodoz, M., Royer-Bertrand, B., Schopflin, R., Chan, W.L., Speck-Martin, C.E., Carvalho, D.R., Farage, L. and Lourenco, C.M. (2021). Non-coding deletions identify Maenli lncRNA as a limb-specific En 1 regulator. *Nature*, 592(7852), 93-98
- Anand, S., Wang, W. C., Powell, D. R., Bolanowski, S. A., Zhang, J., Ledje, C., . . . Shashikant, C. S. (2003). Divergence of Hoxc8 early enhancer parallels diverged axial morphologies between mammals and fishes. *Proceedings of the National Academy of Sciences*, 100(26), 15666-15669.
- Andrusier, N., Nussinov, R., Wolfson, H. J. J. P. S., Function,, & Bioinformatics. (2007). FireDock: fast interaction refinement in molecular docking. 69(1), 139-159.
- Anwar, S., Minhas, R., Ali, S., Lambert, N., Kawakami, Y., Elgar, G., . . . Abbasi, A. A. (2015). Identification and functional characterization of novel transcriptional enhancers involved in regulating human GLI 3 expression during early

- development. *Development growth & differentiation*, 57(8), 570-580. doi:  
<https://doi.org/10.1111/dgd.12239>
- Arnold, S., Pelet, A., Amiel, J., Borrego, S., Hofstra, R., Tam, P., . . . Chakravarti, A. (2009). Interaction between a chromosome 10 RET enhancer and chromosome 21 in the Down syndrome–Hirschsprung disease association. *Human mutation*, 30(5), 771-775.
- Arnosti, D. N., & Kulkarni, M. M. J. J. o. c. b. (2005). Transcriptional enhancers: Intelligent enhanceosomes or flexible billboards? , 94(5), 890-898.
- Bailey, T., Krajewski, P., Ladunga, I., Lefebvre, C., Li, Q., Liu, T., . . . Zhang, J. J. P. C. B. (2013). Practical guidelines for the comprehensive analysis of ChIP-seq data. 9(11), e1003326.
- Banerji, J., Rusconi, S., & Schaffner, W. J. C. (1981). Expression of a  $\beta$ -globin gene is enhanced by remote SV40 DNA sequences. 27(2), 299-308.
- Barber, B. A., Liyanage, V. R., Zachariah, R. M., Olson, C. O., Bailey, M. A., & Rastegar, M. J. A. o. A.-A. A. (2013). Dynamic expression of MEIS1 homeoprotein in E14. 5 forebrain and differentiated forebrain-derived neural stem cells. 195(5), 431-440.
- Bastir, M., Rosas, A., Gunz, P., Peña-Melian, A., Manzi, G., Harvati, K., . . . Hublin, J.-J. (2011). Evolution of the base of the brain in highly encephalized human species. *Nature Communications*, 2(1), 1-8. doi:<https://doi.org/10.1038/ncomms1593>
- Bird, C. P., Stranger, B. E., Liu, M., Thomas, D. J., Ingle, C. E., Beazley, C., . . . Dermitzakis, E. T. (2007). Fast-evolving noncoding sequences in the human genome. *Genome biology*, 8(6), R118.
- Botquin, V., Hess, H., Fuhrmann, G., Anastassiadis, C., Gross, M. K., Vriend, G., . . . development. (1998). New POU dimer configuration mediates antagonistic control of an osteopontin preimplantation enhancer by Oct-4 and Sox-2. 12(13), 2073-2090.
- Boyd, J. L., Skove, S. L., Rouanet, J. P., Pilaz, L.-J., Bepler, T., Gordân, R., . . . Silver, D. L. (2015a). Human-chimpanzee differences in a FZD8 enhancer alter cell-cycle dynamics in the developing neocortex. *Current Biology*, 25(6), 772-779. doi:<https://doi.org/10.1016/j.cub.2015.01.041>

- Boyd, J. L., Skove, S. L., Rouanet, J. P., Pilaz, L.-J., Bepler, T., Gordân, R., . . . Silver, D. L. J. C. B. (2015b). Human-chimpanzee differences in a FZD8 enhancer alter cell-cycle dynamics in the developing neocortex. *25(6)*, 772-779.
- Brand, A. H., Breeden, L., Abraham, J., Sternglanz, R., & Nasmyth, K. J. C. (1985). Characterization of a “silencer” in yeast: a DNA sequence with properties opposite to those of a transcriptional enhancer. *41(1)*, 41-48.
- Britten, R. J., & Davidson, E. H. (1969). Gene regulation for higher cells: a theory. *Science*, *165(3891)*, 349-357.
- Brunskill, E. W., Ehrman, L. A., Williams, M. T., Klanke, J., Hammer, D., Schaefer, T. L., . . . Vorhees, C. V. (2005). Abnormal neurodevelopment, neurosignaling and behaviour in Npas3-deficient mice. *European Journal of Neuroscience*, *22(6)*, 1265-1276.
- Buck, M. J., & Lieb, J. D. J. G. (2004). ChIP-chip: considerations for the design, analysis, and application of genome-wide chromatin immunoprecipitation experiments. *83(3)*, 349-360.
- Bulger, M., & Groudine, M. J. C. (2011). Functional and mechanistic diversity of distal transcription enhancers. *144(3)*, 327-339.
- Burley, S. K., Almo, S. C., Bonanno, J. B., Capel, M., Chance, M. R., Gaasterland, T., . . . Swaminathan, S. (1999). Structural genomics: beyond the human genome project. *Nature genetics*, *23(2)*, 151-157.
- Bylund, M., Andersson, E., Novitch, B. G., & Muhr, J. J. N. n. (2003). Vertebrate neurogenesis is counteracted by Sox1–3 activity. *6(11)*, 1162-1168.
- Cáceres, M., Lachuer, J., Zapala, M. A., Redmond, J. C., Kudo, L., Geschwind, D. H., . . . Barlow, C. (2003a). Elevated gene expression levels distinguish human from non-human primate brains. *Proceedings of the National Academy of Sciences*, *100(22)*, 13030-13035.
- Cáceres, M., Lachuer, J., Zapala, M. A., Redmond, J. C., Kudo, L., Geschwind, D. H., . . . Barlow, C. J. P. o. t. N. A. o. S. (2003b). Elevated gene expression levels distinguish human from non-human primate brains. *100(22)*, 13030-13035.
- Calhoun, V. C., Stathopoulos, A., & Levine, M. J. P. o. t. N. A. o. S. (2002). Promoter-proximal tethering elements regulate enhancer-promoter specificity in the *Drosophila* Antennapedia complex. *99(14)*, 9243-9247.

- Carroll, S. B. (2003). Genetics and the making of Homo sapiens. *Nature*, 422(6934), 849-857.
- Carroll, S. B. (2005). Evolution at two levels: on genes and form. *PLoS biology*, 3(7), e245.
- Cavallaro, M., Mariani, J., Lancini, C., Latorre, E., Caccia, R., Gullo, F., . . . Ronchi, A. J. D. (2008). Impaired generation of mature neurons by neural stem cells from hypomorphic SOX2 mutants. *135*(3), 541-557.
- Charvet, C. J., & Striedter, G. F. (2011). Developmental modes and developmental mechanisms can channel brain evolution. *Frontiers in Neuroanatomy*, 5, 4.
- Cimadamore, F., Shah, M., Amador-Arjona, A., Navarro-Peran, E., Chen, C., Huang, C.-T., . . . research, m. (2012). SOX2 modulates levels of MITF in normal human melanocytes, and melanoma lines in vitro. *25*(4), 533.
- Clark, M. D., Hennig, S., Herwig, R., Clifton, S. W., Marra, M. A., Lehrach, H., . . . Group, W.-G. E. (2001). An oligonucleotide fingerprint normalized and expressed sequence tag characterized zebrafish cDNA library. *Genome Research*, 11(9), 1594-1602.
- Cohen, F. E., & Sternberg, M. J. (1980). On the prediction of protein structure: the significance of the root-mean-square deviation. *Journal of molecular biology*, 138(2), 321-333. doi:[https://doi.org/10.1016/0022-2836\(80\)90289-2](https://doi.org/10.1016/0022-2836(80)90289-2)
- Comeau, S. R., Gatchell, D. W., Vajda, S., & Camacho, C. J. J. N. a. r. (2004). ClusPro: a fully automated algorithm for protein–protein docking. *32*(suppl\_2), W96-W99.
- Comeau, S. R., Kozakov, D., Brenke, R., Shen, Y., Beglov, D., Vajda, S. J. P. S., Function,, & Bioinformatics. (2007). ClusPro: performance in CAPRI rounds 6–11 and the new server. *69*(4), 781-785.
- Connolly, M. L. J. S. (1983). Solvent-accessible surfaces of proteins and nucleic acids. *221*(4612), 709-713.
- Cooper, A. (1976). Thermodynamic fluctuations in protein molecules. *Proceedings of the National Academy of Sciences*, 73(8), 2740-2741. doi:<https://doi.org/10.1073/pnas.73.8.2740>
- Crawford, G. E., Holt, I. E., Mullikin, J. C., Tai, D., Green, E. D., Wolfsberg, T. G., . . . Sciences, N. I. o. H. I. S. C. J. P. o. t. N. A. o. (2004). Identifying gene



