Brain Enhancers and their Role in Distinguishing Human CNS from that of Non-human Primates

By

Rabail Zehra

National Center for Bioinformatics

Faculty of Biological Sciences

Quaid-i-Azam University

Islamabad, Pakistan

2019

Brain Enhancers and their Role in Distinguishing Human CNS from that of Non-human Primates

By

Rabail Zehra

A thesis submitted in the partial fulfillment of

the requirements for the degree of

DOCTOR OF PHILOSOPHY

IN

BIOINFORMATICS

National Center for Bioinformatics

Faculty of Biological Sciences

Quaid-i-Azam University

Islamabad, Pakistan

2019

Author's Declaration

I, **Rabail Zehra,** hereby state that my Ph.D. thesis titled as "**Brain Enhancers and their Role in Distinguishing Human CNS from that of Non-human Primates"** is my own work 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 revoke my Ph.D. degree.

Name: Rabail Zehra

Date: January 28' 2019

Plagiarism Undertaking

I solemnly declare that research work presented in this thesis titled as "**Brain Enhancers and their Role in Distinguishing Human CNS from that of Non-human Primates**" is solely my own research and has been written completely by me with no significant contribution from any other person. Small contribution/help whenever taken has been duly acknowledged.

I understand the zero tolerance policy of the **Higher Education Commission, Pakistan** and **Quaid-i-Azam University Islamabad, Pakistan** towards plagiarism. Therefore, I, as an Author of the above titled thesis declare that no portion of my thesis has been plagiarized and any material used as reference is properly referred to/cited.

I undertake that if I am found guilty of any formal plagiarism in the above titled thesis even after the award of the Ph.D. degree, the University reserves the right to withdraw/revoke my Ph.D. degree. Also, HEC and the University will bear the right to publish my name on the HEC/University Website among names of the students who submitted plagiarized thesis.

Student/Author

Signature: ____________________

Name: Rabail Zehra

DEDICATION

To my mother, ***Atia Batool****, who is the reason I am getting this doctorate and to my father,* ***Ishfaq Hussain*** *(Late), who would have been so happy and proud to see me become a doctor finally.*

ACKNOWLEDGEMENTS

It would not be fair without naming HIM first before whom my head bows in humility and gratefulness. Who would have thought that this day would come, had HE not been there all along? I thank **ALLAH ALMIGHTY** for responding to my desperate prayers, for showing me light in the toughest of times and for making me able to achieve what little I have achieved. I thank HIM and I will never be able to thank HIM enough. Without HIS will and without HIM guiding the way, I would be left stranded on an empty path.

Praise be on the **Prophet Muhammad (P.B.U.H),** his household and his faithful companions who sacrificed so much for us. In our worldly pursuits, we often forget the peace and the order they brought in our lives. We owe it to them for bringing the moral compass, the urge to learn and the awareness to respect humanity. No words will ever be able to do justice to describe what they have done for us.

I would like to thank my teachers and my mentors from the very childhood to this day who made me who I am. I particularly would like to thank my supervisor and chairperson, **Dr. Amir Ali Abbasi**, who led this years' long commitment to fruition. I am forever inspired by his ideas, devotion and dedication. I learnt from him to think bigger and to aim higher. I am indebted to him for his kind input and healthy critique for this entire dissertation. I am greatly thankful to **Dr. Sajid Rashid**, former chairperson, National Center for Bioinformatics for being an outstanding leader of the NCB team and to the entire faculty and staff for being there for us in times of need, especially **Mr. Naseer Ahmed Raja**, our course coordinator. I also would like to extend my gratitude to Higher Education Commission, Pakistan, for its financial assistance.

I would like to thank my lab colleagues (too many to name), of whom I saw many leave and many enter. Each one of you has left an impact, but those who stayed with me till the end have a very kind place in my heart. I would like to extend my gratitude to **Nashaiman Pervaiz**, whom I turned to for my project discussions. Her input has been invaluable to my work and to my problems. I also would like to thank **Shahid Ali** and **Irfan Hussain**, for being the most supportive of colleagues and also for their conducive discussions on various projects.

I am forever indebted to my parents for providing me with the means to be able to get to this level. Thank you for believing in your daughters and for making them believe also that education is the only way to empowerment and independence. I owe this degree to my **Amma**, who was so into pursuing education that this entire journey happened on her insistence. I am thankful to my **Abbu** whose hard work and dedication to his children over the years will forever contribute to all what we will ever achieve in our lives. I am thankful to my sisters, **Farwa, Rimsha, Abeer** and my brother **Ayyan** who practically

raised my kid when I was not around. No words can thank them enough for their contribution in my life. I knew that they were as much worried as I was for my work and were as happier on my success as I was on getting this done. I owe this accomplishment to them.

As much as I owe it to my parents, I am indebted for life to my husband **Raza Abbas** whose unflinching support made it so easy for me that I hardly ever thought while working that I am married and have a household to tend to. I am thankful for his generosity and for being the most appreciative and kindest of mentors. As much as he helped me in getting this done, he remains to be forever excited about my academic journey. I thank him for being such a significant part of this journey which would not have been possible without him.

And in the end, I would like to make the most important acknowledgement to my little man, my bundle of endless joy, my son, **Hasnain**. Thank you for putting up with Mama's absence and for being such a good boy. The amount of guilt that I had to suppress for leaving you behind only pushed me to work even harder. This Ph.D. has been for you and because of you.

Rabail

ABSTRACT

BACKROUND: Human sequence acceleration has been reported to have revamped the status of present-day humans over the course of evolution and has immensely contributed to their efficient adaptation to do highly complicated assessments. Human accelerated DNA fragments are those bits of the genome that have experienced frequent sequential changes after the human-chimp split despite being strongly conserved among mammals. Previous studies have indicated that many such accelerated genomic segments happen to harbor cis-regulatory elements, among which enhancers take up the most portion. Enhancers make up the distal category of cis-regulatory elements that could reside many kilobases away from their target genes and contribute in initiation of cell specific gene expression. Recent findings have also brought to our notice that coding region mutations shared with archaic humans were followed by substitutions in regulatory elements that were *Homo sapien*-unique and hence attributed to anatomically profound modern human traits. Following this deduction, we opted for brain that is the most profoundly adapted organ in the present-day human anatomy, characterizing them as the most cognitively advanced species. We focussed on acceleration of enhancers that express solely in the brain region. With respect to that, craniofacial development due to an increased brain size during the course of primate evolution has also garnered immense attention over the past many years. The relevance of this increase in brain size and its direct impact in formulating the facial mechanics of humans, both archaic and modern, has left many questions unanswered. Climate is one leading factor that imposed evolutionary constraints over the human facial dynamics. While observing such wide variety of facial forms in the present-day human population, it becomes evidently intriguing to probe into genetic factors that might have given in to the forces of natural selection. With the advent of genome wide association studies, we now have a decent collection of single nucleotide polymorphisms that are associated with various facial features. We took nasal morphology as our case study for being nature's profound conditioning system in the human body. By keeping out-of-Africa ancient migrations in mind, we observe a drastic climatic shift from an extremely hot-humid environment of Africa to relatively temperate regions of Asia and extremely cold Europe. Following the pattern of nasal variation on these lines, the aim of our study ensures a link between nasal adaptations to climatic change as wide-bulbous noses are significant features of hot-humid climate and narrower-taller noses represent a much colder climate.

RESULTS: This study relied on empirically confirmed brain exclusive enhancers to avoid any misjudgments about their regulatory status and categorized among them a subset of enhancers with an exceptionally accelerated rate of lineage specific divergence in humans. Among these accelerated enhancers, we found an assorted set of 13 distinct transcription factor binding sites were located that possessed unique existence in humans. 3/13 such sites belonging to transcription factors SOX2, RUNX1/3 and FOS/JUND possessed single nucleotide variants that made them unique to *H. sapiens* upon comparisons with Neanderthal and Denisovan orthologous sequences. These variants modifying the binding sites in modern human lineage were further substantiated as single nucleotide polymorphisms via exploiting 1000 Genomes Project Phase III data. Long range haplotype (LRH) based tests laid out evidence of positive selection to be governing in African population on two of the modern human motif modifying alleles with strongest results for SOX2 binding site. For nasal phenotype assessment on the basis of genetic variation, we gathered a set of SNPs from six GWAS studies till date, each associated with a particular nasal feature and applied tests so as to determine the pattern of contrasting selection over alleles in regions of climatic opposites. We incorporated 2504 individuals' data from 1000 Genomes Project Phase III. We observed 9 such SNPs that made strong cases of positive selection on either of their allelic variants (derived or ancestral). Among them, we also observed SNPs that conspicuously showed varying patterns of selection on either of the alleles in Africa (hot-humid climate) in comparison with four non-African populations (temperate or colder climates), hence, highlighting a climatically driven, contrasting patterns of divergence of alleles that favored a particular nasal phenotype.

CONCLUSION: Our study concludes that sequence divergence in the regulatory repertoire of modern humans underlie their vast phenotypic leverage over other species, brain being the crown of all such adaptations. We also concluded that *Homo sapien*-specific binding site variants in these enhancers are prone to accelerated divergence across the current-day human population and could be involving a functional advantage. We also gauged in this study that nasal type variation in different regions of the world are climatically driven. Our data also highlights the uniqueness of these substitutions, as majority of the human specific substitutions are not shared with Neanderthals and Denisovans. Also, the occurrence of these SNPs in non-coding part of the genome also points towards a new aspect in which cis-regulatory evolution could be playing a significant role in devising the nasal morphological mechanics of the present-day human population.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

BE-HAEs	Brain Exclusive Human Accelerated Regions
CHIP	Chromatin Immunoprecipitation
CREs	Cis-regulatory Regions
DAF	Derived Allele Frequency
DHSSs	DNAse I hypersensitivity Sites
DPE	Downstream Promoter Element
EHH	Extended Haplotype Homozygosity
FISH	Fluorescence *in-situ* Hybridization
GWAS	Genome Wide Association Studies
GXD	Genome Expression Database
HARs	Human Accelerated Regions
HMG	High Mobility Group
HMMs	Hidden Markov Models
INDELs	INsertions-DELetions
Inr	Initiator
LCRs	Locus Control Regions
LD	Linkage Disequilibrium
MARs	Matrix Attachment Regions
MGD	Mouse Genome Database
MGI	Mouse Genome Informatics
NCCs	Neuronal Crest Cells
NCNRs	Non-coding Non-repetitive Regions
NREs	Negative Regulatory Element
PFC	Pre-frontal Cortex
PIC	Pre-initiation Complex
rEHH	Relative Extended Haplotype Homozygosity

SNPs	Single Nucleotide Polymorphisms
SNVs	Single Nucleotide Variants
TFBSs	Transcription Factor Binding Sites
TFs	Transcription Factors
TIC	Transcription Initiation Complex
TRE	TPA Response Element
TSS	Transcription Start site
UTR	Untranslated Regions

INTRODUCTION

Genome of a species in its entirety offers endless information. A remarkable landmark in the roadmap of genomics was achieved by sequencing 2.91 billion base pair of euchromatin human genome (Venter et al., 2001). In this year of 2018, sequencing a genome is a reinvented science where both time and cost of the procedures involved have been eliminated as confounding factors and also the quality generated of a sequence is of decent nature. This propelling feat enabled production of vast amounts of data for exploring various genomic dimensions of innumerable species. This trove of data that reached us in millions of base pairs of strings has been brilliantly utilized in medicinal, physiological, evolutionary and developmental studies over the past few years. In addition to this, several methodologies have successfully cropped up that helped this raw data to be categorized into functional categorizations of a species' genetics.

Human genome of all vertebrate genomes sequenced so far set out to be a source of interest for many as it was the most extensively sequenced genome of all species and also much larger in volume than any other species' genome sequenced prior to it (I. H. G. S. Consortium, 2001). An interest also piques as we belong to the same species and as a conundrum involving us as a center, it invites us to look into its multilayered dynamics. We now know that human genome and pretty much every other genome has two major divisions of labor within a cell. One group of genome sequences are those that code for the proteins and the other present intermittently, either close by to the gene of interest or not, has some role to play in the regulation of these protein-defining genes.

Perhaps the most fascinating of all discoveries made in the area of genomics and related fields in the last decade or two was the rejection of the term "junk DNA". Previously thought out notions, that a relatively larger portion of our genome serves no purpose other than to be of mere presence were brutally disregarded as many researches indicated that 80% of the human genome possessed some kind of biochemical activity (Pennisi, 2012). In eulogy written for "junk DNA" in 2012, it is mentioned that defining the proteins is not the only consumption of the DNA sequence; they also serve as places for binding that could affect the timing and space of the proteins being coded by the genes. These sequences can undergo modifications that will silence the chromosome and can also produce RNA with many pivotal functions (Pennisi, 2012).

Searching for the protein-defining sequences and their corresponding positions in the genome soon after the completion of the Human Genome Project was different in its effect from those of prokaryotes. Features such as larger intergenic regions and adjacently present introns in the eukaryotic genes are absent from the genes of a prokaryote, presented immense challenges paired with lesser computational advancement of that age (Mathé, Sagot, Schiex, & Rouzé, 2002). Gene prediction softwares now employ latest algorithms and efficient model systems that largely decreased the superficiality of the previous prediction pipelines. However, the utmost potential of prediction upto 100 % is still not guaranteed. Presence of a poly-A tail, intron/exon boundaries and an open reading frame comprise some of the predictive signals that help in the overall prediction of a gene's presence but mandatory occurrence of these signals is not assured (Figure 1.1) (Baxevanis, 2004). In sum, three methods make up the protocol of the gene prediction strategies such as 1) site-based methods and 2) content-based methods, widely categorized among the *ab initio* strategies of gene prediction and 3) comparative methods that take into account sequence homology with already predicted coding sequences (Z. Wang, Chen, & Li, 2004). The *ab initio* predictions are strategized via neural networks, Hidden Markov Models (HMMs), dynamic programming and many other advanced algorithms.

In site-based methods, presence or absence of a particular sequence or a consensus is identified. This sequence can correspond to a factor binding site, a poly-A tract, a splice site or the presence of start and stop codons. These signatures of specific sites are also known as signal sensors (Z. Wang, et al., 2004). Content-based methods rely on the sequence properties that are wider in range. From synonymous codons in various species that encode the same codon to characterization of repeats, these features are helpful in assigning properties to a region and categorizing it as a gene (Baxevanis, 2004). Comparative methods include homology based searches in which a query sequence is searched against a database of already curated sequences to see which of these are homologous to the one being queried. This method, although more direct in approach renders limitation to annotate sequences for which no prior homolog in the protein database exists (Baxevanis, 2004). Local and global alignments are the two common approaches of sequence similarity searches that assist in such homology based predictions. In gene structure prediction, these sequence similarity based searches are founded in an idea that exons are evolutionarily more conserved than the non-functional non-coding regions (Z. Wang, et al., 2004).

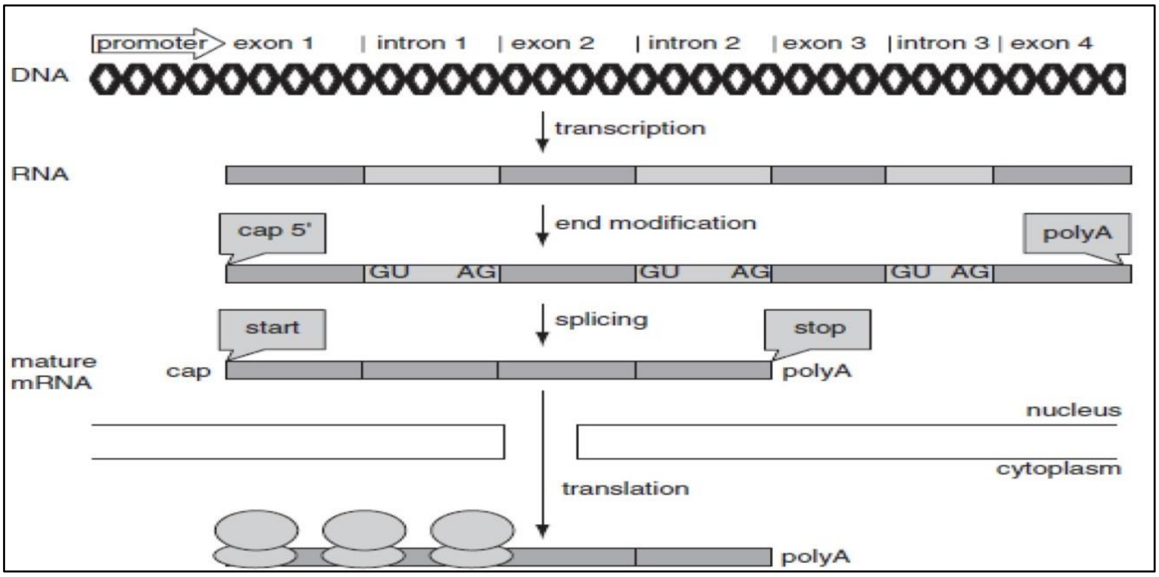


Figure 1.1: Central dogma in Molecular Biology

Central dogma of molecular biology is depicted from DNA being transcribed to RNA and RNA being translated into protein. Computational identification of a gene structure can be made on a number of features preceding or trailing the actual coding sequence. These include features such as start/stop codons, splice sites and poly-A tracts. However, the presence of all such features is not always assured, and if present, do not comply with the same deduction always. Adopted from: (Baxevanis, 2004)

1.1 Regulatory Elements in the Human Genome

Human non-coding regions have no dearth of regulatory elements. Categorizing the gene-coding section of the human genome, however, was much easier in comparison with annotating the regulatory repertoire that controlled it. Unlike in the case of regulatory sequence prediction, there exists a specific triplet code, a transcription start site and a preceding promoter site to strategize prediction of the coding genes. However, annotating gene regulatory regions or the cis-regulatory regions (CREs) was in due part stood the most challenging of the tasks. CREs comprise of promoters and enhancers as major members of the group, majorly influencing gene transcription. Other regulatory elements include silencers, insulator, locus control regions and matrix attachment regions, extending significant contributions to the regulatory landscape of the human genome. All these elements are discussed in detail in the forthcoming sub-sections.

1.1.1 Promoters

Promoters make up an indispensible set of sequences responsible for transcription initiation of protein and RNA coding genes (Umarov & Solovyev, 2017). These 5' flanking sequences contain in them functional motifs for transcription factors (TFs) that upon binding initiate gene expression. The minimal portion of a eukaryotic promoter consists of a core promoter containing a transcription start site (TSS) which has the ability to initiate basal level transcription (Umarov & Solovyev, 2017). Upto 50% of eukaryotic promoters contain a TATA box, some 30bp upstream of the TSS. However, many important genes such as cancer causing genes, housekeeping genes and growth factor genes may have promoters without a TATA box. In such promoters, a recently discovered downstream promoter element (DPE) or the initiator region some 25-30bp downstream of the TSS may act as the positional control of transcription. In prokaryotic promoters, conserved sequence lying approximately 10bp upstream of the TSS initiates transcription, whereas, conserved sequence lying approximately 35bp upstream of the TSS controls the rate of transcription (Umarov & Solovyev, 2017). With the advent of latest sequencing technologies, many genomes have been sequenced and put to public access so far. Within them, correct assessment of gene sequences and efficient prediction of the regulatory networks controlling their expression remain a point of challenge till date. Because of a gene specific architecture of the promoters and lack of an intact conserved sequence in prokaryotes as well as eukaryotes among all their species, their predictability across prokaryotes and eukaryotes still poses serious constraints.

In the human genome, chromatin structure containing cis-regulatory elements like promoters and enhancers require for their activation a combinatorial effort by multitude of TFs and co-factors that bind to these cis-regulatory sequences. As a result of which gene transcription is initiated (Lemon & Tjian, 2000). Various microarray and chromatin immuno-precipitation (ChIP-chip) assays have shown the nature of the chromatin modifications lying in regions where promoters exist that potentially possess the predictive power of elucidating these widespread cis-regulatory features. It has been reported that various histone modifications in the chromatin structure possessing active promoters are indicative of trimethylations in many residues of H3 and H4 and particularly trimethylation of histone 3's lysine 4 (H3K4) (Heintzman et al., 2007). Although, similar signatures also exist in the identification of enhancers, the random location and orientation of the enhancers make these predictions all the more difficult. However, with the notion that promoters exist near the TSS and within a close proximity of the gene its transcription it is controlling, can be utilized as one powerful feature. Among flies and yeast, depletion of the nucleosome is also a powerful characteristic to indicate the presence of an active promoter, this feature though is still to be thoroughly examined in the mammalian system.

1.1.1.1 Role of promoters in expression divergence

Promoters are linked with higher degree of expression divergence. Genes whose promoters contain TATA box have not been associated with more mutation but higher expression divergence has been observed in eukaryotes than those lacking the TATA signature (Tirosh, Barkai, & Verstrepen, 2009). This kind of trend is evidently depictive of a patterned phenomenon that links sequence signatures of regulatory elements with differential gene expression perpetuated in species divergence. The presence of TATA box has been associated with maintaining dynamic gene regulation in eukaryotes. To dissect the above stated facts, the process of transcription can be broken down into two major steps. The first step involves harnessing of the pre-initiation complex (PIC) and the RNA polymerase to the core promoter. The second step involves release of the RNA polymerase to initiate transcription. If the PIC remains bound to the core promoter, a step mainly assisted by TATA box, multiple rounds of transcription can be carried out (Yean & Gralla, 1999). This makes TATA box an extremely important factor in amplifying and re-initiating gene expression when PIC remains bound to the core promoter. Notably, the binding of the PIC onto the core promoter works in close cooperative fashion with TF binding onto other sites and thus largely determines the transcriptional output of the gene. For the very reason, TATA containing genes are more watchful of mutations in their regulatory regions that might modify the TF binding space of the region compared to those which lack it.

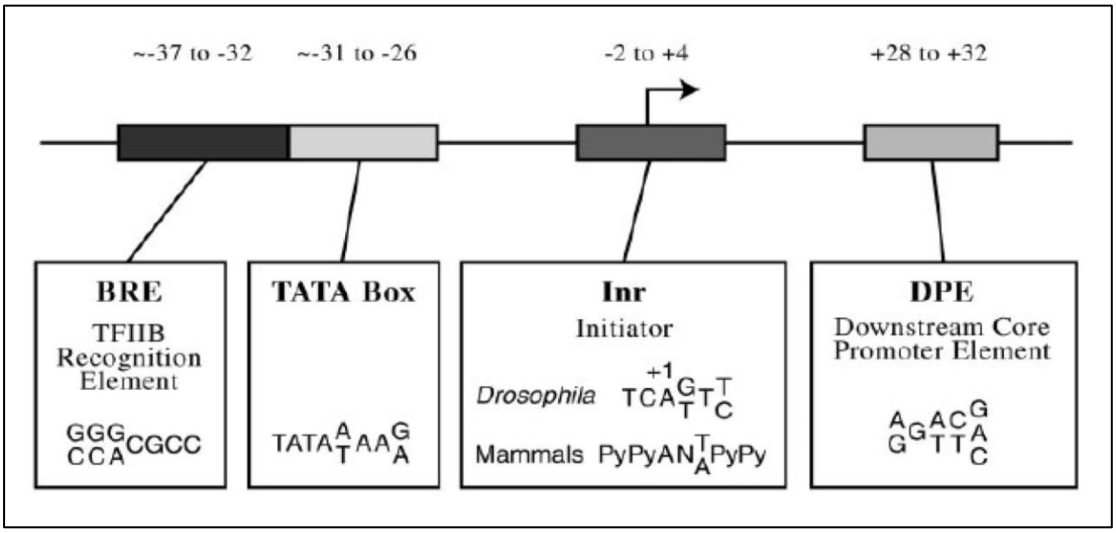


Figure 1.2: Core Promoter of RNA Polymerase II

Composition of a core promoter has been shown. Elements such as BRE, TATA box, Inr and DPE are shown which may or may not be present in all core promoters. DPE motif cannot function without Inr, whereas, TATA box can function even in the absence of the other three elements. DPE consensus has been determined for drosophila. Inr consensus has been shown for drosophila and mammals. Adopted from: (Smale & Kadonaga, 2003)

1.1.2 Enhancers

Enhancers were discovered more than 35 years ago (Banerji, Rusconi, & Schaffner, 1981). They stay dynamic till date as we lack a universal language for their identification. Enhancers have a diversified location that is either in the untranslated regions, introns or gene deserts. They also tend to lie largely irrespective of the orientation of the gene they are transcriptionally controlling (Kolovos, Knoch, Grosveld, Cook, & Papantonis, 2012). One of the initial identifications of the enhancers came from comparative genomic techniques in which various non-coding regions of the genome were seen to be highly conserved among mammals and vertebrates. Upon empirical investigation, several of these highly conserved non-coding regions were detected as developmental enhancers. Although sequence conservation could turn out to be a turning point in their predictive space, evidence suggests that identical expression level of genes was observed between species whose enhancers bore no similarities amongst them (Hare, Peterson, Iyer, Meier, & Eisen, 2008).

1.1.2.1 Models for enhancers' mode of action

Enhancers by far make up the most important category of cis-regulatory elements. They largely increase the transcription of the gene by interacting with one or more promoters. As mentioned earlier, given their distal nature that indicates their occurrence to be many kilobases away from their target promoter/promoters, they also happen to lie in an orientation independent manner of the gene whose transcription it is increasing. Enhancers could be occupying the intron of a gene it is transcribing or could be present in the intergenic region bypassing several close by genes to ultimately help in the transcription of a distal gene. Fluorescent *in situ* hybridization (FISH) and chromosome conformation capture (3C) methodologies have supported the looping mechanism by which the largely spaced enhancers and their target promoter come in contact with each-other via ligation and also to the gene of interest (Pennacchio, Bickmore, Dean, Nobrega, & Bejerano, 2013). There have also been proposed other models for their interaction with promoter to initiate transcription such as the tracking model (TF travelling along the DNA towards the promoter site) , the linking model (Polymerization of TFs towards the promoter site), and the relocation model (gene relocating to make enhancer-promoter interaction feasible) (Figure 1.3) (Kolovos, et al., 2012).'

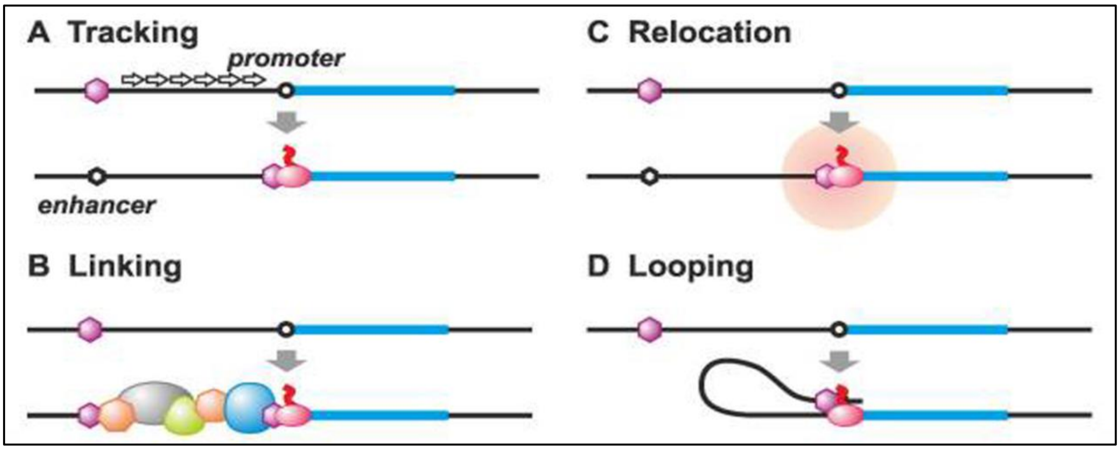


Figure 1.3: Models for Enhancer's Role in Initiating Transcription

(A) *In the first model, a TF colored in pink binds to the enhancer, and propagates along the DNA towards the promoter site where it binds with the polymerase and initiates transcription*

***(B)** In the second model, a TF (in pink) binds to an enhancer site, polymerizes other TFs in the direction of the promoter to initiate transcription **(C)** In the third model, the gene relocates to make enhancer-promoter interaction feasible **(D)** The intervening DNA loops out to make physical interaction between the enhancer and promoter feasible via protein-protein interaction. Adopted from: (Kolovos, et al., 2012)*

1.1.2.2 Relational dynamics of enhancers and transcription factors (TFs)

An enhancer sequence can recruit transcription factors in a variety of ways. TF cooperativity either by direct interaction among the adjacently binding TFs or through indirect co-binding with the co-factor largely determines the transcriptional outcome an enhancer will deliver (Spitz & Furlong, 2012). Functional implications of TF binding could be debated as a TF binding event does not always imply regulatory control of the nearby genes. Many binding events have been termed non-functional and could be due to easier access to chromatin that the TF has occupied or reconfiguration of the nucleosome induced by the binding event for facilitating another TF occupancy leading to gene expression (Spitz & Furlong, 2012). Differences in the transcription factor binding sites (TFBSs) between the species within the regulatory sequences can impart huge impact on the regulation of the associated genes. Substitution in intron 8 of *FOXP2* gene within the vertebrate conserved POU3F2 binding site in the present-day humans when compared with Neanderthals portrayed potential candidacy for driving selective sweep in the entire *FOXP2* gene (Maricic et al., 2013). Selective sweep in a population, therefore, confers a genomic region significant where an allele offering a fitness advantage increases in frequency along with other neighboring alleles (LD: linkage disequilibrium). This phenomenon renders the entire locus less diverse (Cadzow et al., 2014).

1.1.2.3 Role of enhancers in phenotypic evolution

Enhancers make up the category of the most widely assayed cis-regulatory elements. From their discovery to their incessant dissection into controlling phenotypic variation even amongst the human population has been regarded carefully in numerous studies. As it is now easily comprehendible that a wide variety of biochemical modifications exist in the genomes that provide insights into categorizing regulatory elements. Chromatin structure, numerous histone modifications and binding sites for various

TFs have largely been determined for a larger set of TFs and their co-factors in all sorts of virtual cellular environment. It is also important to note that 10-20% of the human genome regulates gene expression that may consist of enhancers, promoters and other regulatory elements (Pennacchio, et al., 2013). It is estimated that enhancers make up the largest content of regulatory repertoire and hence more prone to incorporating changes that could contribute to species specific evolution of a trait.

Another aspect of enhancers' vital importance in driving evolution comes from their modular mode of action. Notably in humans, 80% of the GWAS-associated SNPs are non-coding and a larger percentage must occur in these regulatory elements (Hindorff et al., 2009). For a gene to be expressed in many cells and tissues, a mutation in it can prove detrimental. However, in terms of modularity of enhancers where tissue specific coordination between enhancer and other regulatory elements can drive the expression of a gene in one cellular context can be differentiated from an entirely new expression pattern observed in another context where the enhancer or assisting regulatory regions may not be active (Pennacchio, et al., 2013). Selection and mutation in enhancers therefore can go hand in hand in making regions of choice to be sources of adaptation and fitness. Many examples exist that are evident on enhancer's contribution to loss or gain of a phenotype in a lineage specific manner. In drosophila, for instance, larval trichome formation and wing pigmentation are all such adaptations (Pennacchio, et al., 2013). Within the human population, lactase persistence is a good example of regulatory mutations affecting the phenotype (Fang, Ahn, Wodziak, & Sibley, 2012).