- regulatory elements by genome-wide recovery of DNase hypersensitive sites. *101*(4), 992-997.
- Cui, C.-P., Zhang, Y., Wang, C., Yuan, F., Li, H., Yao, Y., . . . Liu, C. H. J. N. c. (2018). Dynamic ubiquitylation of SOX2 regulates proteostasis and governs neural progenitor cell differentiation. *9*(1), 1-15.
- Davidson, E. H., & Erwin, D. H. (2006). Gene regulatory networks and the evolution of animal body plans. *Science*, *311*(5762), 796-800.
- de Vries, S. J., & Bonvin, A. M. (2011). CPORT: a consensus interface predictor and its performance in prediction-driven docking with HADDOCK. *PloS one*, *6*(3), e17695. doi:<https://doi.org/10.1371/journal.pone.0017695>
- De Vries, S. J., Van Dijk, M., & Bonvin, A. M. (2010). The HADDOCK web server for data-driven biomolecular docking. *Nature protocols*, *5*(5), 883-897.
- Dickmeis, T., & Müller, F. (2005). The identification and functional characterisation of conserved regulatory elements in developmental genes. *Briefings in Functional Genomics*, *3*(4), 332-350.
- Doan, R. N., Bae, B.-I., Cubelos, B., Chang, C., Hossain, A. A., Al-Saad, S., . . . Balkhy, S. (2016). Mutations in human accelerated regions disrupt cognition and social behavior. *Cell*, *167*(2), 341-354. e312. doi:<https://doi.org/10.1016/j.cell.2016.08.071>
- Dominguez, C., Boelens, R., & Bonvin, A. M. (2003). HADDOCK: a protein– protein docking approach based on biochemical or biophysical information. *Journal of the American Chemical Society*, *125*(7), 1731-1737.
- Duhovny, D., Nussinov, R., & Wolfson, H. J. (2002). *Efficient unbound docking of rigid molecules*. Paper presented at the International workshop on algorithms in bioinformatics.
- Dunbar, R. I., & Shultz, S. J. P. T. o. t. R. S. B. B. S. (2007). Understanding primate brain evolution. *362*(1480), 649-658.
- Enard, W., Khaitovich, P., Klose, J., Zöllner, S., Heissig, F., Giavalisco, P., . . . Ravid, R. (2002). Intra-and interspecific variation in primate gene expression patterns. *Science*, *296*(5566), 340-343.
- Enard, W. J. C. B. (2015). Human evolution: enhancing the brain. *25*(10), R421-R423.

- Enattah, N. S., Sahi, T., Savilahti, E., Terwilliger, J. D., Peltonen, L., & Järvelä, I. (2002). Identification of a variant associated with adult-type hypolactasia. *Nature genetics*, *30*(2), 233-237. doi:<https://doi.org/10.1038/ng826>
- Epstein, D. J. (2009). Cis-regulatory mutations in human disease. *Briefings in Functional Genomics and Proteomics*, *8*(4), 310-316.
- Fang, L., Ahn, J. K., Wodziak, D., & Sibley, E. J. H. g. (2012). The human lactase persistence-associated SNP– 13910\* T enables in vivo functional persistence of lactase promoter–reporter transgene expression. *131*(7), 1153-1159.
- Fantes, J., Ragge, N. K., Lynch, S.-A., McGill, N. I., Collin, J. R. O., Howard-Peebles, P. N., . . . van Heyningen, V. (2003). Mutations in SOX2 cause anophthalmia. *Nature genetics*, *33*(4), 462-463. doi:<https://doi.org/10.1038/ng1120>
- Fernández, X. M., & Birney, E. (2010). Ensembl genome browser *Vogel and Motulsky's Human Genetics* (pp. 923-939): Springer.
- Ferri, A. L., Cavallaro, M., Braida, D., Di Cristofano, A., Canta, A., Vezzani, A., . . . DeBiasi, S. J. D. (2004). SOX2 deficiency causes neurodegeneration and impaired neurogenesis in the adult mouse brain. *131*(15), 3805-3819.
- Florio, M., Borrell, V., & Huttner, W. B. (2017). Human-specific genomic signatures of neocortical expansion. *Current opinion in neurobiology*, *42*, 33-44. doi:<https://doi.org/10.1016/j.conb.2016.11.004>
- Franchini, L. F., & Pollard, K. S. (2015). Can a few non-coding mutations make a human brain? *Bioessays*, *37*(10), 1054-1061.
- Franchini, L. F., & Pollard, K. S. (2017). Human evolution: the non-coding revolution. *BMC biology*, *15*(1), 1-12. doi:<https://doi.org/10.1186/s12915-017-0428-9>
- Furey, T. S. J. N. R. G. (2012). ChIP–seq and beyond: new and improved methodologies to detect and characterize protein–DNA interactions. *13*(12), 840-852.
- Gangemi, R. M. R., Griffero, F., Marubbi, D., Perera, M., Capra, M. C., Malatesta, P., . . . Corte, G. J. S. c. (2009). SOX2 silencing in glioblastoma tumor-initiating cells causes stop of proliferation and loss of tumorigenicity. *27*(1), 40-48.
- Geschwind, D. H., & Rakic, P. J. N. (2013). Cortical evolution: judge the brain by its cover. *80*(3), 633-647.
- Gilbert, S. L., Dobyns, W. B., & Lahn, B. T. (2005). Genetic links between brain development and brain evolution. *Nature Reviews Genetics*, *6*(7), 581-590.

- Glinsky, G., & Barakat, T. S. (2019). The evolution of Great Apes has shaped the functional enhancers' landscape in human embryonic stem cells. *Stem cell research*, 37, 101456. doi:<https://doi.org/10.1016/j.scr.2019.101456>
- Goodrich, J. A., & Tjian, R. J. C. o. i. c. b. (1994). TBP-TAF complexes: selectivity factors for eukaryotic transcription. 6(3), 403-409.
- Graham, V., Khudyakov, J., Ellis, P., & Pevny, L. J. N. (2003). SOX2 functions to maintain neural progenitor identity. 39(5), 749-765.
- Gu, J., & Gu, X. (2003). Induced gene expression in human brain after the split from chimpanzee. *Trends in Genetics*, 19(2), 63-65.
- Gu, J., & Gu, X. J. T. i. G. (2003). Induced gene expression in human brain after the split from chimpanzee. 19(2), 63-65.
- Gubbay, J., Collignon, J., Koopman, P., Capel, B., Economou, A., Münsterberg, A., . . . Lovell-Badge, R. J. N. (1990). A gene mapping to the sex-determining region of the mouse Y chromosome is a member of a novel family of embryonically expressed genes. 346(6281), 245-250.
- Guth, S., Wegner, M. J. C., & sciences, m. l. (2008). Having it both ways: Sox protein function between conservation and innovation. 65(19), 3000-3018.
- Hacia, J. G. (2001). Genome of the apes. *Trends in genetics*, 17(11), 637-645.
- Hawrylycz, M. J., Lein, E. S., Guillozet-Bongaarts, A. L., Shen, E. H., Ng, L., Miller, J. A., . . . Riley, Z. L. J. N. (2012). An anatomically comprehensive atlas of the adult human brain transcriptome. 489(7416), 391-399.
- Hellman, L. M., & Fried, M. G. (2007). Electrophoretic mobility shift assay (EMSA) for detecting protein–nucleic acid interactions. *Nature protocols*, 2(8), 1849. doi:<https://doi.org/10.1038/nprot.2007.249>
- Hellman, L. M., & Fried, M. G. J. N. p. (2007). Electrophoretic mobility shift assay (EMSA) for detecting protein–nucleic acid interactions. 2(8), 1849.
- Hindorff, L. A., Sethupathy, P., Junkins, H. A., Ramos, E. M., Mehta, J. P., Collins, F. S., & Manolio, T. A. J. P. o. t. N. A. o. S. (2009). Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. 106(23), 9362-9367.
- Hnisz, D., Abraham, B. J., Lee, T. I., Lau, A., Saint-André, V., Sigova, A. A., . . . Young, R. A. (2013). Super-enhancers in the control of cell identity and disease. *Cell*, 155(4), 934-947.