1.1.3 Silencers

Suppression of gene expression in eukaryotes is encountered by employing silencers. Silencers like enhancers and insulators make up compositional components of metazoan regulatory landscape (Kolovos, et al., 2012). Silencers were first discovered to affect the the mating type loci in yeast in 1985 (Brand, Breeden, Abraham, Sternglanz, & Nasmyth, 1985). Silencers, as the name indicates, silence transcription by working antagonistically to enhancers which enhance expression. Silencers have two important categories, of which one are the classical silencers that cause gene suppression irrespective of their position. The second are negative regulatory elements (NREs) that are position dependent and create passive suppression by interfering with the upstream elements (Ogbourne & Antalis, 1998). These silencer sequences are DNA sequences that act as binding sites for various repressor proteins that work in a variety of ways. A repressor can mimic an activator and may compete over the binding site required for a gene's transcription. Repressors by binding within a close range as that of an activator may interfere with its activity. However, by binding with silencers, they can inhibit the formation of transcription initiation complex (TIC or the GTF assembly) and its respective activity through protein-protein interactions (Narlikar & Ovcharenko, 2009). Silencers for their repressing effect have largely been known to work in an orientation free manner from the gene they are repressing the transcription of. However, cases have also been reported in which silencers were seen to be present among enhancers and for some they work only within the untranslated regions (UTRs) and promoters (Narlikar & Ovcharenko, 2009). Silencers can also exist as independent entities. They are often called as insulators for their ability to confine the expression within specific chromatin boundaries. Instances have been reported in which silencers and enhancers were reported for long-range interactions form distances as long as 130kb with promoter of the *MECP2* gene, a gene widely implicated for X-linked dominant neurodevelopmental disorder, Rett Syndrome (J. Liu & Francke, 2006). This gene has the highest degree of expression in human brain compared to other tissues.

1.1.4 Insulators

Enhancers were initially thought to work for specific target promoters over long range distances in a eukaryotic system. However, transgene experimentation showed that this was not always the case. Enhancer-promoter pairing to activate a gene can be restricted by intervening sequences that can prevent this interaction. Insulators in effect are those fragments of DNA that protect genes by blocking the activity of signals stemming from their surroundings. They prevent the expression of genes by blocking interaction between the enhancer/s and a promoter of a gene only if they are present between them and not anywhere else. The other way they can operate is by creating barriers or hurdles to curtail the spread of heterochromatin that otherwise would silence the expression of a gene. Thus, insulators are those regulatory sequences that prevent both the activation and repression of a gene by either posing a hindrance for interaction between two adjacently lying chromatin structures interaction or by acting as a barrier respectively (West, Gaszner, & Felsenfeld, 2002). The insulator sequences contain multiple sites for transcription factors and the extent to which they insulate is directly linked to the number of sites present.

1.1.5 Locus Control Regions

Cell lineage-specific regulation of a gene is not only dependent on the individual cis-regulatory elements (enhancers, promoters, silencers and insulators) but their collective localization in a chromatin structure that can independently regulate the expression of a gene is also of importance. Locus control regions (LCRs) are such regions containing multiple CREs that bind to their own set of TFs (Li, Peterson, Fang, & Stamatoyannopoulos, 2002). LCR was initially identified for the globin gene loci, and its function was designated to its ability to enhance expression of a downstream cluster of genes in a tissue specific manner at unusual chromatin structures (Grosveld, van Assendelft, Greaves, & Kollias, 1987). In expressing cells, these regions make up a sovereign chromatin identity that contains DNAse I hypersensitive sites (DHSSs), a feature highly explored in the identification of CREs. These sites also have binding sites of various TFs. The LCRs are mainly associated with the eukaryotic gene systems and apart from controlling the expression of a downstream gene can also reside in the introns of some genes. They also confer an open chromatin structure on a locus which is indicative of the DNA being accessible to various binding factors (Li, et al., 2002). In a study, none of the globin genes were expressed when a 35kb upstream region of the globin locus was deleted that configured the entire locus in a closed chromatin structure (Forrester et al., 1990).

1.1.6 Matrix Attachment Regions

Matrix attachment regions or MARs are also known as the matrix associated regions or the scaffold associated regions. It has been reported that chromatin structure in eukaryotes also acts as a regulator of gene function (Dillon & Grosveld, 1994). Chromatin in eukaryotes is arranged in various domains and loops (Paulson & Laemmli, 1977). These loops are defined by specific DNA fragments that help chromatin fiber to bind to nuclear matrices outside of the histone-chromosomal complex. The nuclear matrices are isolated protein structures created as a result of histone depletion and get bound to specific genomic DNA fragments (MARs) *in vitro* (Rollini, Namciu, Marsden, & Fournier, 1999) . As MARs were originally discovered through *in vitro* biochemical studies, they subsequently implicated for their regulatory role as insulators in confining the chromatin structures from acting on the cognate genes (Dillon & Grosveld, 1994; Rollini, et al., 1999).

1.2 CREs' Evolution in the Human Genome

It was argued decades ago when the operator sequence in *lac* operon was discovered that the environmental condition in which a gene product is formed is as much important as the product itself (Jacob & Monod, 1978; Monod & Jacob, 1961). Mutations as heritable source of all variation among and between the species, therefore, were in large part correlated with their significant phenotypic impact when present in the regulatory regions. Two ground-breaking studies in the 1970's further bolstered these speculations and paved ways for future studies that today serve as the basis of cis-regulatory divergence. In the first study, Britten and Davidson identified that repetitive elements control gene transcription and mutations in them can largely induce variable phenotypic effects in the organism (Britten & Davidson, 1971). In the second study, King and Wilson stated that homologous proteins of human and chimpanzee are almost identical. Therefore, smaller changes in the divergence of these proteins cannot make grounds for such large amount of differences between the two species and hence, gene regulation, the condition and time in which a gene product is made must govern for the phenotypic differences existing between the two (King & Wilson, 1975).

Over the past years, mutations in the coding regions have pointed out consequences directly on the protein product being made. This being done, was later evaluated to be an easier to undertake phenomenon, where to identify such mutation, lesser complications are faced. For instance, non-synonymous mutations, frameshifts as a result of gross mutations, and non-sense mutations that induce pre-mature stop codons, consequently producing shorter polypeptides, are relatively easier to locate even with comparison of the DNA sequences. However, mutations that have the potential to fully disrupt the transcriptional profile of a gene are difficult to decipher for which several functional and biochemical evaluations have to be carried out (Wray, 2007).

To estimate the evolutionary extent of these cis-regulatory mutations, their effect on the phenotype of an organism is required. One hypothesis states that phenotypic effects introduced via such kind of cis-regulatory mutations are more pronounced in some traits. Since, transcription is a dynamic process prone to be fine-tuned according to organismal context demands in processes such as reproduction, adaptability, immunity, development and behavior; it is easier to accommodate the required changes in these processes by reinventing the regulatory regions and through that altering the spatiotemporal expression of the genes (Wray, 2007). As a result, processes that require rapid change in the phenotype can be satiated by cis-regulatory mutations then by an altered macromolecular structure of a polypeptide as a result of a coding sequence mutation, which is more of a slower and static process. On similar lines, selection acting on these cis-regulatory mutations is more efficient (Wray, 2007).

In humans, several mutations in the regulatory regions have been identified in processes such as immunity, diet and most importantly cognition where evidence of positive and balancing selection has also been reported. In a study, a SNP was identified in the binding site of a GATA1 protein (Tournamille, Colin, Cartron, & Le Van Kim, 1995). This mutation peculiarly inhibited the transcription of *DARC* gene in erythrocytes that subsequently made the cells resistant to the infection of *Plasmodium vivax*. This mutation, however, did not constrict the gene to be expressed in other cells. The haplotypes of the modern human population carrying the GATA1 binding site modifying mutation also showed positive selection in places where malaria was prevalent (Hadley & Peiper, 1997; Hamblin & Di Rienzo, 2000). Dietary shifts to omnivorous eating habits have also been attributed to mutations in the regulatory regions that introduced lactase persistence in the humans, a dietary ability apparently missing in the great apes (Fang, et al., 2012; Olds & Sibley, 2003).

1.3 Human Trait Advancement

Soon after a full-length discovery of the human genome, its annotation in terms of functional categorization went drastically forward. But things did not stop here.

Having had a complete set of genes known and regulatory elements predicted and also empirically verified, efforts were drawn to estimate these genomic regions for their prospective role in species' evolution and development. Genome evolution owes its dynamics to two things: mutation and patterns of natural selection (Pennacchio, et al., 2013). Almost 85% of the human genome undergoing selection constraint comprises of the non-coding regions (Ward & Kellis, 2012), which contain a reasonable number of cis-regulatory elements. A likely consequence of this factual unraveling is that mutation occurring in a cis-regulatory region and the prospective effect of selection could prove to be either detrimental or highly favorable for the organism's evolutionary fitness. In a study by Ward and Kellis, 95% of the human genome reported to be non-conserved across mammals is somehow biochemically active. This non-conserved but biochemically active DNA contributed to lesser diversity in humans, depicting a lineage specific purifying selection. In contrast, the remaining 5% inactive, conserved portion seemed to have contributed to the human variability also indicating recent non-functionality (Ward & Kellis, 2012).

Over the past few years, from precisely acting promoters to distantly spread enhancers in a genome, these elements have come to light in terms of massive implications they brought with them in the form of regulatory variants. These variants have not only been a target of natural selection to increase the stature of human's adaptability as the most thriving creature of its time but resistance to various disease, changes in immune responses along with highly cognitive brains, inclusion of language and the *Homo sapien* evolution itself from its predecessors have enlightened and amazed the scientific community to great measures. Such features comprise the highly attuned features of the modern human lineage with the climatic, ecological and physiological demands. Perhaps the competition of survival among the species of genus Homo could largely be placed onto the strand of evolution that fine-tuned the regulatory mechanism of the present-day humans to affect a set phenotype in a set of defined cells.

1.4 Human Brain Evolution

Brain folding provides a significant lowdown on unique human brain structure. Mammalian ancestral brain was evidenced to also have a folded brain but during the course of evolution many mammals lost this feature and now possess smooth brains, for example, mice. However, human brain in its current form has the largest number of furrows and folding which provide it with larger surface area, more neuronal accommodation and hence more cognitive power. Though we see greater extent of gyrification (brain folding) in primate clade as well, humans act as outliers even within this most intelligent clade of non-human primates (Atkinson, Rogers, Mahaney, Cox, & Cheverud, 2015). The brain size varies greatly among the mammalian species. This increase in brain size owes it to enlarged cerebellum, neocortex, olfactory cortex, and enlarged olfactory bulbs (Rowe, Macrini, & Luo, 2011). To devise the pattern of early brain evolution in mammals, it is believed that brain expansion to mammalian levels happened in various phases. The major reasons of brain expansion lie with sophisticated attainment of the power of odor i.e. olfaction, an overall sensory innovation and also because of increased neuromuscular coordination (Rowe, et al., 2011). Because of lack of fossil proof, comparisons among the living mammalian species reveal that encephalization and development of neocortex along with increased power of smell, hearing, metabolism, nocturnality and nutrition have a lot of evolutionary drivers (Rowe, et al., 2011).

Lying at the topmost offshoot of the animal kingdom, human brain paved intriguing ways for attaining this place with its decision making power, advanced cognition and also by adopting other physiological and mechanical advantages such as bipedalism, dexterity of the hands, usage of tools etc. Human brain is three times larger in size than the brain of apes (Enard, 2015). This increase in brain size when on one hand helped a lot in increasing its faculties; it also became energetically more demanding for humans. It has been reported that at infancy in humans, cerebral volume greatly increases along with a dramatic increase in the white matter but the same is not witnessed in chimpanzees (Sakai et al., 2013).

To apply a logical consequence to the increase in brain size, more neurons should be present in a bigger brain. This idea became the basis for ‘radial unit hypothesis’, that stated that in order to have more neurons, more neural progenitors should be present and actively dividing (Enard, 2015). Owing to more than 20 million changes in the human genome after its divergence from chimpanzees, changes must have been incorporated that sped up the proliferation of these cells. This called for probing into molecular and cellular mechanisms of evolution that are studied with great precision in mouse model systems. Efforts have also been directed onto gauging parts of the brain that have increased preferentially in humans. For the very reasons, it became imperative to determine the underpinnings of such traits in terms of both coding and regulatory sequences. Lately a study by Boyd et al. uncovered increased cell cycle in neural progenitors that markedly increase brain size in humans. However, orthologous chimpanzee sequence upon insertion could not produce the suspected result (Boyd et al., 2015).

Among the brain regions, forebrain takes up an executive seat in the anatomy of brain and various neuropsychological disorders occur due to problems arising in this region. Human forebrain has been categorized into pallium, sub-pallium, hypothalamaus and thalamus which control and coordinate the higher level function of the human brain (Nord, Pattabiraman, Visel, & Rubenstein, 2015). Forebrain formation and organization is associated with higher level transcriptional circuitry. Over the years, efforts have been directed to devise a TF code that works only in a cell or tissue specific manner in embryonic brain, however, empirical evaluation of such TFs in terms of their complex, spatiotemporal interaction with cis-regulatory elements remains a point of elucidation till date (Nord, et al., 2015).

1.5 Gene Regulation in Human Brain Development

Gene regulation has long been playing a role in fine-tuning the brain circuits that distinguish the highly cognitive human brain from that of the protein comparatively lesser adaptive non-human primate brain function (Cáceres et al., 2003). Primate brain evolution displays a disproportionate enlargement of neocortex, frontal lobe and an overall larger brain volume, properties that underpin its intelligent workings (Dunbar & Shultz, 2007). Human brain is triple in size and more efficiently adapted to do highly complicated assessments through language and cognitive skills than that of great apes (Geschwind & Rakic, 2013). Evidence also suggests that human neocortex possesses a greater volume and significant cell cycle differences that lead to increased corticogenesis (Boyd, et al., 2015). At molecular level, little evidence has been uncovered to relate gene sequence change with the phenotypic traits that bifurcate humans and the closest relative chimpanzee into two different strata of intelligence. It is however established that gene regulation, the spatiotemporal expression of genes play a defining role in making up the current form of highly adaptive brain of present-day humans (Cáceres, et al., 2003; Enard et al., 2002; Gu & Gu, 2003). Previous study stated that the human-chimp cerebral cortex relies on a special patterning of gene expression. Out of a gene pool considered in the study, 169 genes were observed to have expressed differently between human and chimpanzee. Among them, 91 genes hinted at being differently expressed in the human lineage alone, with macaque as an out-group (Cáceres, et al., 2003). About 90% of the genes that were differentially expressed in human lineage belonged to brain, whereas in liver and heart, nearly an equal number of genes were upregulated and downregulated between human and chimpanzee. (Cáceres, et al., 2003). Another analysis sums up the number to 54 pre-frontal cortex (PFC) genes having a lineage specific upregulation in human PFC after divergence from other hominoids (Geschwind & Rakic, 2013).

1.5.1 Enhancers and human brain development

Recent findings have highlighted that human specific mutations in enhancers can impart huge changes in gene regulatory mechanisms and eventually produce brain size differences (Boyd, et al., 2015). Enhancers despite of their proximal existence to promoters of some genes are widely catalogued as also the distal category of cis-regulatory elements, residing many kilobases (kb) away from their target genes; and contribute to gene regulatory networks in terms of initiating cell specific gene expression together with TF occupancy (Choukrallah, Song, Rolink, Burger, & Matthias, 2015; Spitz & Furlong, 2012). In mammals, enhancers are either active or primed. Active enhancers possess biochemical signatures of H3K27ac and H3K4me1 and are associated with actively expressing genes whereas primed enhancers possess only the latter methylation mark and are most likely to get activated later on by a developmental or environmental stimulus once a cell has acquired its tissue specific identity (Choukrallah, et al., 2015).

1.5.2 Role of enhancer sequence acceleration in human cognition

Many of the accelerated portions of the genomes harbor developmental enhancers and genomic changes within them can impart huge alterations in brain function (Burbano et al., 2012; Hubisz & Pollard, 2014; Prabhakar et al., 2008). Evolutionary studies have also endorsed acceleration in enhancer sequences compared to coding and non-coding/non-enhancer genomic blocks in vertebrates during land adaptation (Yousaf, Raza, & Abbasi, 2015). A recent study has therefore consolidated this view where human specific changes in a neuro-developmental enhancer of *FZD8* gene produced immense differences in the size of the brain (Franchini & Pollard, 2015). Necessitating enhancers and their role in predominantly controlling the spatiotemporal expression of the genes, sequential changes that rapidly accumulated in human brain enhancers should be evaluated (Maston, Evans, & Green, 2006). For that a strong limiting criterion to include brain specific enhancers that are already functionally confirmed should be observed that brings forth the safety of eliminating any genomic non-coding portions that failed to act as enhancers during functional verifications (Kvon, 2015). This criterion is in line with recent studies that have rendered the use of biochemical signatures such as H3K4 monomethylation for enhancer function and prediction useless (Dorighi et al., 2017). Bulk of data has been introduced in the form of individual studies as well as publicly accessible databases to acquire such empirically confirmed enhancers. VISTA enhancer browser remains one such publicly accessible, widely utilized repertoire of verified enhancers (Visel, Minovitsky, Dubchak, & Pennacchio, 2007).

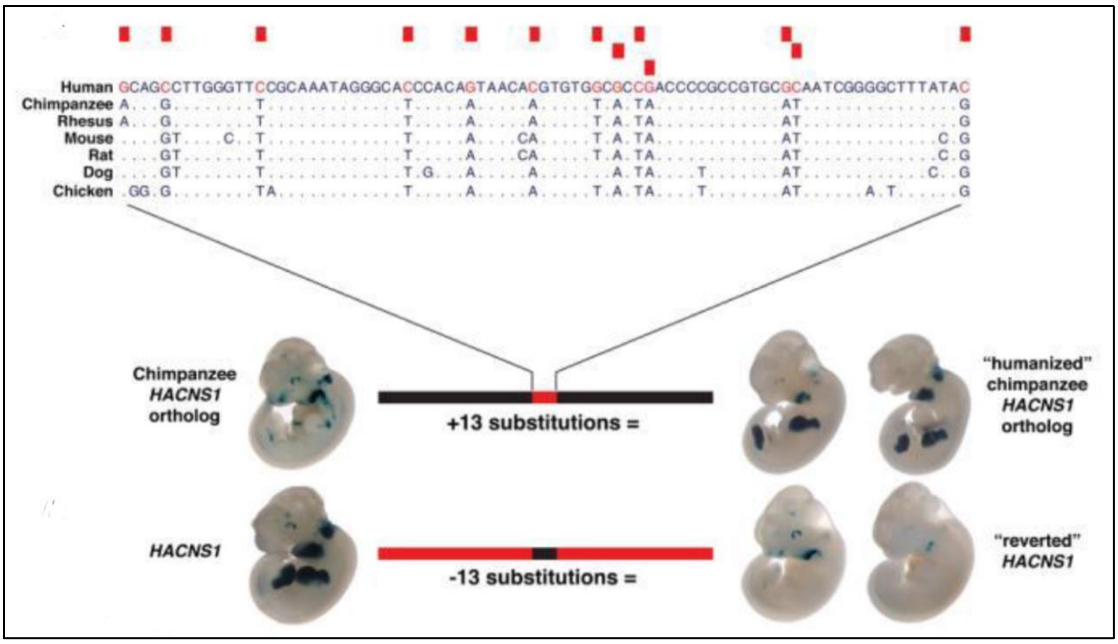

Human GCAGCCTTGGGTTCCGCAAATAGGGCACCCACAGTAACACGTGTGGCGCCGACCCCGCCGTGCGCAATCGGGGCTTTATAC
Chimpanzee
Rhesus
Mouse
Rat
Dog
Chicken
Chimpanzee HACNS1 ortholog
+13 substitutions =
"humanized" chimpanzee HACNS1 ortholog
HACNS1
-13 substitutions =
"reverted" HACNS1

Figure 1.4: Gain of function in human accelerated region

An 81bp region containing 13 substitutions in the human genome was compared to other vertebrates' orthologous sequences. Each substitution is indicated via red boxes on top of the alignment. These substitutions were part of an accelerated region (HACNS)1 that was verified as a developmental enhancer. The humanized version (containing the 13 substitutions) was inserted into the mouse resulted positive for expression in limbs. Whereas, when these substitutions were reverted and non-humanized version of the sequence was inserted into mouse, no detectable expression in limbs was obtained. Adopted and Modified from: (Prabhakar, et al., 2008)

1.6 Archaic Humans of the Genus Homo

Various amounts of data from genetics, archaeology and paleontology revealed that our ancestors, extinct species of the genus Homo, also possessed specialized brains. Although they could not survive in the game of being the fittest and eventually succumbed to climatic or intellectual factors that modern human beings excelled at, they still present evidence if one gets down to tracking changes that eventually made us this modern and behaviorally advance. As fossilized brains are absent from our collection of evidence, we can still get an idea from the bony braincases (Neubauer, Hublin, & Gunz, 2018). Prior studies indicated while studying and comparing endocasts of both modern and archaic humans that modern humans possess retracted smaller faces, but larger brain cases. Modern humans have globular brains and globular endocasts, with round, enlarged cerebellar areas, and protruding parietal and steep frontal regions. However in our ancestors such as Neanderthals and others, an anterior-posterior elongation existed (Neubauer, et al., 2018). Evidence from craniodental data of our ancestors also shows the extent of dynamics to which facial, mandibular, cranial and dental advancements in the current-day humans originated and persisted (Richter et al., 2017). Exploiting this information to determine what went differently in the evolution of brain and its associated features provide insightful information into modern human evolution.

1.7 Human Facial Features

When brain size enlarged in humans, it also changed the correlated mechanics of the human facial features and size. Craniofacial development in humans, in essence, is a very intricate phenomenon in which inductive and directive molecular interactions come together to help differentiate cells into different facial layers (Evans & Francis-West, 2005). In vertebrates, craniofacial development is largely determined by the neural crest cells (NCCs) that contribute cartilage, connective tissue and bone to the developing head. There has been reported integrative interplay of intrinsic program of the NCCs and some outside cues that guide facial morphogenesis. NCC's in effect are migratory cells that originate from dorsal part of the neural tube under development. After induction, these cells migrate to different cranial regions and help in the formation of facial, pharyngeal skeletons and also contribute to bony and cartilaginous parts of the braincase (Minoux & Rijli, 2010) (Rada-Iglesias et al., 2012).

As much as the head is regarded as the most refined creation in the vertebrate history, it had to be complemented with equally sophisticated facial features. Auditory, nasal, vision related and dental reinventions in the human lineage were mandatory. In humans, many craniofacial abnormalities, in large part have been reported as repercussions of the mutations in genes of limb development. Abnormalities in limb and skull are associated with craniofacial malformations as cases have been observed for many developmentally important genes ,like sonic hedgehog (*SHH*), *ALX4* and *TWIST* (Wilkie & Morriss-Kay, 2001). The role of non-coding regions in terms of cis-regulatory elements has also been proven. In one study, enhancers potentially controlling regulation of facial genes such as *LHX8* whose implication in human palate abnormalities have also been reported (Malt, Cesario, Tang, Brown, & Jeong, 2014).

1.7.1 Nasal Morphology

Human nasal form comes in two major varieties. For African noses, we see enlarged nasal cavity and bulbous nasal mass. For European noses, heighted, aquiline noses are observed. This is in accordance with the climatic constraints both extremes of the regional demarcations face. Africa being the hottest and the most humid continent needs ventilation at much faster and efficient rate. This was therefore assisted with a larger nasal cavity and larger nostrils so that quick and efficient ventilation could be assured. In contrast, extremely cold and dry Europe needs to moisten and warm the air first so that lung damage could be prevented. This was done by keeping a constricted nasal cavity (Zaidi et al., 2017). Although this quantitative data also presents outliers as regions in extremely cold northern China does not indicate any such resemblance to European nasal cavity, we can consider them as outliers for which factors other than climate could be playing a part.

Certain facial features are associated with certain populations, for instance nasal bone flatness is associated with East Asian population. There have been cases of admixture (migration) that exposes people living in one region to an outer region's environmental constraints. Inter-racial marriages also pose complexity when studying about distinct facial forms. However, with evolution, not just human genome got refined to its best-fit form, adaptations are continuously being made according to climatic or other external stimuli and genetic basis of the traits can't be investigated by isolating populations. Gene flow is a major factor for overlap of facial features between the populations. However, certain facial features stay distinct and depictive of a certain population and its regional whereabouts. One such population cohort of heterogeneous population of Latin America is included in the study to address such concerns (Adhikari et al.,2016).

The extent of nose shape and size variation exists not only between the members of modern humans but comparing their nasal cavity with that of Neanderthals also presents some fascinating revelations. As Neanderthals inhabited colder areas of Europe, their nasal cavity should be according to one observed in Europe. However, this is not the case as they have larger nasal apertures, an anomaly that cannot be explained within the constraints of European climatic (Evteev, Cardini, Morozova, & O'Higgins, 2014). An alternative narrative that Neanderthals must have adapted their wider nasal apertures prior to migration to Europe also prevails for which concrete evidence remains to be brought up yet. Therefore, background genetic interplay and operatives of natural selection on the alleles shaping the nasal form become crucial in understanding the overall adaptation pattern of present-day humans. Several features such as nares width, nasal protrusion, nasion position, and nasal bridge breadth have been extrapolated and investigated in several genome wide association studies (see Figure 1.5). Apart from quantitative data that measure and compares these features, it is important to align it with signatures of natural selection so as to know what genetic factors largely shaped up such versatile nasal morphology among human population.

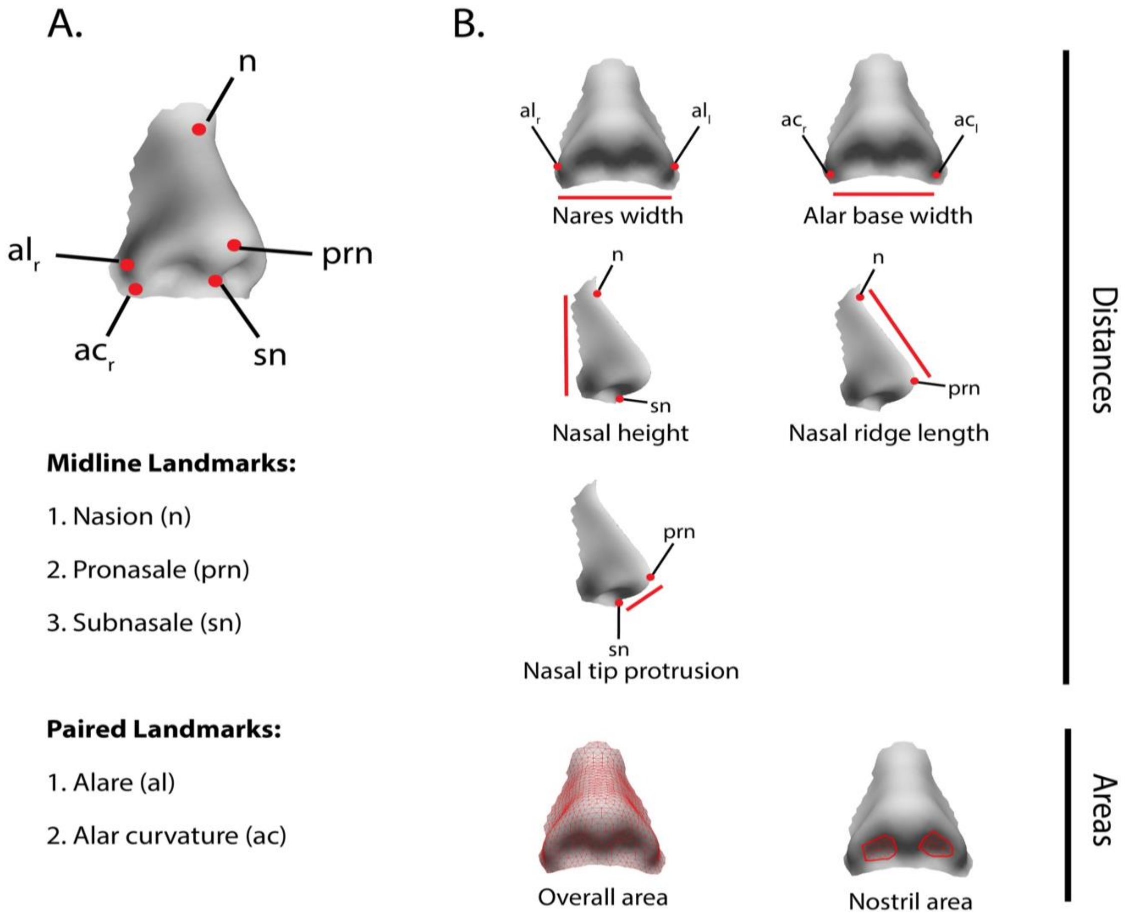


Figure 1.5: Nasal Midline and Paired Landmarks of the morphology

In the above figure, nasal features have been highlighted that have been usually considered in various previously reported studies in order to calculate nasal measurements. ***A)*** *Landmark points are highlighted with red dots that are usually taken into account for nasal measurements. These include midline and paired marks.* ***B)*** *Nasal measurements such as distance between the two extremes of nares (nares width), nasal height, nasal tip protrusion and overall areas are covered. Adopted from: (Zaidi, et al., 2017)*

1.8 Regime of Natural Selection

The regionally diverse changes in craniometric analyses among living human populations are often correlated with climatic change (Roseman, 2004). These morphological variations were in part speculated to have arisen as a result of neutral evolutionary forces such as genetic drift that paved ways for a diversified cranial anatomy and along with it facial diversity (Roseman, 2004). Several studies have been put forth that identified brain evolution in terms of either positive selection or purifying selection. Acceleration in a modern human genome however could be logically connected to mutations that first arose in a genomic region, got sustained (overlooked by the repair mechanisms) and then accelerated to an extent that a small region developed a large amount of changes in a short span of time i.e. 6 Myrs after the split from chimpanzees. In a study, upon comparison with archaic human data, 84% of the human specific substitutions in HARs had at least one allele in sharing with the ancestral hominins. 8% of the remaining substitutions however were categorized to be recent, that is they originated in the common ancestor of the *Homo sapiens* and early hominins (Burbano, et al., 2012). It was also observed, that these substitutions in HARs tend to get fixed much rapidly than those present elsewhere (Burbano, et al., 2012). However, owing to plethora of nasal morphological forms, each serving a particular climatic condition, we can expect forces of natural selection to have acted in manipulating genome level information and using it to its advantage (Roseman, 2004). By keeping in mind the ancient migrations that exposed humans to various climatic extremes, and the extent to which these drastic climatic changes must have helped shape the nasal architecture, it is incumbently necessary to probe into mechanisms that have helped genome level changes to adopt a region specific pattern of selection (Lieberman, 2008).

1.9 Aims of the Study

As discussed in the preceding sections, our aims enlisted gathering of the most widely distributed non-coding functional enhancers that were empirically verified for sole expression in the human brain. This confinement for their functional testing assured their bona fide status as enhancers. By applying various selection and rate analysis tests and by adding orthologous sequences of closely related primates, these enhancers were to be rigorously tested for human lineage specific sequence acceleration. This acceleration in enhancer sequences was then to be compared with transcription factors binding site analysis. Among them, *Homo sapien*-specific sites were to be gauged under the parameters of natural selection among the present-day human population. In extension to this, the possible outcomes of an increased brain on facial features and particularly on the nasal forms which are vastly distributed in the present-day human population were evolutionarily gauged. GWAS-associated SNPs for nasal morphological features were to be gathered and put to test for regionally driven selection among various genomes of *Homo sapiens*.

In summation the goals of the study can be highlighted as:

- Enhancer sequences are prone to lineage specific acceleration in the human brain tissue. This acceleration can act as a crucial indicator of modern human specific anatomy and higher degree cognitive function of the brains.
- This cis-regulatory accelerated environment must incorporate assistance to trans-regulatory factors in which substitutions in binding sites along with a faster sequence divergence rate must be devising newer binding site patterns in modern humans.
- Evolution in the cis-regulatory environment is more pronounced, particularly in the functionally relevant, accelerated regions between species of the primates taking modern humans as the foreground branch.
- Enhancer sequences should diverge from the ancestral archaic human sequences as well in functionally important domains such as TFBSs among the cis-regulatory sequences so as to devise specifically a trans-cis code that worked especially in the benefit of *Homo sapiens*.
- Ongoing evolutionary signatures among the present-day human population can be tracked by rendering such binding site modifying alleles to population genetics and determining which region offers a relaxed constraint or special selection to the site modifying allele in the present-day human genome.
- Brain expansion and associated higher number of neurons or progenitor cells must have resulted due to fine-tuning of the cis-regulatory elements. This altogether increase in size should have physical ramifications on the braincase. An enlarged braincase must have reshaped the face size and facial features in the current-day humans. Apart from this, climatic effects are one of the primary drivers of adaptation and evolution, therefore, nasal shape which acts as body’s natural conditioning system was to be opted as a case study to study evolutionary parameters of SNPs associated with various nasal phenotypes in a region wise manner.