- Holian, B. (1995). Atomistic computer simulations of shock waves. *Shock waves*, 5(3), 149-157.
- Hornshøj, H., Nielsen, M. M., Sinnott-Armstrong, N. A., Świtnicki, M. P., Juul, M., Madsen, T., . . . Hobolth, A. (2018). Pan-cancer screen for mutations in non-coding elements with conservation and cancer specificity reveals correlations with expression and survival. *NPJ genomic medicine*, 3(1), 1-14.
- Hou, T., Wang, J., Li, Y., & Wang, W. (2011). Assessing the performance of the MM/PBSA and MM/GBSA methods. 1. The accuracy of binding free energy calculations based on molecular dynamics simulations. *Journal of chemical information and modeling*, 51(1), 69-82. doi:<https://doi.org/10.1021/ci100275a>
- Indjeian, V. B., Kingman, G. A., Jones, F. C., Guenther, C. A., Grimwood, J., Schmutz, J., . . . Kingsley, D. M. (2016). Evolving new skeletal traits by cis-regulatory changes in bone morphogenetic proteins. *Cell*, 164(1-2), 45-56. doi:<https://doi.org/10.1016/j.cell.2015.12.007>
- Jeong, S., Rebeiz, M., Andolfatto, P., Werner, T., True, J., & Carroll, S. B. (2008). The evolution of gene regulation underlies a morphological difference between two *Drosophila* sister species. *Cell*, 132(5), 783-793. doi:<https://doi.org/10.1016/j.cell.2008.01.014>
- John, S., Sabo, P. J., Canfield, T. K., Lee, K., Vong, S., Weaver, M., . . . Thurman, R. E. J. C. p. i. m. b. (2013). Genome-scale mapping of DNase I hypersensitivity. *103*(1), 21.27. 21-21.27. 20.
- Johnson, D. S., Li, W., Gordon, D. B., Bhattacharjee, A., Curry, B., Ghosh, J., . . . Flicek, P. (2008). Systematic evaluation of variability in ChIP-chip experiments using predefined DNA targets. *Genome research*, 18(3), 393-403.
- Joshi, T., Joshi, T., Sharma, P., Chandra, S., & Pande, V. (2021). Molecular docking and molecular dynamics simulation approach to screen natural compounds for inhibition of *Xanthomonas oryzae* pv. *Oryzae* by targeting peptide deformylase. *Journal of Biomolecular Structure and Dynamics*, 39(3), 823-840. doi:<https://doi.org/10.1080/07391102.2020.1719200>
- Junion, G., Spivakov, M., Girardot, C., Braun, M., Gustafson, E. H., Birney, E., & Furlong, E. E. J. C. (2012). A transcription factor collective defines cardiac cell fate and reflects lineage history. *148*(3), 473-486.

- Kamachi, Y., Cheah, K. S., Kondoh, H. J. M., & Biology, C. (1999). Mechanism of regulatory target selection by the SOX high-mobility-group domain proteins as revealed by comparison of SOX1/2/3 and SOX9. *19*(1), 107-120.
- Kamachi, Y., Uchikawa, M., Tanouchi, A., Sekido, R., & Kondoh, H. (2001). Pax6 and SOX2 form a co-DNA-binding partner complex that regulates initiation of lens development. *Genes & development*, *15*(10), 1272-1286.
- Kamm, G. B., Pisciotano, F., Kliger, R., & Franchini, L. F. (2013). The developmental brain gene NPAS3 contains the largest number of accelerated regulatory sequences in the human genome. *Molecular biology and evolution*, *30*(5), 1088-1102.
- Kanginakudru, S., Metta, M., Jakati, R., & Nagaraju, J. (2008). Genetic evidence from Indian red jungle fowl corroborates multiple domestication of modern day chicken. *BMC evolutionary biology*, *8*(1), 1-14. doi:<https://doi.org/10.1186/1471-2148-8-174>
- Karolchik, D., Baertsch, R., Diekhans, M., Furey, T. S., Hinrichs, A., Lu, Y., . . . Thomas, D. J. (2003). The UCSC genome browser database. *Nucleic acids research*, *31*(1), 51-54.
- Kent, J., Wheatley, S. C., Andrews, J. E., Sinclair, A. H., & Koopman, P. J. D. (1996). A male-specific role for SOX9 in vertebrate sex determination. *122*(9), 2813-2822.
- Khan, A., Fornes, O., Stigliani, A., Gheorghe, M., Castro-Mondragon, J. A., Van Der Lee, R., . . . Tan, G. J. N. a. r. (2018). JASPAR 2018: update of the open-access database of transcription factor binding profiles and its web framework. *46*(D1), D260-D266.
- Khan, A., Khan, S., Saleem, S., Nizam-Uddin, N., Mohammad, A., Khan, T., . . . Suleman, M. (2021). Immunogenomics guided design of immunomodulatory multi-epitope subunit vaccine against the SARS-CoV-2 new variants, and its validation through in silico cloning and immune simulation. *Computers in biology and medicine*, *133*, 104420.
- Khan, A., Zia, T., Suleman, M., Khan, T., Ali, S. S., Abbasi, A. A., . . . Wei, D. Q. (2021). Higher infectivity of the SARS-CoV-2 new variants is associated with K417N/T, E484K, and N501Y mutants: An insight from structural data. *Journal of cellular physiology*.

- Khan, M. T., Ali, S., Zeb, M. T., Kaushik, A. C., Malik, S. I., & Wei, D.-Q. (2020). Gibbs Free Energy Calculation of Mutation in PncA and RpsA Associated With Pyrazinamide Resistance. *Frontiers in molecular biosciences*, 7, 52.
- King, N., Westbrook, M. J., Young, S. L., Kuo, A., Abedin, M., Chapman, J., . . . Letunic, I. J. N. (2008). The genome of the choanoflagellate *Monosiga brevicollis* and the origin of metazoans. *451*(7180), 783-788.
- Kishi, M., Mizuseki, K., Sasai, N., Yamazaki, H., Shiota, K., Nakanishi, S., & Sasai, Y. J. D. (2000). Requirement of SOX2-mediated signaling for differentiation of early *Xenopus* neuroectoderm. *127*(4), 791-800.
- Kolovos, P., Knoch, T. A., Grosveld, F. G., Cook, P. R., & Papantonis, A. (2012). Enhancers and silencers: an integrated and simple model for their function. *Epigenetics & chromatin*, 5(1), 1.
- Kolovos, P., Knoch, T. A., Grosveld, F. G., Cook, P. R., Papantonis, A. J. E., & chromatin. (2012). Enhancers and silencers: an integrated and simple model for their function. 5(1), 1-8.
- Koshikawa, S. (2020). Evolution of wing pigmentation in *Drosophila*: Diversity, physiological regulation, and cis-regulatory evolution. *Development, growth & differentiation*, 62(5), 269-278. doi:<https://doi.org/10.1111/dgd.12661>
- Kostka, D., Hubisz, M. J., Siepel, A., & Pollard, K. S. (2012). The role of GC-biased gene conversion in shaping the fastest evolving regions of the human genome. *Molecular biology and evolution*, 29(3), 1047-1057. doi:<https://doi.org/10.1093/molbev/msr279>
- Kozakov, D., Beglov, D., Bohnuud, T., Mottarella, S. E., Xia, B., Hall, D. R., . . . Bioinformatics. (2013). How good is automated protein docking? , *81*(12), 2159-2166.
- Kräutler, V., Van Gunsteren, W. F., & Hünenberger, P. H. (2001). A fast SHAKE algorithm to solve distance constraint equations for small molecules in molecular dynamics simulations. *Journal of computational chemistry*, 22(5), 501-508.
- Krissinel, E. (2010). Crystal contacts as nature's docking solutions. *Journal of computational chemistry*, 31(1), 133-143. doi:<https://doi.org/10.1002/jcc.21303>

- Kuhlbrodt, K., Herbarth, B., Sock, E., Hermans-Borgmeyer, I., & Wegner, M. (1998). Sox10, a novel transcriptional modulator in glial cells. *Journal of Neuroscience*, *18*(1), 237-250.
- Kvon, E. Z., Waymack, R., Gad, M., & Wunderlich, Z. (2021). Enhancer redundancy in development and disease. *Nature Reviews Genetics*, *22*(5), 324-336. doi:<https://doi.org/10.1038/s41576-020-00311-x>
- Larkin, M. A., Blackshields, G., Brown, N. P., Chenna, R., McGettigan, P. A., McWilliam, H., . . . Lopez, R. (2007). Clustal W and Clustal X version 2.0. *bioinformatics*, *23*(21), 2947-2948. doi:<https://doi.org/10.1093/bioinformatics/btm404>
- Laskowski, R. A., Jabłońska, J., Pravda, L., Vařeková, R. S., & Thornton, J. M. (2018). PDBsum: Structural summaries of PDB entries. *Protein science*, *27*(1), 129-134.
- Laudet, V., Stehelin, D., & Clevers, H. J. N. a. r. (1993). Ancestry and diversity of the HMG box superfamily. *21*(10), 2493-2501.
- Lee, Y. I., Lee, S., Das, G. C., Park, U. S., Park, S. M., & Lee, Y. I. J. O. (2000). Activation of the insulin-like growth factor II transcription by aflatoxin B1 induced p53 mutant 249 is caused by activation of transcription complexes; implications for a gain-of-function during the formation of hepatocellular carcinoma. *19*(33), 3717-3726.
- Lefebvre, V., Dumitriu, B., Penzo-Méndez, A., Han, Y., Pallavi, B. J. T. i. j. o. b., & biology, c. (2007). Control of cell fate and differentiation by Sry-related high-mobility-group box (Sox) transcription factors. *39*(12), 2195-2214.
- Lelli, K. M., Slattery, M., & Mann, R. S. J. A. r. o. g. (2012). Disentangling the many layers of eukaryotic transcriptional regulation. *46*, 43-68.
- Leonhard, K., & Deiters, U. (2002). Monte Carlo simulations of nitrogen using an ab initio potential. *Molecular Physics*, *100*(15), 2571-2585.
- Lettice, L. A., Heaney, S. J., Purdie, L. A., Li, L., de Beer, P., Oostra, B. A., . . . de Graaff, E. (2003). A long-range Shh enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly. *Human molecular genetics*, *12*(14), 1725-1735.
- Levchenko, A., Kanapin, A., Samsonova, A., & Gainetdinov, R. R. (2017). Human accelerated regions and other human-specific sequence variations in the context