MATERIALS AND METHODS

2.1 SECTION 1: Enhancer Divergence

Enhancers make up an important category of accelerated cis-regulatory elements that efficiently control the spatiotemporal expression of many developmental genes. Establishing plausible reasons for accelerated enhancer sequence divergence in *Homo sapiens* has been termed significant in various previously published studies. Commensurate with such evidence, our first round of work in this section encompassed methods that helped gauge signals of acceleration in brain specific enhancers as part of the brain evolutionary process in humans. For this we relied on an empirically confirmed set of brain specific enhancers and subjected them to an inter-species analysis to shortlist those enhancers that have diverged relatively faster in humans. Combined with transcription factor binding site (TFBS) analysis on the set of accelerated enhancers, we then set out to see the sequence level changes within these TFBSs that may have been the cause of cis-regulatory evolution. These changes within the TFBSs accounted to single nucleotide polymorphisms (SNPs), fulfiiling the classcial definition i.e. the occurrence of an allele has to be >1% in a population (G. P. Consortium, 2010; Karki, Pandya, Elston, & Ferlini, 2015). These SNPs were then analyzed for selection signatures among the present-day human population by employing 1000 Genomes Project Phase III data (G. P. Consortium, 2015).

2.1.1 Sequence collection of empirically tested human brain enhancers

We initiated our search for functionally confirmed enhancers by employing an *in vivo* repertoire of VISTA enhancer browser (Visel, et al., 2007). Our collection limited enhancers that expressed solely in the human brain. In sum, from an available total of 1393 elements at VISTA with enhancer activity confirmed in different kinds of tissues (at the time of this study), we collected only 271 enhancers that showed endogenous expression profiles exclusively in brain regions**.** Out of the total collected brain enhancers, exclusive subset in which enhancers expressing solely in the forebrain (104), midbrain (55) and hindbrain (38) tissues were placed. The other subset incorporated enhancers expressing in either two (62) or three (12) of the aforementioned brain domains.

VISTA enhancer browser: VISTA enhancer browser (https://enhancer.lbl.gov/) is a largely employed *in-vivo* dataset of empirically verified tissue-specific enhancers. The current number of total enhancers available in the repository accounted to 1393 at the time of study. These enhancers were initially chosen based upon very long distance conservation between humans and non-mammalian vertebrates. They also include enhancers that showed extreme ultra-conservation among mammals (human, mouse and rat). The candidate putative enhancers typically ranging in length from 200 to 2000bp were then tested using transgenic mice assay. Embryos were collected at embryonic day 11.5. Those enhancers were called as true positive hits for which results were consistently positive in at least three of the embryos (Visel, et al., 2007). It is however, important to note that enhancers that did not show reporter gene expression in three or more embryos at embryonic day 11.5 cannot be disregarded as true negatives, given the fact that the enhancers could be expressing in a different spatio-temporal setting, or could be assisted by the presence of additional cist-regulatory elements.

2.1.2 Sequence alignment and alignment segmentation

We collected enhancer orthologous non-human primate sequences through UCSC genome browser via BLAT (Karolchik et al., 2003; Kent, 2002). MAFFT was used to generate alignments for human and chimpanzee orthologous enhancer sequences by keeping macaque as an outgroup (Katoh, Misawa, Kuma, & Miyata, 2002). Gapped columns pertaining to a gap percentage of 1 were removed both from reference proxy and target enhancer regions for positive selection tests. Our strategy to probe enhancers for their relatively accelerated rate of evolution in human consisted of a segmented approach. As enhancer alignments show variant patterns of substitution rate over their entire length, from parts which are highly conserved among the three species to those which are highly variable in one or more of the lineages. Based upon the number of substitutions in the human lineage by keeping chimpanzee and macaque for orthologous comparisons, we partitioned the length of each enhancer into segments or blocks, each one to be of at least

300bp in length. Enhancers with no invariant patterns of substitution rate as a whole were kept in their original length. However, for each enhancer with variant substitution range, all partitioned blocks were checked for signals of positive selection and ones with closest positive values were shortlisted for further investigation.

2.1.3 Determining fast evolving enhancers via proxies

For determining the accelerated rate of evolution in human lineage with that of aforementioned non-human primate orthologous enhancer sequences, we used the strategy provided by Haygood and co-workers (Haygood, Fedrigo, Hanson, Yokoyama, & Wray, 2007), originally used to expound signals of positive selection on promoter sequences (Pond, Frost, & Muse, 2004). Our analysis carried three-species alignment (human-chimp-macaque), the minimum number of sequences allowed by the methodology. The methodology uses phylogenetic, branch specific approach for estimating positive selection over the sequences (J. Zhang, 2005). Intronic and non-coding, non-repetitive, loosely conserved sites (NCNRS) were used as proxies to determine signals of positive selection.

Wong and Neilsen Approach to detect selection on non-coding regions: In a coding region, ω (omega) represents the ratio of non-synonymous substitution rate (d_N) to synonymous substitution rate (d_S). Positive selection is termed on a coding region for which the rate of non-synonymous substitutions is greater than synonymous substitution rate. To expound this, in a codon TCC coding for serine amino acid, a change C -> G at the wobble position would change the codon to TCG but would not affect the amino acid it is coding and would be narrated as a synonymous change (Wong & Nielsen, 2004). For any codon site that undergoes positive selection, the beneficial mutation is sustained and the rate of non-synonymous substitution becomes much faster than the synonymous substitution rate, hence $\omega > 1$. This model of codon substitution was introduced in pioneer studies from 1994 to 2000 in which maximum likelihood approach was employed to estimate the parameter values (Goldman & Yang, 1994; Muse & Gaut, 1994; Yang, 1997; Yang, Nielsen, Goldman, & Pedersen, 2000).

Positive selection however cannot be confined to coding regions as there are studies that indicated positive selection is also acting on the non-coding regions. In order to estimate selection on non-coding sites, Wong and Neilsen extended the previous models, by assuming that synonymous substitution rate is constant in both coding and non-coding regions and a parameter ζ (zeta) was introduced (Wong & Nielsen, 2004; Jianzhi Zhang, Nielsen, & Yang, 2005), also adopted in the work of Haygood and co-workers (Haygood, et al., 2007). ζ models the substitution rate in non-coding regions to synonymous substitution rate in the coding regions (Haygood, et al., 2007). The values of ζ can be interpreted on analogous terms with ω, such as:

$\zeta < 1$ if the site is undergoing negative selection

$\zeta > 1$ if the site is undergoing negative selection

$\zeta = 1$ if the site is undergoing neutral selection

2.1.3.1 Global Proxy

For a preliminary inquiry of signals of positive selection in the candidate enhancer regions, those with a probable chance were narrowed down using a global proxy. In contrast to Haygood and colleagues' approach, where a locus specific proxy residing within a 100kb distance from the region of interest to make sure substitution rate does not vary among the proxy and target regions, proxy regions comprising conserved introns 1 and 5 of *FHL1* gene were chosen. This proxy choice made the screening independent of considering any genomic mutational hot and cold spots and also the chromosomal context enabling us to identify regions that could be evolving fast in the human lineage under positive selection (Chuang & Li, 2004). It was estimated that high conservation of our proxies will affect the results as lesser number of substitutions in the proxy region compared with the enhancer region will result in many false positives and so was the case.

Shorter Intron 5 of *FHL1* gene: At first, shorter intron 5 of *FHL1* gene, highly conserved among the three species was used as a proxy. 86 enhancers (Forebrain: 31, Midbrain: 19, Hindbrain: 10, Midbrain/Forebrain: 13, Forebrain/Hindbrain: 5, Hindbrain/Midbrain: 5, Midbrain/Forebrain/Hindbrain: 3) passed this stage of test. Test statistic of P-value with 95% confidence level implies all enhancers to be under positive selection with a value less than 0.05. P-Values were corrected for false discovery rate (Q-values) for this first round of analysis. Complete table on all analyzed 271 enhancers with *FHL1* intron 5 for signals of positive selection can be viewed in Appendix -I: Table A1.

Longer Intron 1 of *FHL1* gene: 52/86 enhancers greater in alignment length than the proxy region intron 5 of *FHL1* also existed and to address the length parameters that state proxy and target region should at least be equal (Haygood, et al., 2007), we applied a bigger 35.4kb proxy region of intron 1 of *FHL1* gene to all of the 86 enhancers. We see an under estimation of results when enhancers' alignments less than the length of *FHL1* intron 5 were treated with much longer intron 1 of *FHL1*. Therefore, comparable proxy-target associations were looked for to maintain 52/86 enhancers greater than *FHL1*'s intron 5's length to be treated with *FHL1* intron 1. These adjustments resulted in a set of 44 enhancers with likely chances of positive selection. Since first introns are the longest and most conserved relative to other introns in transcriptionally active proteins, they harbor active regulatory entities and higher proportion of epigenomic marks (Park, Hannenhalli, & Choi, 2014). Also, greater the length of the intron, more chances there are for it be evolving under purifying selection. Consistent with the said facts, including *FHL1* intron 1 also posed effects on biasing the overall results. Complete table on all analyzed 86 enhancers with longer *FHL1* intron 1 for signals of positive selection can be viewed in Appendix -I: Table A2.

2.1.3.2 Local Proxy

The aforementioned strategy helped us in an initial shortlisting of an overestimated set of brain enhancers, one with a highly likely chance to result in signals of positive selection. To determine the extent of false positives, the 86 predicted fast evolving enhancers were subjected to a more rigorous, context based approach in which introns of a nearby gene residing within 100kb of an enhancer were selected to be the locus specific intronic proxies to compare them with the enhancer of interest. To avoid cis-regulatory entities that might lie in the center or start of longer introns, ends of introns, approximately 2500bp in length were chosen with upto 150bp from the end sequence removed to avoid splice signals that might affect the status of the proxy (Haygood, et al., 2007). First introns of the gene were also eliminated (Haygood, et al., 2007; Park, et al., 2014).

Non-coding, non-repetitive sequences (NCNRS): For enhancers bracketed by longer gene deserts, random, loosely conserved NCNRS were preferred. To see if comparable lengths of the global proxies and target sequence have actually affected the results, we performed locus specific proxy test on all of the initially gathered 86 enhancers. We found an accordance of most enhancers resulting through application of locus specific proxy with the ones found to be under signals of positive selection when comparable lengths of *FHL1* introns were used i.e. the set of 44 enhancers. This stringent criterion curtailed the set of brain exclusive human accelerated enhancers (BE-HAEs) to 15 (see Table 2.1). These 15 enhancers overlapping with the ones in the set of 44 enhancers include enhancers with positive selection signals can be viewed in Figure 2.1. Complete table for all 86 enhancers which were compared with locus specific reference proxy within a 100kb range for determining signals of positive selection can be viewed in Appendix -II: Table A3.

Table 2.1: Results with locus specific intronic/NCNRS proxy region for previously shortlisted 86 Enhancers**

VISTA ID	VISTA Coordinates (GRCh37/hg19)	Expression Domain	Alignment Length (bp)	Proxy Within 100kb	Proxy Coordinates	Proxy Alignment Length (bp)	Distance From Proxy (kb)	P-Value
hs192	chr3:180773639-180775802	Forebrain	889	FXR1	chr3:180585929-180700541	21996	73.1	0.04
hs1301	chr11:16423269-16426037	Forebrain	885	SOX6*	chr11:15987995-16761138	31688	Intragenic	0.02
hs526	chr4:1613479-1614106	Forebrain	620	SLBP	chr4:1694527-1714282	4644	80.4	0.03
hs540	chr13:71358093-71359507	Forebrain	452	NCNRS	chr13:71343593-71345593	1990	12.5	0.03
hs37	chr16:54650598-54651882	Forebrain	601	NCNRS	chr16:54687882-54690000	2111	36	0.02
hs1210	chr2:66762515-66765088	Forebrain	419	MEIS1	chr2:66660584-66801001	12410	Intragenic	0.03
hs847	chr4:42150091-42151064	Forebrain	327	BEND4	chr4:42112955-42154895	9570	Intragenic	0.03
hs1019	chr7:20838843-20840395	Forebrain	471	ABCB5	chr7:20654830-20816658	32041	22.2	0.006
hs1526	chr2:104353933-104357342	Forebrain	1381	NCNRS	chr2:104388797-104390900	2096	31.5	0.03
hs430	chr19:30840299-30843536	Midbrain	632	ZNF536 *	chr19:30719197-31204445	7435	Intragenic	0.0007
hs1366	chr6:38358690-38360084	Midbrain	1380	BTBD9	chr6:38136227-38607924	17735	Intragenic	0.03
hs1632	chr11:116521882-116522627	Midbrain	630	BUD13	chr11:116618886-116643704	10609	96.3	0.04
hs1726	chr18:49279374-49281480	Hindbrain	1092	NCNRS	chr18:49291974-49293480	1496	10.5	0.02

hs563	chr6:98491829-98493238	Hindbrain	416	NCNRS	chr6:98467400-98470038	2627	21.8	0.03
hs304	chr9:8095553-8096166	Midbrain/Fore brain	614	NCNRS	chr9:8107387 - 8108217	828	11.2	0.04

*** Proxy coordinates are given for non-coding, non-repetitive sequences (NCNRS) and genes lying within 100kb distance from the enhancer region are obtained for genome build GRCh37/hg19 from UCSC and Ensembl respectively*

** Proxy genes harboring other VISTA elements in their introns*

***SOX6**: hs1720, hs883, hs236, hs518, hs717, hs1301; **ZNF536**: hs384, hs82*

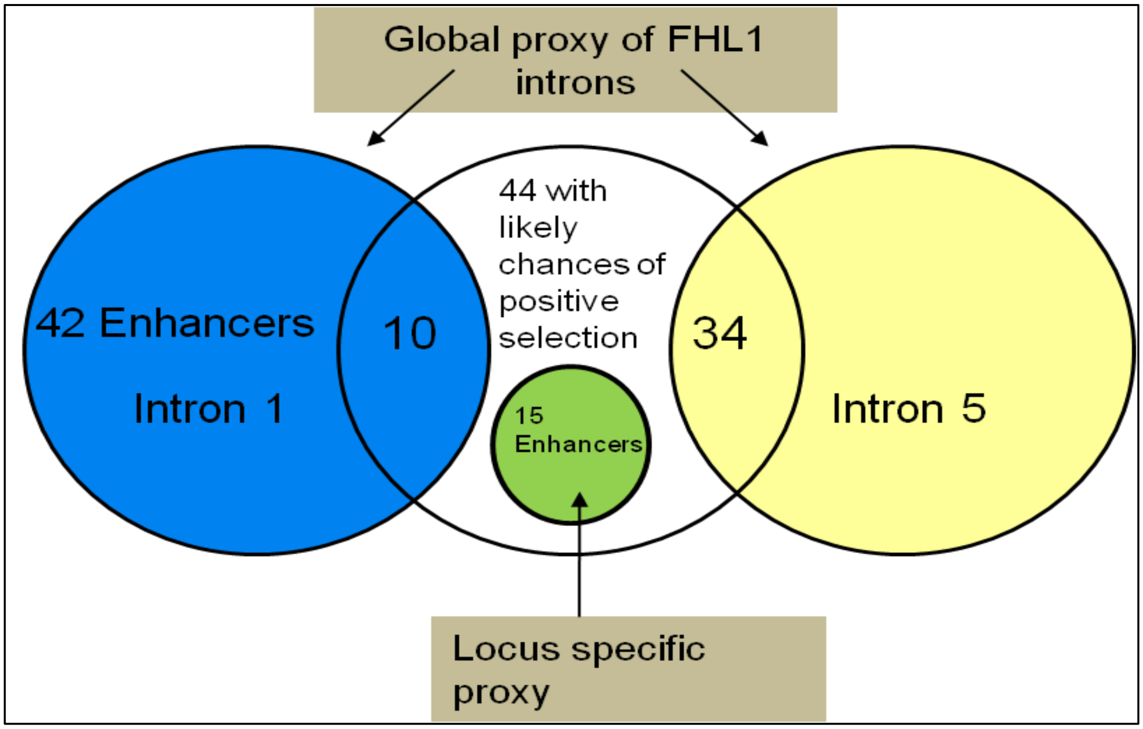

Global proxy of FHL1 introns
42 Enhancers
Intron 1
10
44 with likely chances of positive selection
15 Enhancers
34
Intron 5
Locus specific proxy

Figure 2.1: Results with Intronic Proxies of *FHL1* gene

86 positively selected brain enhancers were gained when FHL1-intron 5 of 1228bp length was applied to all 271 brain expressing enhancers. To see for length mismatches, we see shorter 1228 bp intron 5 has 34/86 enhancers resulted in terms of positive selection with length equal or less than the intron 5. Rest of the 52 enhancers checked had 10 enhancers with signals of positive selection with longer FHL1-intron 1 of 35.4kb. Locus specific proxy was applied on all 86 to avoid under or overestimation of the results. The resultant 15 enhancers through locus specific proxy in the next step came as a subset of previously collected 44 positively selected enhancers treated with introns 1 and 5 of FHL1 gene.

2.1.4 Associating target genes to accelerated brain enhancers

Enhancers gained from the aforementioned analysis were subjected to their surrounding genic environment and based upon orthologous genomic intervals in teleost fish, amphibians, reptiles, birds and monotremes, conserved target genes were predicted for each of the positively selected enhancer that the enhancer sequence could have a regulatory control on (Parveen et al., 2013)**.** The genes expression pattern was also confirmed through Mouse Genome Informatics (MGI: *in situ* RNA hybridization).

MGI (Mouse Genome Informatics): MGI (http://www.informatics.jax.org/) is an international consortium, an integrated effort that combines many component knowledegbases such as Mouse Genome Database (MGD) Project, Gene Expression Database (GXD) and others (Blake et al., 2016; Finger et al., 2016). It is a well-equipped platform employing mouse as a model organism to study disease and genetics in human. We employed this platform to confirm the expression of transcription factors in respective brain tissues and also for the target genes associated with the positively selected enhancers via RNA *in-situ* hybridization. RNA in situ hybridization is a technique that engages quantification of RNA transcripts by hybridizing it with complementary probes (Thomsen, Nielsen, & Jensen, 2005).

2.1.5 Assigning binding motifs to accelerated brain enhancers

On the resultant enhancers with signals of positive selection when compared with introns of the nearby genes and NCNRs within the 100kb vicinity, TFBS analysis was carried to develop a link between sequence acceleration with functional implications in terms of TFBS evolution. TRANSFAC repository was employed to collect transcription factor binding motifs of 142 brain expressing TFs. These TFs were confirmed via literature for their role in human brain development and were cross checked on MGI and Human Protein Atlas for their expression validity in any of the brain domain

TRANSFAC: TRANSFAC serves as one of the largest repositories containing information of eukaryotic gene regulation in the form of transcription factors and their binding sites (Matys et al., 2003). This regulation is mediated by the standalone or combinatorial action of various transcription factors that bind to specific sites on a DNA regulatory region. This repository refers to two main structures, one describes about the factors based upon their DNA binding domains and seconds encompasses similar binding sites occupied by different factors for the base preference depending upon the nucleotide frequency in each position of a binding site.

Human Protein Atlas: Human Protein Atlas (http://www.proteinatlas.org/) is a concentrated effort by Uhlen et al. that provides a unified platform for accessing human RNA and protein expression data at the single cell level (Uhlén et al., 2015). Tissue specificity and organ level information of human transcriptome was complemented with protein profiling via microarray based immuno-histochemisty. In this database, it was observed that almost half of the total number of putative genes under study expressed in all of the tissues, potentially controlling the housekeeping function and basic metabolic properties such as blood circulation, nerve function etc. In order to investigate the expression of the collected 142 TFs in brain, whose binding sites were gained from TRANSFAC, we referred to Human protein Atlas and found almost all of the factors expression in at least one of the aforementioned brain domains. This was done along with conformational data from MGI database for all of the TFs, and together reliable information about the expressional space of all the collected TFs was obtained and confirmed for brain tissue (Blake, et al., 2016; Matys, et al., 2003; Uhlén, et al., 2015).

2.1.6 Locating *Homo sapien* unique TFBSs in BE-HAEs

To see for modern human unique TFBSs, orthologous genomic segments from archaic humans (Neanderthal and Denisovan) were gathered (Meyer et al., 2012; Prüfer et al., 2014). The TFBSs found on human enhancers were catalogued against archaic human orthologs and non-human primate orthologs to enlist TFBSs that have evolved only among hominins or modern *Homo sapiens* owing to a substitution in the human lineage.

2.1.7 Human Population Genetics

To explore population dynamics over the allelic variants among the *Homo sapien*-unique TFBSs within the three BE-HAEs, 1000 Genomes Project Phase III data was employed to see the trend of natural selection among the human population (G. P. Consortium, 2015). Unphased VCF files from 1000 Genomes Project were converted to phased haplotype files through fastPHASE under default settings (Scheet & Stephens, 2006). In order to generate analysis that highlights the segregating alleles to be under the influence of positive selection, extended haplotype homozygosity (EHH) plots and relative EHH (rEHH) score were generated through package ‘rehh’ (version 2.0.0) and Sweep software respectively (Gautier & Vitalis, 2012; Sabeti et al., 2007). Weir and Cockerham F_{st} values were computed through VCFtools to estimate significantly differentiated SNPs between populations (Danecek et al., 2011; Weir & Cockerham, 1984). The haplotype range defined had 300kb region at either ends of the enhancer making up an entire region under consideration to be of approximately 600kb. Bearing in mind that human populations belonging to different ethnicities hone different adaptive mechanisms because of being exposed to variable climatic differences and changeable adaptive pressures (Tekola-Ayele et al., 2015), we catered to such vast yet delicate regional inconsistencies by dissecting our allelic deductions into regional and worldwide graphical representations.

1000 Genomes Project: A DNA sequence of a species at any given locus can present variable sequence forms. This sequence change that is common in a population is known as polymorphism. The classical definition of polymorphism, as also mentioned before, states that the least common allele at a given variant locus has to be >1% in a population (G. P. Consortium, 2010; Karki, et al., 2015). Human DNA sequence level polymorphism comes in multiple forms of SNPs and INDELS (INsertion/DELetion). 1000 Genomes Project is a platform in which relationship between genotype and phenotype is assisted by taking into account complete human genetic variation.

This publicly available repository aims at providing >95% of human genetic variants whose alleles fulfill the classical requirement of being present at >1% frequency in each of the five major regional demarcations. The five major groups include 2504 individuals' data in Phase III (dataset employed in this work) from Europe, Africa, South Asia, East Asia and America. Given the complete haplotype information is available on all the five populations at 1000 Genomes Project, this repository has been widely employed to associate common variants with disease combined with the LD structure in many GWAS studies. To associate a variant with a disease, the minor allele frequency has to be >5% in a population.

Haplotype: Haplotype is a combination of two words, haploid and genotype. While the former accounts for cells containing a single set of chromosomes, the latter means the entire genetic compliment of an individual. Haplotype, therefore refers to a set of markers on a chromosme that are inherited together from a single parent (O'Connell et al., 2014). Humans are diploid/biallelic beings. Each one of us receives two sets of chromosomes from both the parents (see Figure 2.2). Hence, we carry for a single chromosme two haplotypes, one from each parent. This accounts to haplotype inference or phasing which is a very crucial step in polulation genetics studies (Salem, Wessel, & Schork, 2005).

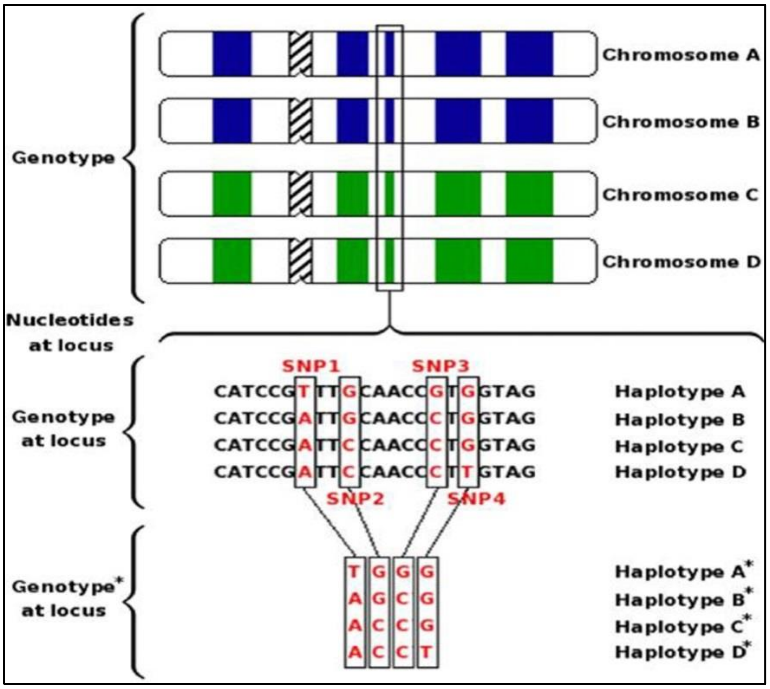


Figure 2.2: Haplotype structure for SNP data

For a heterogeneous genotype at a location in which two alleles are present, phased haplotypes help identify which allele belongs to which chromosomes. It also indicates the case of linkage disequilibrium in which alleles are inherited together. Adopted from: (Neigenfind et al., 2008)

fastPHASE: For a heterogenous genotype at a place in a genome that we may as well call a SNP location, it is necessary to asses which allele is coming from which chromosome (either metarnal or paternal). This information was attained via fastPHASE to order the genotypes so as not only to know which allele is coming from which side but also to know the set of alleles that have been inherited together. This helps estimate the idea of linkage disequilibrium , a crucial indicator of positive selection, in which alleles surrounding a positively selected allele are inherited together as agroup. Linkage disequilibrium can be defined where an allele offering a fitness advantage increases in frequency along with other neighboring alleles (Cadzow, et al., 2014).

Till date, several programs for haplotype phase inference on diploid genomes such as PHASE, BEAGLE, fastPHASE, IMPUTE2 and MACH have been developed (Browning

& Browning, 2011). Each one of them uses a different statsitical approcah and differs in measures of accuracy and efficiency according to the number of markers being analyzed and the sample size being used. fastPHASE and BEAGLE, however, are the most efficient softwares to analyze genome wide SNP data unlike their predecessor PHASE that accounted for 100 markers at maximum and a much smaller individual sample size (Browning & Browning, 2011). Our choice of fastPHASE relied on our relatively smaller sample size i.e. less than 1000 for each population (661 was the maximum number of individuals for African population at 1000 Genomes Phase III) which BEAGLE could not accurately handle (sample size should be >1000) (Browning & Browning, 2011).

EHH method: EHH or extended haplotype homozygosity test is a method designed by Sabeti and co-workers (Sabeti et al., 2002) to investigate the signals of recent positive selection on SNP data. This method was categorically employed in human population studies but was also successfully used in several other animals including cattle (Bomba et al., 2015; Qanbari et al., 2010). The neutral evolution theory states that allelic variants in a certain place in a genome can randomly increase or decrease in their frequencies in a population, the phenomenon called as the genetic drift. Under this assumption, an allele would have to undergo multiple rounds of recombination events that would subsequently decay the LD (inheritance of the neighboring alleles together) (Bomba, et al., 2015). However, for an allele chosen by natural selection, the sudden uncommon rise in its frequency does not have to undergo multiple rounds of recombination due to lesser numebr of generations and therefore LD is preserved. This makes the entire locus less diverse, which can be detected via EHH method (Bomba, et al., 2015).

rEHH method: Because of several shortcomings of the EHH test in resulting a number of false positives, rEHH or relative haplotype homozygosity test was designed as an extension to EHH test by Sabeti and co-workers, to assess the significance of selection signals (Sabeti, et al., 2007). This test takes into account all haplotypes that are made in a region with strong LD. In order to assess a haplotype with an alelle of interest, this haplotype will be compared to other control haplotyopes made in the same locus and the signal would be assessed for its true positivity (Bomba, et al., 2015). Haplotypes carrying alleles that are undergoing positive selection are reported signficant.

Haplotype Bifurcation diagrams: Bifurcation diagrams in human population genetics were introduced by Sabeti et al. in order to view breakdowns of linkage disequilibrium around an allele of interest (Gautier & Vitalis, 2012; Sabeti, et al., 2002). Little branching at the nodes depict lesser rounds of recombination events and hence longer unbroken haplotypes, maintaining linkage disequilibrium with the allele of interest. More branching at the nodes depict otherwise. These diagrams are an excellent visual aid to evidence long range homozygosity around the focal allele of interest (C. Zhang et al., 2015). The schematic illustration of the workflow design is shown in the folloiwng Figure 2.3.

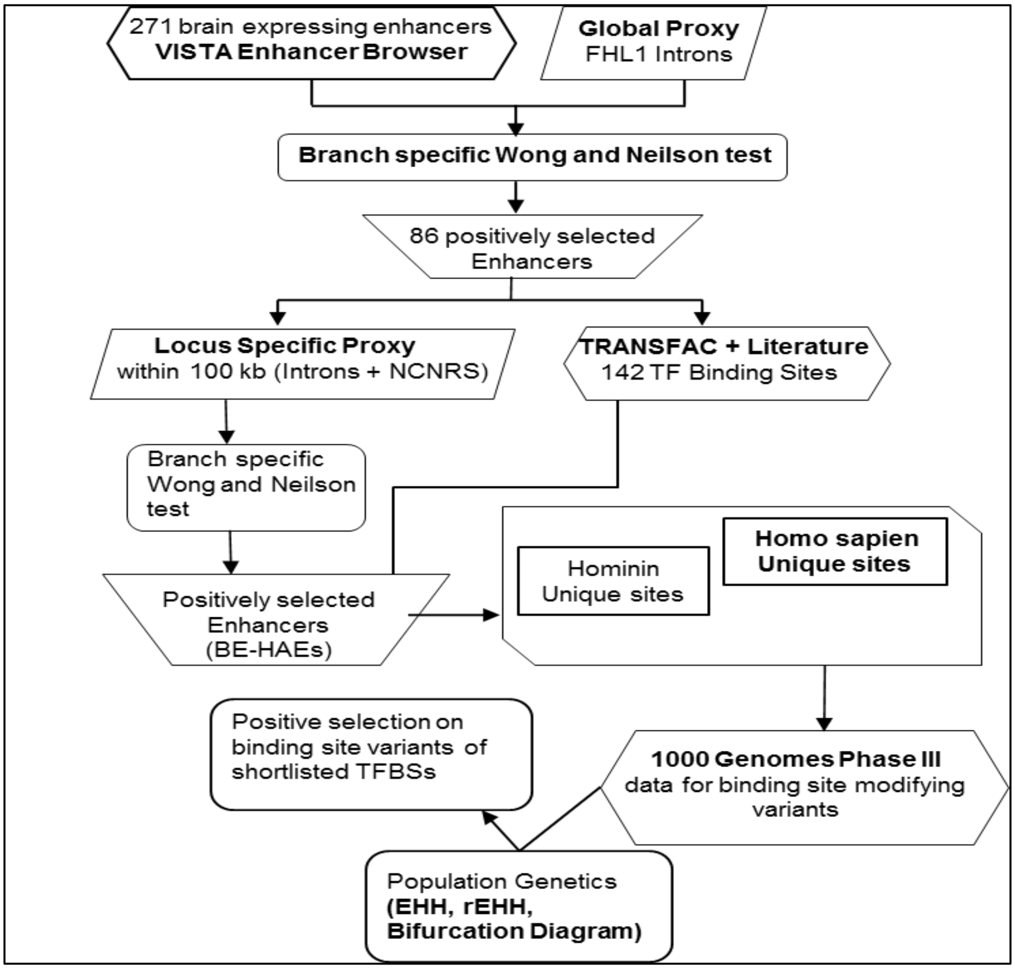

271 brain expressing enhancers
VISTA Enhancer Browser
Global Proxy
FHL1 Introns
Branch specific Wong and Neilson test
86 positively selected Enhancers
Locus Specific Proxy
within 100 kb (Introns + NCNRS)
TRANSFAC + Literature
142 TF Binding Sites
Branch specific Wong and Neilson test
Positively selected Enhancers (BE-HAEs)
Hominin Unique sites
Homo sapien Unique sites
1000 Genomes Phase III
data for binding site modifying variants
Positive selection on binding site variants of shortlisted TFBSs
Population Genetics
(EHH, rEHH, Bifurcation Diagram)

Figure 2.3: Schematic display of the carried out steps in the work design

All 271 enhancers collected from VISTA enhancer browser were subjected to global and local proxies. The subsequent TFBS analysis on the accelerated enhancers showed Homo sapien-specific sites on which different tests of population genetics were applied via using 1000 Genomes Phase III data.