- of evolution and their relevance for brain development. *Genome biology and evolution*, *10*(1), 166-188.
- Levine, M., Cattoglio, C., & Tjian, R. J. C. (2014). Looping back to leap forward: transcription enters a new era. *157*(1), 13-25.
- Levine, M., & Tjian, R. (2003). Transcription regulation and animal diversity. *Nature*, *424*(6945), 147-151.
- Li, Q., Peterson, K. R., Fang, X., & Stamatoyannopoulos, G. J. B., The Journal of the American Society of Hematology. (2002). Locus control regions. *100*(9), 3077-3086.
- Lobanov, M. Y., Bogatyreva, N., & Galzitskaya, O. (2008). Radius of gyration as an indicator of protein structure compactness. *Molecular Biology*, *42*(4), 623-628. doi:<https://doi.org/10.1134/S0026893308040195>
- Long, H. K., Prescott, S. L., & Wysocka, J. (2016). Ever-changing landscapes: transcriptional enhancers in development and evolution. *Cell*, *167*(5), 1170-1187. doi:<https://doi.org/10.1016/j.cell.2016.09.018>
- Marais, G. (2003). Biased gene conversion: implications for genome and sex evolution. *TRENDS in Genetics*, *19*(6), 330-338. doi:[https://doi.org/10.1016/S0168-9525\(03\)00116-1](https://doi.org/10.1016/S0168-9525(03)00116-1)
- Marcelli, G., & Sadus, R. J. (1999). Molecular simulation of the phase behavior of noble gases using accurate two-body and three-body intermolecular potentials. *The Journal of chemical physics*, *111*(4), 1533-1540.
- Maricic, T., Günther, V., Georgiev, O., Gehre, S., Čurlin, M., Schreiweis, C., . . . Lalueza-Fox, C. (2013). A recent evolutionary change affects a regulatory element in the human FOXP2 gene. *Molecular Biology and Evolution*, *30*(4), 844-852. doi:<https://doi.org/10.1093/molbev/mss271>
- Marshall, N. F., Peng, J., Xie, Z., & Price, D. H. J. J. o. B. C. (1996). Control of RNA polymerase II elongation potential by a novel carboxyl-terminal domain kinase. *271*(43), 27176-27183.
- Marshall, N. F., & Price, D. H. J. J. o. B. C. (1995). Purification of P-TEFb, a transcription factor required for the transition into productive elongation. *270*(21), 12335-12338.
- Martin, M. J., Rayner, J. C., Gagneux, P., Barnwell, J. W., & Varki, A. (2005). Evolution of human-chimpanzee differences in malaria susceptibility:



- relationship to human genetic loss of N-glycolylneuraminic acid. *Proceedings of the National Academy of Sciences*, 102(36), 12819-12824.
- Maston, G. A., Evans, S. K., & Green, M. R. (2006a). Transcriptional regulatory elements in the human genome. *Annu. Rev. Genomics Hum. Genet.*, 7, 29-59.
- Maston, G. A., Evans, S. K., & Green, M. R. J. A. R. G. H. G. (2006b). Transcriptional regulatory elements in the human genome. 7, 29-59.
- Matsushima, D., Heavner, W., & Pevny, L. H. J. D. (2011). Combinatorial regulation of optic cup progenitor cell fate by SOX2 and PAX6. *138*(3), 443-454.
- McConkey, B. J., Sobolev, V., & Edelman, M. J. C. S. (2002). The performance of current methods in ligand–protein docking. 845-856.
- McHenry, H. M. (1994). Behavioral ecological implications of early hominid body size. *Journal of Human Evolution*, 27(1-3), 77-87.
- McLean, C. Y., Reno, P. L., Pollen, A. A., Bassan, A. I., Capellini, T. D., Guenther, C., . . . Schaar, B. T. (2011). Human-specific loss of regulatory DNA and the evolution of human-specific traits. *Nature*, 471(7337), 216-219. doi:<https://doi.org/10.1038/nature09774>
- Meza, J. C. (2010). Steepest descent. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2(6), 719-722.
- Morris, G. M., Huey, R., Lindstrom, W., Sanner, M. F., Belew, R. K., Goodsell, D. S., & Olson, A. J. J. o. c. c. (2009). AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility. *30*(16), 2785-2791.
- Müller, F., & Tora, L. J. B. e. B. A.-G. R. M. (2014). Chromatin and DNA sequences in defining promoters for transcription initiation. *1839*(3), 118-128.
- Mundade, R., Ozer, H. G., Wei, H., Prabhu, L., & Lu, T. J. C. C. (2014). Role of ChIP-seq in the discovery of transcription factor binding sites, differential gene regulation mechanism, epigenetic marks and beyond. *13*(18), 2847-2852.
- Murshudov, G. N., Skubák, P., Lebedev, A. A., Pannu, N. S., Steiner, R. A., Nicholls, R. A., . . . Vagin, A. A. (2011). REFMAC5 for the refinement of macromolecular crystal structures. *Acta Crystallographica Section D: Biological Crystallography*, 67(4), 355-367. doi:<https://doi.org/10.1107/S0907444911001314>
- Neubauer, S., Hublin, J.-J., & Gunz, P. (2018). The evolution of modern human brain shape. *Science advances*, 4(1), eaao5961.

- Noonan, J. P., & McCallion, A. S. (2010). Genomics of long-range regulatory elements. *Annual review of genomics and human genetics*, *11*, 1-23.
- Ogbourne, S., & Antalis, T. M. J. B. J. (1998). Transcriptional control and the role of silencers in transcriptional regulation in eukaryotes. *331*(1), 1-14.
- Osterwalder, M., Barozzi, I., Tissières, V., Fukuda-Yuzawa, Y., Mannion, B. J., Afzal, S. Y., . . . Pickle, C. S. (2018). Enhancer redundancy provides phenotypic robustness in mammalian development. *Nature*, *554*(7691), 239-243.
- Padhi, E. M., Mehinovic, E., Sams, E. I., Ng, J. K., & Turner, T. N. (2021). ACES: Analysis of Conservation with Expansive Species. *bioRxiv*. doi: <https://doi.org/10.1101/2021.06.16.448733>
- Panne, D. (2008). The enhanceosome. *Current opinion in structural biology*, *18*(2), 236-242.
- Pearce, E., Stringer, C., & Dunbar, R. I. (2013). New insights into differences in brain organization between Neanderthals and anatomically modern humans. *Proceedings of the Royal Society B: Biological Sciences*, *280*(1758), 20130168.
- Pennacchio, L. A., Bickmore, W., Dean, A., Nobrega, M. A., & Bejerano, G. (2013a). Enhancers: five essential questions. *Nature Reviews Genetics*, *14*(4), 288-295. doi:<https://doi.org/10.1038/nrg3458>
- Pennacchio, L. A., Bickmore, W., Dean, A., Nobrega, M. A., & Bejerano, G. J. N. R. G. (2013b). Enhancers: five essential questions. *14*(4), 288-295.
- Pepke, S., Wold, B., & Mortazavi, A. J. N. m. (2009). Computation for ChIP-seq and RNA-seq studies. *6*(11), S22-S32.
- Pervaiz, N., Kang, H., Bao, Y., & Abbasi, A. A. (2021). Molecular evolutionary analysis of human primary microcephaly genes. *BMC Ecology and Evolution*, *21*(1), 1-9. doi:<https://doi.org/10.1186/s12862-021-01801-0>
- Pettersen, E. F., Goddard, T. D., Huang, C. C., Couch, G. S., Greenblatt, D. M., Meng, E. C., & Ferrin, T. E. (2004). UCSF Chimera—a visualization system for exploratory research and analysis. *Journal of computational chemistry*, *25*(13), 1605-1612. doi:<https://doi.org/10.1002/jcc.20084>
- Pevny, L. H., & Lovell-Badge, R. (1997). Sox genes find their feet. *Current opinion in genetics & development*, *7*(3), 338-344. doi:[https://doi.org/10.1016/S0959-437X\(97\)80147-5](https://doi.org/10.1016/S0959-437X(97)80147-5)