2.2 SECTION 2: Nasal Morphological Variation

The genomic changes, either in the coding or non-coding parts of the genome, have manifested in a variety of morphological and anatomical traits that gave *Homo sapiens* a profound leverage over other hominoids. The present-day humans are unique in many aspects. To name a few, bipedalism, furless skin, brilliantly accessorized cognitive brain and associated reflexes, ability to analyze and assess danger and complications, evapotranspiration and nasal heat exchange make up a power set of skills that make humans an evolutionary dominant species (Lieberman, 2015). Probing into evolutionary dynamics of such traits and physiological features is undoubtedly a fascinating research area for scientists to explore.

In the second round of work in this section, we focussed on human nasal morphology that serves as a center stage to probe into highly variable facial features among modern humans. The correlation between an increased brain size in human evolution and its direct impact in reorienting the facial mechanics has been a topic of debate for many years. It is also an established fact that climate provides a strong basis for human adaptability. Thus, in order to see the climatic turn of events in shaping nature’s natural conditioning system in humans by also keeping in mind the ancient migrations from Africa, we collected single nucleotide polymorphisms (SNPs) that were successfully associated with important nasal traits in present-day human population and performed an intra-species analysis in present-day human population to see the trend of evolution.

2.2.1 Collection of SNPs associated with nasal traits

SNPs largely contribute to genetic variability. Many genome wide association studies have successfully accounted for a reasonable amount of genetic variation in human population that also governs many complex traits. These breakthrough studies have majorly contributed to elucidating the nature of dark matter in our genome that is still largely unaddressed. Nasal morphology is a highly variable trait in human population and climate is seemingly an important driving factor of various nasal forms in different regions of the world (Adhikari et al., 2016). From large, bulbous noses in Africa to narrower, taller and heighted nasal statures in Europe, reason lies in a crossroads territory of evolution, climate and adaptability. In order to gauge the genetic variability, owing to hundreds of SNPs that may govern the plethora of variable nasal forms, their heritability and a likely chance to spread in a population on an extraordinary pace, we sought out those SNPs that were strongly associated with prominent nasal traits. For this, we referred to six GWAS studies till date that significantly associated 25 SNPs with 8 nasal traits of nasal bridge breadth, columella inclination, midfacial height, nasion position, nasal width, nasal protrusion, nasal wing breadth and nasal height, exceeding the conventional threshold (Adhikari, et al., 2016; Lee et al., 2017; F. Liu et al., 2012; Paternoster et al., 2012; Pickrell et al., 2016; Shaffer et al., 2016).

2.2.2 Shortlisting of SNPs based upon derived allele frequency

In order to choose SNPs among the previously collected 25 SNPs that are potential candidates of being driven under the force of natural selection, and hence contribute to a region specific nasal shape, we exploited the instance of derived allele frequency (DAF). For an apparent difference in nasal morphology between two climatic extremes of hot-humid Africa and cold-dry Europe, we first enlisted traits that render visible differences between prominent nasal types of the two regions. Traits such as nasal width, nose size and nasal wing breadth are more prevalent in African population compared to opposite traits of elevated mid facial height, nasal protrusion observed in higher number in Europe. 10/25 SNPs were shortlisted as a result of this criterion. There also existed SNPs for traits like columella inclination, nasion position, nasal bridge breadth for which no data of higher occurrence in either of the climatic extremes exists. Therefore, we included all 4 SNPs contributing to these traits along with the 10 shortlisted SNPs for further analysis, hence totaling the number of analyzed SNPs to 14.

2.2.3 Human Population Genetics

In order to see the trends of positive selection on either of the variants of the shortlisted 14 SNPs, we referred to 1000 Genomes Phase III SNP data pertaining to 2504 individuals. Unphased VCF files from 1000 Genomes Project were converted to phased haplotype files through fastPHASE (Scheet & Stephens, 2006). For details of the methods used, see section 2.1.7.

RESULTS

3.1 SECTION 1: Enhancer Divergence

As of recent findings, human specific mutations in enhancers have brought to light the massive implications gene regulation can have on brain size and eventually on highly developed brain function in humans (Boyd, et al., 2015). We codified a strategy to find out the extent to which these human specific enhancer changes manifest in reshaping human brain circuits, and eventually characterizing *Homo sapiens* as the most successfully thriving members of the genus Homo. To pursue the investigation, we incumbently relied on an empirically verified, *in vivo* catalog of human brain-specific enhancers derived by Visel and colleagues for the root dataset of this study (Visel, et al., 2007). We conducted prioritized enhancer assortment obtained via transgenic mice assay to maintain reliabilty over ChIP-seq predicted putative enhancers that render a possibility of being eliminated as non-enhancers due to experimental artifacts or dubious nature of TF binding (Kvon, 2015). We then set out to construe sequence mutations within these enhancers and the rate at which they have proliferated in the human lineage, upon comparison with the closest relative chimpanzee taking macaque as an outgroup. As a result, we determined 15 such enhancers that consistently showed signals of acceleration in the human lineage when compared to orthologous non-human primate sequences upon both kinds of reference proxies. We term these accelerated enhancers as brain exclusive human accelerated enhancers (BE-HAEs) for future analogies (see Figure 3.1).

These 15 BE-HAEs presented a number of evolutionarily significant functional dynamics that could further be probed. For instance, identifying the gene body of the enhancer and determining its role in brain development can be resourceful in tracking down the whole mechanism with which the gene regulatory circuits are being evolved in the human lineage. Along with that, plethora of information can be attained on identifying the active binding sites within these fast evolving enhancers. Some very recent breakthroughs have successfully identified human specific variants in such active binding sites of transcription factors within the enhancers that markedly changed the phenotypes in mice.

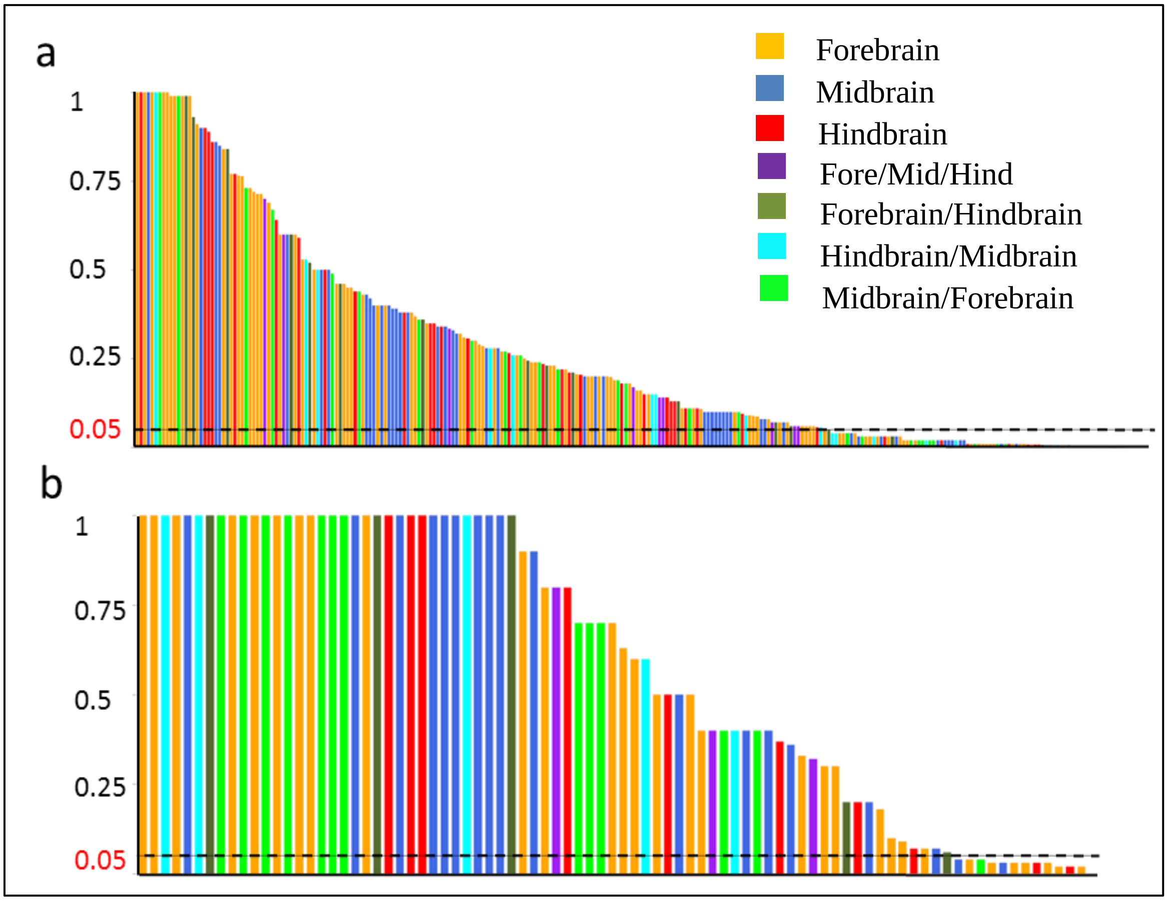


Figure 3.1: Test for positive selection using branch specific Wong and Nielson method with foreground branch human

(a) *Y-axis contains P-values. X-axis contains a total of 271 Enhancers. Each enhancer was compared and analyzed with conserved intron 5 of human FHL1 gene. 86/271 enhancers significantly indicated signals of positive selection (enhancers under the bar= P-value < 0.05).* ***(b)*** *Previously collected 86 enhancers in* ***(a)*** *were subjected to a robust analysis. Each enhancer was compared and analyzed with a locus specific intronic proxy from a nearby gene. This analysis contracted the previous findings to a number of 15 enhancers that were persistent in showing signals of positive selection (enhancers under the bar=P-value < 0.05).*

3.1.1 Associating target gene bodies with BE-HAEs

It is nonetheless surprising that less than 2 % of the human genome sequence comprised of the protein coding exons (I. H. G. S. Consortium, 2001; Venter, et al., 2001). The remaining much larger percentage of the non-coding part of the genome therefore went unnoticed. It is now in the age of functional genomics, that non-coding part of the genome is now being annotated and properly made use of in terms of its role in gene expression and regulation (Alexander, Fang, Rozowsky, Snyder, & Gerstein, 2010). In our findings of accelerated enhancers, also transgenically verified in mice, it became evidently intriguing to find gene bodies that these accelerated enhancers could spatio-temporally possess the control of. These enhancers, in combination with various transcription factors, precisely control the expression of genes in metazoans (Nolis et al., 2009). In quest of the gene bodies for our accelerated set of brain enhancers, we employed an approach in which comparisons were being made of the DNA blocks harboring our enhancer with the neighboring conserved syntenic blocks among vertebrates. Data reveals that many regions seen conserved among vertebrate genomes often harbor cis-regulatory elements and more often also reside within the bounds of a gene that has a significant role to play in early development (Amir Ali Abbasi et al., 2007). This approach is also superimposed with comparison of gene reporter expression pattern induced by enhancer in transgenic animal model with endogenous expression profile of the nearby genes. Coherence of the expression pattern of the enhancer with that of the gene, lying at least within 1 Mb on either side of the element, more often suggests functional association (Amir A Abbasi et al., 2010; Parveen, et al., 2013). Comparative genomics and expression pattern analysis of the genes and BE-HAEs therefore resulted in a putative set of target gene bodies. Complete illustrations of the syntenic conservation of the DNA blocks harboring the enhancer sequences among vertebrates are shown in figures below followed by a summary table of the 15 BE-HAEs with putative target genes whose endogenous expression profiles were confirmed from MGI and syntenic comparisons. Table 3.1 summarizes the gene associations with enhancers together with evidence from MG1. Figures 3.2 to 3.16 are detailed illustrations of all such associations.

Table 3.1: Evidence for Enhancer and Target gene Association

SN	VISTA ID	VISTA Coordinates (GRCh37/hg19)	Location	Expression domain	Target Genes	inter & intra genomic conserved synteny	MGI Expression data
1	hs37	chr16:54650598-54651882	Intergenic	Forebrain	IRX3/IRX5/IRX6/TOX3*	√	√
2	hs526	chr4:1613479-1614106	Intergenic	Forebrain	FGFRL1/CTBP1/SLBP/TACC/FGFR3*	√	√
3	hs847	chr4:42150091-42151064	Intragenic	Forebrain	LIMCH1/PHOX2B/TMEM33/SHISA3*	√	√
4	hs1019	chr7:20838843-20840395	Intergenic	Forebrain	ITGB8/FERD3L/CDCA7L/RAPGEF5*	√	√
5	hs1210	chr2:66762515-66765088	Intragenic	Forebrain	SPRED2/MEIS1*	√	√
6	hs1526	chr2:104353933-104357342	Intergenic	Forebrain	MAP4K4/MRSP9/POU3F3*	√	√
7	hs540	chr13:71358093-71359507	Intergenic	Forebrain	DNAJC19/SOX2*	√	√
8	hs192	chr3:180773639-180775802	Intergenic	Forebrain	SOX6	√	√
9	hs1301	chr11:16423269-16426037	Intergenic	Forebrain	DACH1	√	√
10	hs430	chr19:30840299-30843536	Intergenic	Midbrain	CCNE1	√	√
11	hs1366	chr6:38358690-38360084	Intragenic	Midbrain	CMTR1/GLO1*	√	√
12	hs1632	chr11:116521882-116522627	Intergenic	Midbrain	CADM1	√	√
13	hs1726	chr18:49279374-49281480	Intergenic	Hindbrain	ME2/DCC*	√	√
14	hs563	chr6:98491829-98493238	Intergenic	Hindbrain	POU3F2	√	√
15	hs304	chr9:8095553-8096166	Intergenic	Mid/Fore	PTPRD	√	√

*Represents enhancers which are associated with more than one target gene

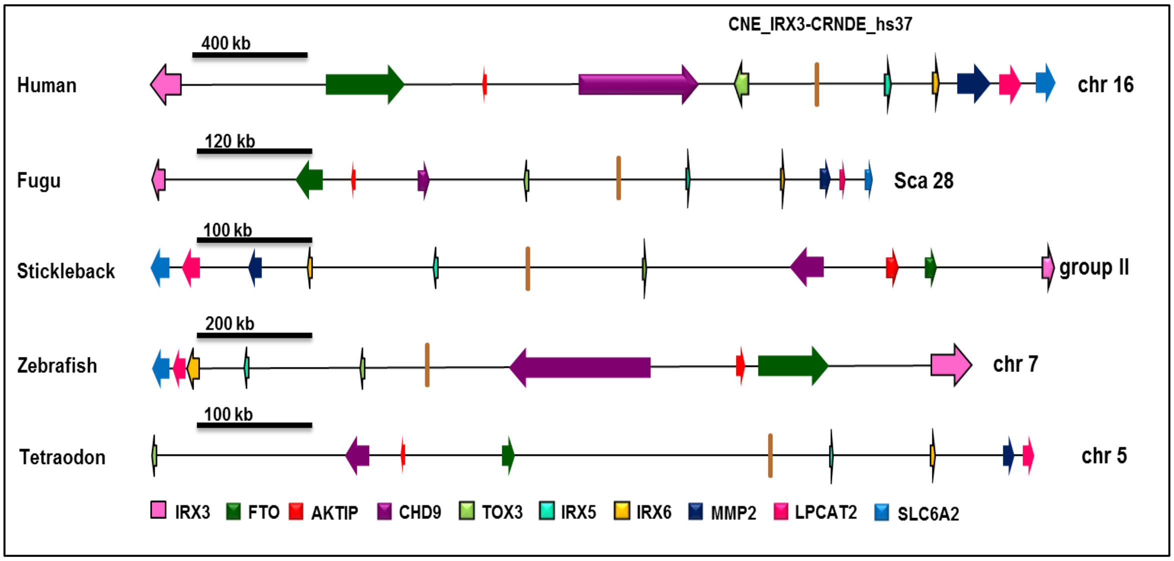


Figure 3.2: Syntenic evidence of associating target genes to a positively selected enhancer hs37

Genes controlled by a specific regulatory element have been identified through systematic analysis of the surrounding genomic content of the orthologous tetrapod-teleost lineages which may also harbor the functionally identified CNE enhancer. Keen analysis of expression pattern and conservation in tetrapod and fish lineages show the aforementioned enhancer is regulating the expression of TOX3 and IRX genes including IRX3, IRX5, IRX6.