- Pevny, L. H., & Nicolis, S. K. (2010). SOX2 roles in neural stem cells. *The international journal of biochemistry & cell biology*, 42(3), 421-424. doi:<https://doi.org/10.1016/j.biocel.2009.08.018>
- Pevny, L. H., Nicolis, S. K. J. T. i. j. o. b., & biology, c. (2010). SOX2 roles in neural stem cells. 42(3), 421-424.
- Pollard, K. S., Salama, S. R., King, B., Kern, A. D., Dreszer, T., Katzman, S., . . . Baertsch, R. (2006). Forces shaping the fastest evolving regions in the human genome. *PLoS genetics*, 2(10), e168. doi:<https://doi.org/10.1371/journal.pgen.0020168>
- Prabhakar, S., Noonan, J. P., Pääbo, S., & Rubin, E. M. (2006). Accelerated evolution of conserved noncoding sequences in humans. *science*, 314(5800), 786-786. doi:<https://doi.org/10.1126/science.1130738>
- Prescott, S. L., Srinivasan, R., Marchetto, M. C., Grishina, I., Narvaiza, I., Selleri, L., . . . Wsocka, J. (2015). Enhancer divergence and cis-regulatory evolution in the human and chimp neural crest. *Cell*, 163(1), 68-83. doi:<https://doi.org/10.1016/j.cell.2015.08.036>
- Price, D. J., & Brooks III, C. L. (2004). A modified TIP3P water potential for simulation with Ewald summation. *The Journal of chemical physics*, 121(20), 10096-10103.
- Prud'Homme, B., Gompel, N., Rokas, A., Kassner, V. A., Williams, T. M., Yeh, S.-D., . . . Carroll, S. B. (2006). Repeated morphological evolution through cis-regulatory changes in a pleiotropic gene. *Nature*, 440(7087), 1050-1053.
- Prüfer, K., Racimo, F., Patterson, N., Jay, F., Sankararaman, S., Sawyer, S., . . . De Filippo, C. (2014). The complete genome sequence of a Neanderthal from the Altai Mountains. *Nature*, 505(7481), 43-49. doi:<https://doi.org/10.1038/nature12886>
- Reis, L. M., Tyler, R. C., Schneider, A., Bardakjian, T., & Semina, E. V. J. M. v. (2010). Examination of SOX2 in variable ocular conditions identifies a recurrent deletion in microphthalmia and lack of mutations in other phenotypes. 16, 768.
- Reményi, A., Lins, K., Nissen, L. J., Reinbold, R., Schöler, H. R., & Wilmanns, M. (2003). Crystal structure of a POU/HMG/DNA ternary complex suggests differential assembly of Oct4 and SOX2 on two enhancers. *Genes & development*, 17(16), 2048-2059. doi:

---

<https://doi.org/10.1101/gad.269303>

- Richter, D., Grün, R., Joannes-Boyau, R., Steele, T. E., Amani, F., Rué, M., . . . Ben-Ncer, A. (2017). The age of the hominin fossils from Jebel Irhoud, Morocco, and the origins of the Middle Stone Age. *Nature*, *546*(7657), 293-296.
- Roe, D. R., & Cheatham III, T. E. (2013). PTRAJ and CPPTRAJ: software for processing and analysis of molecular dynamics trajectory data. *Journal of chemical theory and computation*, *9*(7), 3084-3095.
- Sagendorf, J. M., Berman, H. M., & Rohs, R. (2017). DNAProDB: an interactive tool for structural analysis of DNA–protein complexes. *Nucleic acids research*, *45*(W1), W89-W97. doi:<https://doi.org/10.1093/nar/gkx272>
- Salomon-Ferrer, R., Case, D. A., & Walker, R. C. (2013). An overview of the Amber biomolecular simulation package. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, *3*(2), 198-210.
- Scacheri, C. A., & Scacheri, P. C. (2015). Mutations in the non-coding genome. *Current opinion in pediatrics*, *27*(6), 659. doi:<https://doi.org/10.1097/MOP.0000000000000283>
- Scaffidi, P., & Bianchi, M. E. (2001). Spatially precise DNA bending is an essential activity of the SOX2 transcription factor. *Journal of Biological Chemistry*, *276*(50), 47296-47302. doi:<https://doi.org/10.1074/jbc.m107619200>
- Schaefer, T., Steiner, R., & Lengerke, C. J. I. J. o. M. S. (2020). SOX2 and p53 expression control converges in PI3K/AKT signaling with versatile implications for stemness and cancer. *21*(14), 4902.
- Schepers, G. E., Teasdale, R. D., & Koopman, P. J. D. c. (2002). Twenty pairs of sox: extent, homology, and nomenclature of the mouse and human sox transcription factor gene families. *3*(2), 167-170.
- Schilham, M. W., Oosterwegel, M. A., Moerer, P., Ya, J., de Boer, P. A., van de Wetering, M., . . . Cumano, A. (1996a). Defects in cardiac outflow tract formation and pro-B-lymphocyte expansion in mice lacking Sox-4. *Nature*, *380*(6576), 711-714. doi:<https://doi.org/10.1038/380711a0>
- Schilham, M. W., Oosterwegel, M. A., Moerer, P., Ya, J., de Boer, P. A., van de Wetering, M., . . . Cumano, A. J. N. (1996b). Defects in cardiac outflow tract formation and pro-B-lymphocyte expansion in mice lacking Sox-4. *380*(6576), 711-714.

- Seeliger, D., & De Groot, B. L. (2010). Conformational transitions upon ligand binding: holo-structure prediction from apo conformations. *PLoS computational biology*, 6(1), e1000634. doi: <https://doi.org/10.1371/journal.pcbi.1000634>
- Senger, K., Armstrong, G. W., Rowell, W. J., Kwan, J. M., Markstein, M., & Levine, M. J. M. c. (2004). Immunity regulatory DNAs share common organizational features in *Drosophila*. *13*(1), 19-32.
- Shashikant, C. S., Bolanowsky, S. A., Anand, S., & Anderson, S. M. (2007). Comparison of diverged Hoxc8 early enhancer activities reveals modification of regulatory interactions at conserved cis-acting elements. *Journal of Experimental Zoology Part B: Molecular and Developmental Evolution*, 308(3), 242-249.
- Shlyueva, D., Stampfel, G., & Stark, A. J. N. R. G. (2014). Transcriptional enhancers: from properties to genome-wide predictions. *15*(4), 272-286.
- Sinclair, A. H., Berta, P., Palmer, M. S., Hawkins, J. R., Griffiths, B. L., Smith, M. J., . . . Goodfellow, P. N. J. N. (1990). A gene from the human sex-determining region encodes a protein with homology to a conserved DNA-binding motif. *346*(6281), 240-244.
- Sisodiya, S. M., Ragge, N. K., Cavalleri, G. L., Hever, A., Lorenz, B., Schneider, A., . . . Thompson, P. J. (2006). Role of SOX2 mutations in human hippocampal malformations and epilepsy. *Epilepsia*, 47(3), 534-542. doi:<https://doi.org/10.1111/j.1528-1167.2006.00464.x>
- Skene, P.J. and Henikoff, S., 2017. An efficient targeted nuclease strategy for high-resolution mapping of DNA binding sites. *elife*, 6, p.e21856.
- Spitz, F., & Furlong, E. E. (2012). Transcription factors: from enhancer binding to developmental control. *Nature reviews genetics*, 13(9), 613-626.
- Südbeck, P., Schmitz, M. L., Baeuerle, P. A., & Scherer, G. J. N. g. (1996). Sex reversal by loss of the C-terminal transactivation domain of human SOX9. *13*(2), 230-232.
- Suleman, M., Yousafi, Q., Ali, J., Ali, S. S., Hussain, Z., Ali, S., . . . Khan, A. (2021). Bioinformatics analysis of the differences in the binding profile of the wild-type

- and mutants of the SARS-CoV-2 spike protein variants with the ACE2 receptor. *Computers in Biology and Medicine*, *138*, 104936.
- Sun, F.-L., & Elgin, S. C. J. C. (1999). Putting boundaries on silence. *99*(5), 459-462.
- Takahashi, A., Takahashi, K., Ueda, R., & Takano-Shimizu, T. (2007). Natural variation of ebony gene controlling thoracic pigmentation in *Drosophila melanogaster*. *Genetics*, *177*(2), 1233-1237. doi:<https://doi.org/10.1534/genetics.107.075283>
- Teare, J., Islam, R., Flanagan, R., Gallagher, S., Davies, M., & Grabau, C. (1997). Measurement of nucleic acid concentrations using the DyNA Quant™ and the GeneQuant™. *Biotechniques*, *22*(6), 1170-1174.
- Tsai, J., Taylor, R., Chothia, C., & Gerstein, M. (1999). The packing density in proteins: standard radii and volumes. *Journal of molecular biology*, *290*(1), 253-266. doi:<https://doi.org/10.1006/jmbi.1999.2829>
- Tycko, J., Wainberg, M., Marinov, G. K., Ursu, O., Hess, G. T., Ego, B. K., . . . Spees, K. (2019). Mitigation of off-target toxicity in CRISPR-Cas9 screens for essential non-coding elements. *Nature communications*, *10*(1), 1-14.
- Uchikawa, M., Kamachi, Y., & Kondoh, H. (1999). Two distinct subgroups of Group B Sox genes for transcriptional activators and repressors: their expression during embryonic organogenesis of the chicken. *Mechanisms of development*, *84*(1-2), 103-120.
- ul Qamar, M. T., Maryam, A., Muneer, I., Xing, F., Ashfaq, U. A., Khan, F. A., . . . Rauf, S. A. (2019). Computational screening of medicinal plant phytochemicals to discover potent pan-serotype inhibitors against dengue virus. *Scientific reports*, *9*(1), 1-16.
- Umarov, R. K., & Solovyev, V. V. (2017). Recognition of prokaryotic and eukaryotic promoters using convolutional deep learning neural networks. *PloS one*, *12*(2), e0171410.
- Umarov, R. K., & Solovyev, V. V. J. P. o. (2017). Recognition of prokaryotic and eukaryotic promoters using convolutional deep learning neural networks. *12*(2), e0171410.
- Uwanogho, D., Rex, M., Cartwright, E. J., Pearl, G., Healy, C., Scotting, P. J., & Sharpe, P. T. J. M. o. d. (1995). Embryonic expression of the chicken SOX2,

- Sox3 and Sox11 genes suggests an interactive role in neuronal development. *49*(1-2), 23-36.
- van Dijk, M., & Bonvin, A. M. (2009). 3D-DART: a DNA structure modelling server. *Nucleic acids research*, *37*(suppl\_2), W235-W239.
- Van Zundert, G., Rodrigues, J., Trellet, M., Schmitz, C., Kastritis, P., Karaca, E., . . . Bonvin, A. (2016). The HADDOCK2. 2 web server: user-friendly integrative modeling of biomolecular complexes. *Journal of molecular biology*, *428*(4), 720-725. doi:<https://doi.org/10.1016/j.jmb.2015.09.014>
- Vernot, B., & Akey, J. M. (2015). Complex history of admixture between modern humans and Neandertals. *The American Journal of Human Genetics*, *96*(3), 448-453. doi:<https://doi.org/10.1016/j.ajhg.2015.01.006>
- Visel, A., Blow, M. J., Li, Z., Zhang, T., Akiyama, J. A., Holt, A., . . . Chen, F. J. N. (2009). ChIP-seq accurately predicts tissue-specific activity of enhancers. *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*(Database), D88-D92. doi: 10.1093/nar/gkl822
- Visel, A., Taher, L., Girgis, H., May, D., Golonzhka, O., Hoch, R. V., . . . Blow, M. J. (2013). A high-resolution enhancer atlas of the developing telencephalon. *Cell*, *152*(4), 895-908. doi:<https://doi.org/10.1016/j.cell.2012.12.041>
- Watowich, S. J., Meyer, E. S., Hagstrom, R., & Josephs, R. (1988). A stable, rapidly converging conjugate gradient method for energy minimization. *Journal of computational chemistry*, *9*(6), 650-661.
- White, M. D., Angiolini, J. F., Alvarez, Y. D., Kaur, G., Zhao, Z. W., Mocskos, E., . . . Plachta, N. (2016). Long-lived binding of SOX2 to DNA predicts cell fate in the four-cell mouse embryo. *Cell*, *165*(1), 75-87. doi:<https://doi.org/10.1016/j.cell.2016.02.032>
- + Williams, D. C., Cai, M., & Clore, G. M. (2004). Molecular basis for synergistic transcriptional activation by Oct1 and SOX2 revealed from the solution structure of the 42-kDa Oct1·SOX2·Hoxb1-DNA ternary transcription factor complex. *Journal of Biological Chemistry*, *279*(2), 1449-1457. doi:<https://doi.org/10.1074/jbc.M309790200>

- Williams, S. A., Middleton, E. R., Villamil, C. I., & Shattuck, M. R. (2016). Vertebral numbers and human evolution. *American Journal of Physical Anthropology*, *159*, 19-36.
- Wilson, M., Koopman, P. J. C. o. i. g., & development. (2002). Matching SOX: partner proteins and co-factors of the SOX family of transcriptional regulators. *12*(4), 441-446.
- Wood, H. B., & Episkopou, V. J. M. o. d. (1999). Comparative expression of the mouse Sox1, SOX2 and Sox3 genes from pre-gastrulation to early somite stages. *86*(1-2), 197-201.
- Yan, S. M., & McCoy, R. C. (2020). Archaic hominin genomics provides a window into gene expression evolution. *Current opinion in genetics & development*, *62*, 44-49. doi:<https://doi.org/10.1016/j.gde.2020.05.014>
- Yousaf, A., Sohail Raza, M., & Ali Abbasi, A. (2015). The evolution of bony vertebrate enhancers at odds with their coding sequence landscape. *Genome biology and evolution*, *7*(8), 2333-2343. doi:<https://doi.org/10.1093/gbe/evv146>
- Yuan, H., Corbi, N., Basilico, C., Dailey, L. J. G., & development. (1995). Developmental-specific activity of the FGF-4 enhancer requires the synergistic action of SOX2 and Oct-3. *9*(21), 2635-2645.
- Zehra, R., & Abbasi, A. A. (2018). Homo sapiens-specific binding site variants within brain exclusive enhancers are subject to accelerated divergence across human population. *Genome biology and evolution*, *10*(3), 956-966. doi:<https://doi.org/10.1093/gbe/evy052>
- Zhang, C., Vasmatzis, G., Cornette, J. L., & DeLisi, C. J. J. o. m. b. (1997). Determination of atomic desolvation energies from the structures of crystallized proteins. *267*(3), 707-726.
- Zuin, J., Dixon, J. R., van der Reijden, M. I., Ye, Z., Kolovos, P., Brouwer, R. W., . . . van IJcken, W. F. J. P. o. t. N. A. o. S. (2014). Cohesin and CTCF differentially affect chromatin architecture and gene expression in human cells. *111*(3), 996-1001.



**Appendices**

**Appendix 1**

**Solutions Composition**

Solution A	0.32M sucrose
	10Mm Tris-HCL (PH7.5)
	1% (v/v) Triton X-100
Solution B	10Mm Tris (PH 7.5)
	400Mm NaCl
	2mM EDTA (PH 8.0)
Solution C	400µl Phenol
Solution D	Isoamyl alcohol 1 volume
	Chloroform 24 volumes

**Appendix 2**

**10X TBE Buffer Composition**

**Reagent Quantity**

Tris Base	54 grams
Boric Acid	27.5 grams
EDTA	3.65 grams
Distilled water up to	500 ml

### **Appendix 3**

#### **LB media composition**

##### **Reagent Quantity**

Bactotrypton	5 grams
Yeast Extract	2.5 grams
Sodium Chloride	5 grams
PH	7.5 (with the help of 2N NaOH)
Total Volume 500 ml (with autoclaved distilled water).	

### **Appendix-4**

#### **Buffer Composition**

P-1 Buffer 15Mm Tris Chloride 10Mm EDTA 10µg/ml RNase 200ml distilled water

P-2 Buffer 0.2% NaOH 1ml 1%SDS 1ml

P-3 Buffer 3M Potassium acetate (PH 5.5)

### **Appendix 5**

#### **IPTG Composition**

IPTG (100mM) 2.5µl

Kanamycin 2 µl

Media 2ml

Incubate at 37°C for 10 minutes.

### **Appendix 6**

#### **Protein Extracting Dye Composition.**

20% SDS 2 ml

1M Tris HCl (6.8) 2 ml

Glycerol 1.2 ml

Water up to 10 ml.

Bromophenol Blue 10µl

**Appendix 7**

**SDS-PAGE Composition**

**Separating Gel 12%**

Total Volume	32 ml
Water	6.7 ml
30% Acrylamide	12.8 ml
1M Tris (PH 8.8)	12 ml
20% SDS	0.16 ml
10% APS	0.32 ml
TEMED	0.32 ml

**Stacking Gel 6%**

Total Volume	10 ml
Water	6.6 ml
30% Acrylamide	2 ml
1M tris (PH 6.8)	1.25 ml
20% SDS	0.05ml
10% APS	0.1 ml
TEMED	10 $\mu$ l

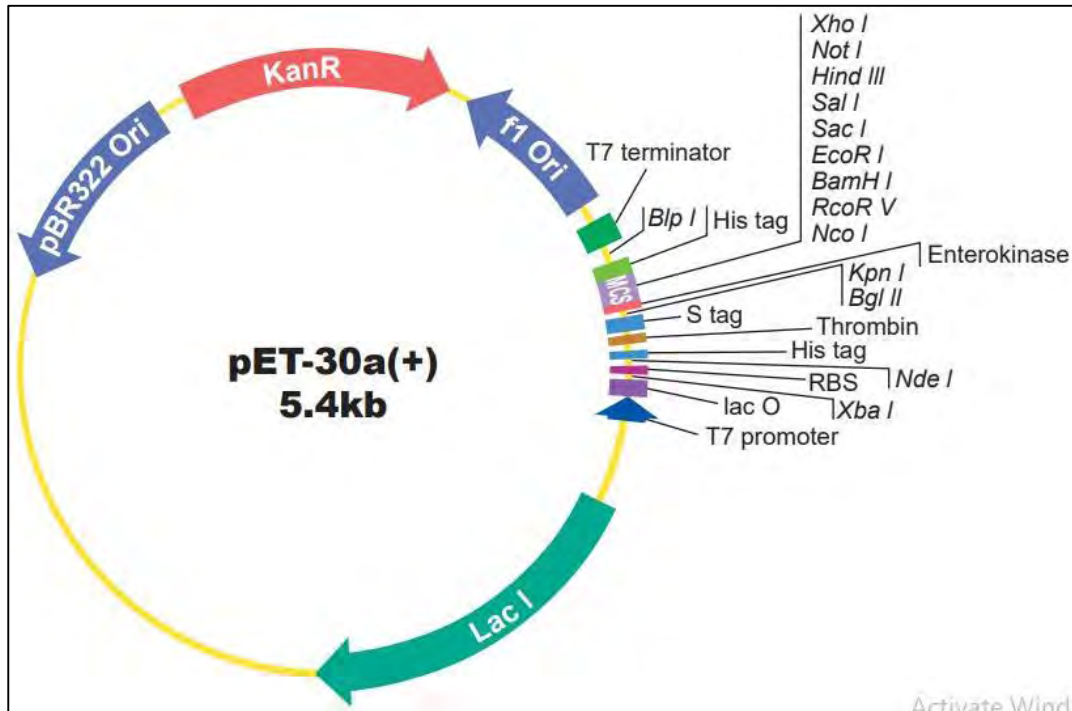
**Appendix 8**

**>PCR\_SOX2\_CCSD**

ATGTACAACATGATGGAGACGGAGCTGAAGCCGCCGGGCCCCGCAGCAA  
CTTCGGGGGGCGGGCGGCGGCAACTCCACCGCGGGCGGCGGCCCCGGCGGCAA  
CCAGAAAAACAGCCCGGACCGCGTCAAGCGGCCCATGAATGCCTTCATGG  
TGTGGTCCCGCGGGCAGCGGCGCAAGATGGCCCAGGAGAACCCCAAGAT  
GCACAACCTCGGAGATCAGCAAGCGCCTGGGGCGCCGAGTGGAACTTTTGT  
CGGAGACGGAGAAGCGGCCGTTTCATCGACGAGGCTAAGCGGCTGCGAGC  
GCTGCACATGAAGGAGCACCCGGATTATAAATACCGGCCCCGGCGGAAA  
ACCAAGACGCTCATGAAGAAGGATAAGTACACGCTGCCCGGCGGGCTGCT  
GGCCCCCGGCGGCAATAGCATGGCGAGCGGGGTCGGGGTGGGCGCCGGC  
CTGGGCGCGGGCGTGAACCAGCGCATGGACAGTTACGCGCACATGAACG  
GCTGGAGCAACGGCAGCTACAGCATGATGCAGGACCAGCTGGGCTACCC  
GCAGCACCCGGGCCTCAATGCGCACGGCGCAGCGCAGATGCAGCCCATGC  
ACCGCTACGACGTGAGCGCCCTGCAGTACAACCTCCATGACCAGCTCGCAG  
ACCTACATGAACGGCTCGCCACCTACAGCATGTCCTACTCGCAGCAGGG  
CACCCCTGGCATGGCTCTTGGCTCCATGGGTTCGGTGGTCAAGTCCGAGGC  
CAGCTCCAGCCCCCTGTGGTTACCTCTTCCCTCCCACTCCAGGGCGCCCTG  
CCAGGCCGGGGACCTCCGGGACATGATCAGCATGTATCTCCCCGGCGCCG  
AGGTGCCGGAACCCGCCGCCCCAGCAGACTTCACATGTCCCAGCACTAC  
CAGAGCGGCCCGGTGCCCGGCACGGCCATTAACGGCACACTGCCCTCTC  
ACACATGT

Appendix 9

Circular Map of Pet-30a (+) Vector





AATTACATTCCCAACCGCGTGGCACAACAACCTGGCGGGCAAACAGTCGTT  
GCTGATTGGCGTTGCCACCTCCAGTCTGGC

CCTGCACGCGCCGTCGCAAATTGTGCGGGCGATTAAATCTCGCGCCGATC  
AACTGGGTGCCAGCGTGGTGGTGTGCGATGG

TAGAACGAAGCGGCGTTCGAAGCCTGTAAAGCGGCGGTGCACAATCTTCTC  
GCGCAACGCGTCAGTGGGCTGATCATTAAC

TATCCGCTGGATGACCAGGATGCCATTGCTGTGGAAGCTGCCTGCACTAA  
TGTTCCGGCGTTATTTCTTGATGTCTCTGA

CCAGACACCCATCAACAGTATTATTTTCTCCCATGAAGACGGTACGCGACT  
GGGCGTGGAGCATCTGGTTCGCATTGGGTC

ACCAGCAAATCGCGCTGTTAGCGGGCCCATTAAGTTCTGTCTCGGCGCGTC  
TGCGTCTGGCTGGCTGGCATAAATATCTC

ACTCGCAATCAAATTCAGCCGATAGCGGAACGGGAAGGCGACTGGAGTG  
CCATGTCCGGTTTTCAACAAACCATGCAAAT

GCTGAATGAGGGCATCGTTCCCACTGCGATGCTGGTTGCCAACGATCAGA  
TGCGCTGGGCGCAATGCGCGCCATTACCG

AGTCCGGGCTGCGCGTTGGTTCGGACATCTCGGTAGTGGGATACGACGAT  
ACCGAAGACAGCTCATGTTATATCCCGCCG

TTAACCACCATCAAACAGGATTTTCGCCTGCTGGGGCAAACCAGCGTGGA  
CCGCTTGCTGCAACTCTCTCAGGGCCAGGC

GGTGAAGGGCAATCAGCTGTTGCCCGTCTCACTGGTGAAAAGAAAACCA  
CCCTGGCGCCCAATACGCAAACCGCCTCTC

CCCGCGCGTTGGCCGATTCATTAATGCAGCTGGCACGACAGGTTTCCCGA  
CTGGAAAGCGGGCAGTGAGCGCAACGCAAT

TAATGTAAGTTAGCTCACTCATTAGGCACCGGGATCTCGACCGATGCCCTT  
GAGAGCCTTCAACCCAGTCAGCTCCTTCC

GGTGGGCGCGGGGCATGACTATCGTCGCCGCACTTATGACTGTCTTCTTTA  
TCATGCAACTCGTAGGACAGGTGCCGGCA

GCGCTCTGGGTCATTTTCGGCGAGGACCGCTTTCGCTGGAGCGCGACGAT  
GATCGGCCTGTCGCTTGCGGTATTCGGAAT

CTTGACGCCCCTCGCTCAAGCCTTCGTCACTGGTCCCGCCACCAAACGTTT  
CGGCGAGAAGCAGGCCATTATCGCCGGCA

TGGCGGCCCCACGGGTGCGCATGATCGTGCTCCTGTCGTTGAGGACCCGG  
CTAGGCTGGCGGGGTTGCCTTACTGGTTAG

CAGAATGAATCACCGATACGCGAGCGAACGTGAAGCGACTGCTGCTGCAA  
AACGTCTGCGACCTGAGCAACAACATGAAT

GGTCTTCGGTTTCCGTGTTTCGTAAAGTCTGGAAACGCGGAAGTCAGCGCC  
CTGCACCATTATGTTCCGGATCTGCATCG

CAGGATGCTGCTGGCTACCCTGTGGAACACCTACATCTGTATTAACGAAG  
CGCTGGCATTGACCCTGAGTGATTTTTCTC

TGGTCCCGCCGCATCCATAACCGCCAGTTGTTTACCCTCACAACGTTCCAGT  
AACCGGGCATGTTCATCATCAGTAACCCG

TATCGTGAGCATCCTCTCTCGTTTCATCGGTATCATTACCCCATGAACAG  
AAATCCCCCTTACACGGAGGCATCAGTGA

CCAAACAGGAAAAAACCGCCCTTAACATGGCCCGCTTTATCAGAAGCCAG  
ACATTAACGCTTCTGGAGAAACTCAACGAG

CTGGACGCGGATGAACAGGCAGACATCTGTGAATCGCTTCACGACCACGC  
TGATGAGCTTTACCGCAGCTGCCTCGCGCG

TTTCGGTGATGACGGTGAAAACCTCTGACACATGCAGCTCCCGGAGACGG  
TCACAGCTTGCTGTGTAAGCGGATGCCGGGA

GCAGACAAGCCCGTCAGGGCGCGTCAGCGGGTGTGGCGGGTGTGCGGGGC  
GCAGCCATGACCCAGTCACGTAGCGATAGC

GGAGTGATACTGGCTTAACTATGCGGCATCAGAGCAGATTGTACTGAGA  
GTGCACCATATATGCGGTGTGAAATACCGC

ACAGATGCGTAAGGAGAAAATACCGCATCAGGCGCTCTTCCGCTTCCTCG  
CTCACTGACTCGCTGCGCTCGGTGCTTCGG



CTGCGGCGAGCGGTATCAGCTCACTCAAAGGCGGTAATACGGTTATCCAC  
AGAATCAGGGGATAACGCAGGAAAGAACAT

GTGAGCAAAAGGCCAGCAAAAGGCCAGGAACCGTAAAAAGGCCGCGTTG  
CTGGCGTTTTTCCATAGGCTCCGCCCCCTG

ACGAGCATCACAAAATCGACGCTCAAGTCAGAGGTGGCGAAACCCGAC  
AGGACTATAAAGATAACCAGGCGTTTCCCCCT

GGAAGCTCCCTCGTGCGCTCTCCTGTTCCGACCCTGCCGCTTACCGGATAC  
CTGTCCGCCTTTCTCCCTTCGGGAAGCGT

GGCGCTTTCTCATAGCTCACGCTGTAGGTATCTCAGTTCGGTGTAGGTCGT  
TCGCTCCAAGCTGGGCTGTGTGCACGAAC

CCCCCGTTCAGCCCGACCGCTGCGCCTTATCCGGTAACTATCGTCTTGAGT  
CCAACCCGGTAAGACACGACTTATCGCCA

CTGGCAGCAGCCACTGGTAACAGGATTAGCAGAGCGAGGTATGTAGGCG  
GTGCTACAGAGTTCTTGAAGTGGTGGCCTAA

CTACGGCTACACTAGAAGGACAGTATTTGGTATCTGCGCTCTGCTGAAGC  
CAGTTACCTTCGGAAAAAGAGTTGGTAGCT

CTTGATCCGGCAAACAAACCACCGCTGGTAGCGGTGGTTTTTTTTGTTTGCA  
AGCAGCAGATTACGCGCAGAAAAAAAGGA

TCTCAAGAAGATCCTTTGATCTTTTCTACGGGGTCTGACGCTCAGTGGAAC  
GAAAACACGTTAAGGGATTTTGGTCAT

GAACAATAAAACTGTCTGCTTACATAAACAGTAATACAAGGGGTGTTATG  
AGCCATATTCAACGGGAAACGTCTTGCTCT

AGGCCGCGATTAAATTCCAACATGGATGCTGATTTATATGGGTATAAATG  
GGCTCGCGATAATGTCGGGCAATCAGGTGC

GACAATCTATCGATTGTATGGGAAGCCCGATGCGCCAGAGTTGTTTCTGA  
AACATGGCAAAGGTAGCGTTGCCAATGATG

TTACAGATGAGATGGTCAGACTAAACTGGCTGACGGAATTTATGCCTCTTC  
CGACCATCAAGCATTTTATCCGTACTCCT


GATGATGCATGGTTACTCACCCTGCGATCCCCGGGAAAACAGCATTCCA  
GGTATTAGAAGAATATCCTGATTCAGGTGA  
AAATATTGTTGATGCGCTGGCAGTGTTCCCTGCGCCGGTTGCATTTCGATTCC  
TGTTTGTAATTGTCCTTTTAACAGCGATC  
GCGTATTTTCGTCTCGCTCAGGCGCAATCACGAATGAATAACGGTTTGGTTG  
ATGCGAGTGATTTTGATGACGAGCGTAAT  
GGCTGGCCTGTTGAACAAGTCTGGAAAGAAATGCATAAACTTTTGCCATT  
CTCACCGGATTCAGTCGTCACCTCATGGTGA  
TTTCTCACTTGATAACCTTATTTTTGACGAGGGGAAATTAATAGGTTGTAT  
TGATGTTGGACGAGTCGGAATCGCAGACC  
GATACCAGGATCTTGCCATCCTATGGAAGTGCCTCGGTGAGTTTTCTCCTT  
CATTACAGAAACGGCTTTTTCAAAAATAT  
GGTATTGATAATCCTGATATGAATAAATTGCAGTTTCATTTGATGCTCGAT  
GAGTTTTTCTAAGAATTAATTCATGAGCG  
GATACATATTTGAATGTATTTAGAAAAATAAACAAATAGGGGTTCCGCGC  
ACATTTCCCCGAAAAGTGCCACCTGAAATT  
GTAAACGTTAATATTTTGTTAAAATTCGCGTTAAATTTTTGTTAAATCAGC  
TCATTTTTTAACCAATAGGCCGAAATCGG  
CAAAATCCCTTATAAATCAAAGAATAGACCGAGATAGGGTTGAGTGTTG  
TTCCAGTTTGAACAAGAGTCCACTATTAA  
AGAACGTGGACTCCAACGTCAAAGGGCGAAAAACCGTCTATCAGGGCGA  
TGGCCCACTACGTGAACCATCACCTAATCA  
AGTTTTTTGGGGTCGAGGTGCCGTAAAGCACTAAATCGGAACCCTAAAGG  
GAGCCCCCGATTTAGAGCTTGACGGGGAAA  
GCCGGCGAACGTGGCGAGAAAGGAAGGGAAGAAAGCGAAAGGAGCGGG  
CGCTAGGGCGCTGGCAAGTGTAGCGGTCACGC  
TGCGCGTAACCACCACACCCGCCGCGCTTAATGCGCCGCTACAGGGCGCG  
TCCCATTGCGCA

RESEARCH ARTICLE

Open Access



# Evolutionary relevance of single nucleotide variants within the forebrain exclusive human accelerated enhancer regions

Hizran Khatoon<sup>1</sup>, Rabail Zehra Raza<sup>2</sup>, Shoaib Saleem<sup>1</sup>, Fatima Batool<sup>1</sup>, Saba Arshad<sup>1</sup>, Muhammad Abrar<sup>1</sup>, Shahid Ali<sup>3</sup>, Irfan Hussain<sup>1</sup>, Neil H. Shubin<sup>3\*</sup> and Amir Ali Abbasi<sup>1\*</sup> 

## Abstract

**Background** Human accelerated regions (HARs) are short conserved genomic sequences that have acquired significantly more nucleotide substitutions than expected in the human lineage after divergence from chimpanzees. The fast evolution of HARs may reflect their roles in the origin of human-specific traits. A recent study has reported positively-selected single nucleotide variants (SNVs) within brain-exclusive human accelerated enhancers (BE-HAEs) hs1210 (forebrain), hs563 (hindbrain) and hs304 (midbrain/forebrain). By including data from archaic hominins, these SNVs were shown to be *Homo sapiens*-specific, residing within transcriptional factors binding sites (TFBSs) for SOX2 (hs1210), RUNX1/3 (hs563), and FOS/JUND (hs304). Although these findings suggest that the predicted modifications in TFBSs may have some role in present-day brain structure, work is required to verify the extent to which these changes translate into functional variation.

**Results** To start to fill this gap, we investigate the SOX2 SNV, with both forebrain expression and strong signal of positive selection in humans. We demonstrate that the HMG box of SOX2 binds *in vitro* with *Homo sapiens*-specific derived A-allele and ancestral T-allele carrying DNA sites in BE-HAE hs1210. Molecular docking and simulation analysis indicated highly favourable binding of HMG box with derived A-allele containing DNA site when compared to site carrying ancestral T-allele.

**Conclusion** These results suggest that adoptive changes in TF affinity within BE-HAE hs1210 and other HAR enhancers in the evolutionary history of *Homo sapiens* might have brought about changes in gene expression patterns and have functional consequences on forebrain formation and evolution.

**Methods** The present study employ electrophoretic mobility shift assays (EMSA) and molecular docking and molecular dynamics simulations approaches.

**Keywords** Evolution, Enhancers, Human accelerated regions, Forebrain, Archaic hominins, TFBS, SNVs, SOX2, HMG box, EMSA

\*Correspondence:

Neil H. Shubin  
nshubin@uchicago.edu  
Amir Ali Abbasi  
abbasiam@qau.edu.pk

Full list of author information is available at the end of the article



© The Author(s) 2023, corrected publication 2023. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

Turnitin Originality Report

Elucidating the events of transcription factor binding in human brain enhancers  
Hizran Khatoon .

by



From PhD (PhD DRSSL)

- Processed on 21-Aug-2023 08:14 PKT
- ID: 2148642772
- Word Count: 20624

Similarity Index

15%

Similarity by Source

Internet Sources:

12%

Publications:

9%

Student Papers:

4%

*[Handwritten Signature]*  
**SUPERVISOR**  
 National Centre for Bioinformatics  
 Quaid-i-Azam University, Islamabad

**sources:**

- 1 1% match (Internet from 10-Dec-2022)  
[http://pr.hec.gov.pk/jspui/bitstream/123456789/11244/1/Rabail%20Zehra\\_Bioinfo\\_2019\\_QAU\\_PRR.pdf](http://pr.hec.gov.pk/jspui/bitstream/123456789/11244/1/Rabail%20Zehra_Bioinfo_2019_QAU_PRR.pdf)
- 2 1% match (Internet from 10-Feb-2023)  
[http://pr.hec.gov.pk/jspui/bitstream/123456789/2966/1/Rashid\\_Minhas\\_Bioinformatics\\_2016\\_QAU\\_02.03.2016.pdf](http://pr.hec.gov.pk/jspui/bitstream/123456789/2966/1/Rashid_Minhas_Bioinformatics_2016_QAU_02.03.2016.pdf)
- 3 1% match ()  
[Abbas Khan, Tauqir Zia, Muhammad Suleman, Taimoor Khan, Syed Shujait Ali, Aamir Ali Abbasi, Anwar Mohammad, Dong-Qing Wei. "Higher infectivity of the SARS-CoV-2 new variants is associated with K417N/T, E484K, and N501Y mutants: An insight from structural data", Journal of Cellular Physiology](#)
- 4 1% match (Internet from 17-Jul-2023)  
<https://pesquisa.bvsalud.org/bvsms/?lang=pt&q=au%3A%22Saleem%2C+Shoab%22>
- 5 1% match (Internet from 11-Jan-2023)  
[https://theses.cz/id/ohcvgo/PhD\\_Thesis-Morteza\\_Khabiri.pdf](https://theses.cz/id/ohcvgo/PhD_Thesis-Morteza_Khabiri.pdf)
- 6 1% match (student papers from 31-Oct-2021)  
[Submitted to J S S University on 2021-10-31](#)
- 7 1% match ()  
[Liu, Yu-Ru. "Function of Sox2 as a transcriptional repressor"](#)
- 8 1% match (Internet from 23-Jul-2022)  
<https://dash.harvard.edu/bitstream/handle/1/17467240/SARKAR-DISSERTATION-2015.pdf?isAllowed=y&sequence=10>
- 9 < 1% match ()  
[Shagufta Shafique, Saiid Rashid. "Structural basis of  \$\beta\$ TrCP1-associated GLI3 processing", Scientific Reports](#)
- 10 < 1% match ()