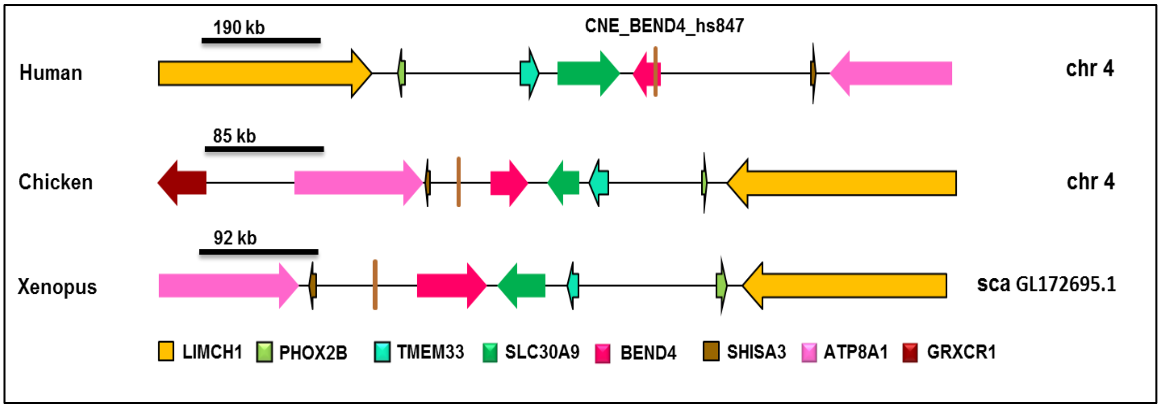


Figure 3.3: Comparison of human-amphibian syntenic conservation to help identify target genes of human enhancer hs847

Enhancer hs847 lies in an intron of BEND4 gene. Careful inspection of the corresponding genomic interval in an amphibian lineage strongly suggests BEND4 to be a target candidate of enhancer hs847. However, presence of other neighboring genes such as LIMCH1, PHOX2B, TMEM33 and SHISA3 and their positive endogenous expression profiles also make these genes suitable candidates as potential target genes of enhancer hs847.

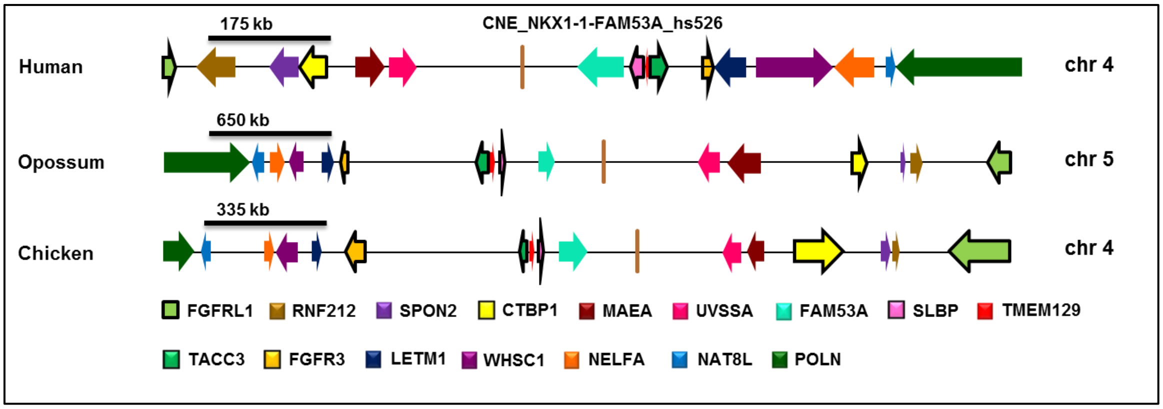


Figure 3.4: Comparison of bird-human syntenic conservation along with reporter gene expression data to help identify target genes of accelerated enhancer hs526

Systematic analysis of the surrounding region of enhancer hs526, conserved across the bird lineages helped identify a gene rich region surrounding the enhancer. Therefore, surrounding genic conservation and its verification through endogenous expression analysis potentially associates enhancer hs526 with FGFRL1, CTBP1, SLBP, TACC3 and FGFR3 genes.

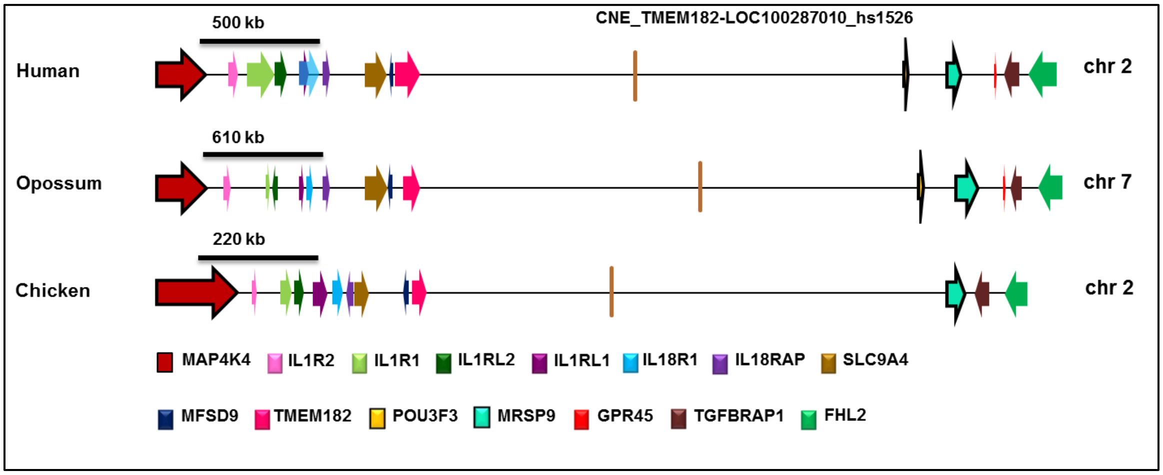


Figure 3.5: Syntenic evidence of associating target genes to a positively selected enhancer hs1526

*Comparative genomic analysis of functionally confirmed enhancer hs1526 across human-bird lineages and corresponding endogenous expression profiles of the genes associate the enhancer with MAP4K4, MRSP9 and POU3F3 genes***.**

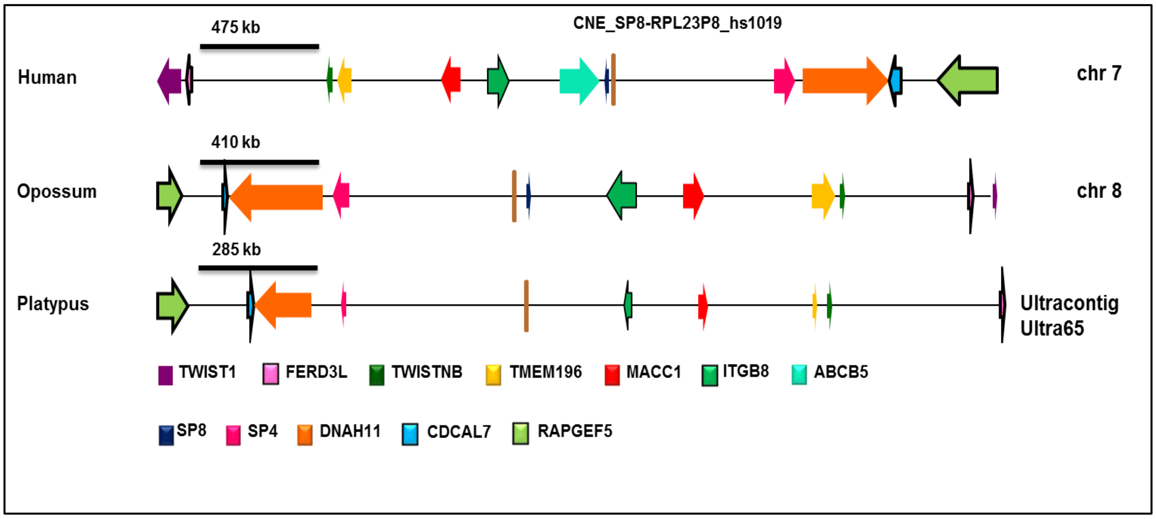


Figure 3.6: Assigning target genes to human enhancer hs1019 via analyzing genic content and expression pattern between human and platypus (monotremata) lineages

Orthology mapping was conducted down to Platypus. Despite of enhancer's closer proximity to gene SP8, conserved genic synteny and supporting endogenous expression profiles suggest FERD3L, CDCA7L and RAPGEF5 to be the target genes of enhancer hs1019.

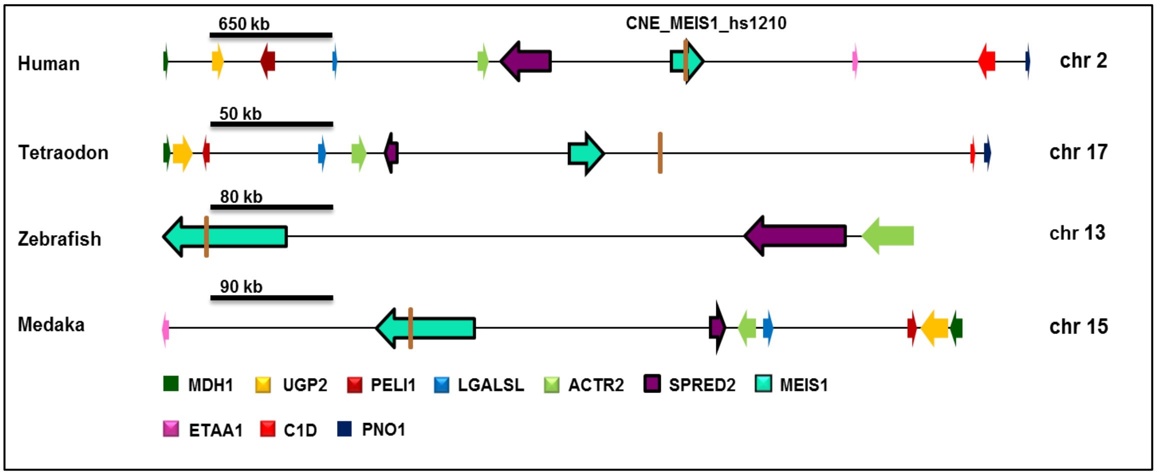


Figure 3.7: Syntenic evidence of associating target genes to a positively selected enhancer hs1210

The enhancer hs1210 was seen conserved down to teleost fish and is located within intron of MEIS1gene. The comparative syntenic analysis of human locus with multiple fish lineages strongly suggests the enhancer's association with MEIS1 and SPRED2 genes. Reporter gene expressions via MGI also support the expression of genes in the corresponding brain domain.

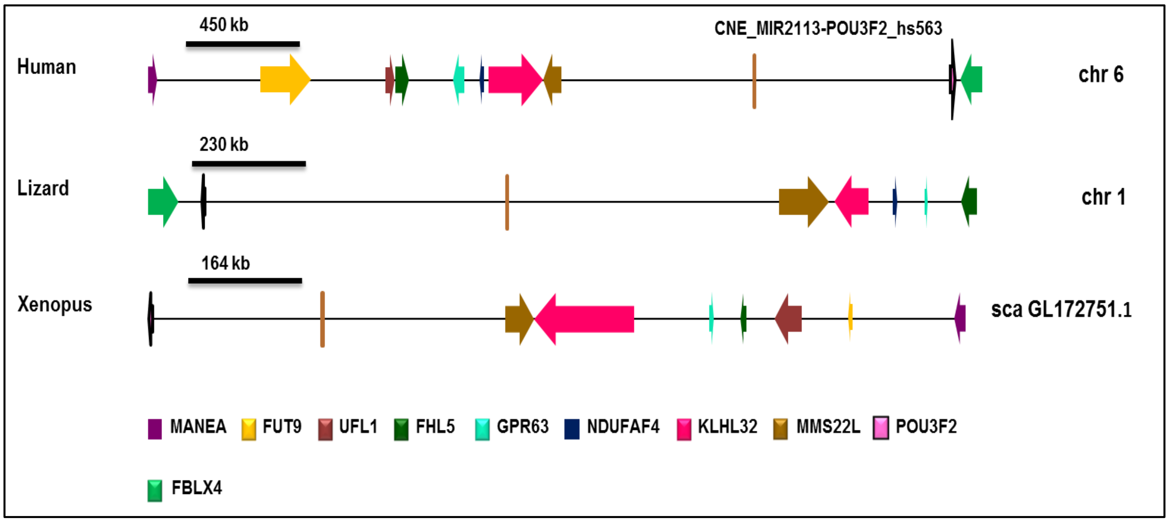


Figure 3.8: Syntenic evidence of associating target genes to a positively selected enhancer hs563

The interspecies genic conservation surrounding the enhancer hs563 and positive endogenous expression profiles of the genes via MGI suggest POU3F2 gene to be a potential target of enhancer hs563.

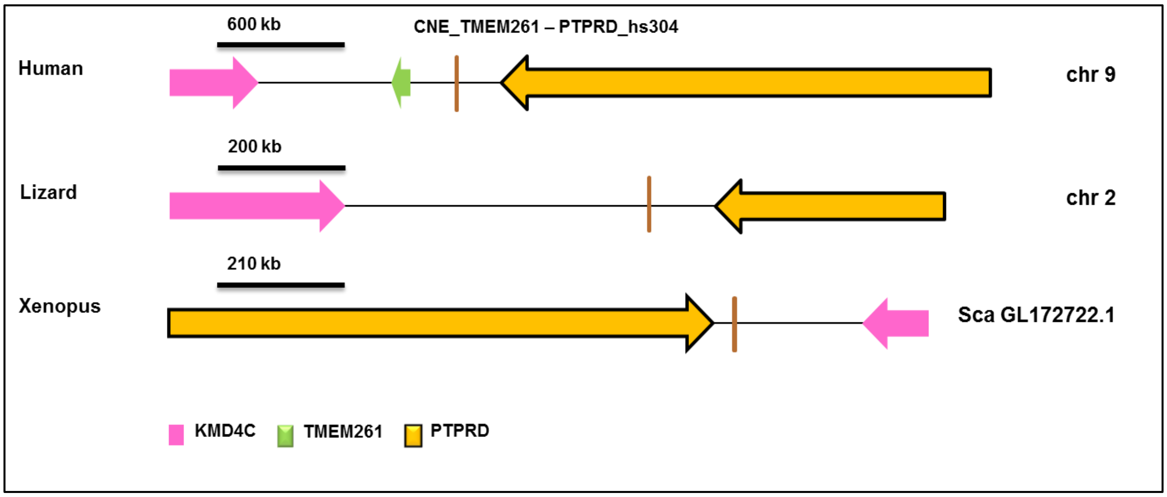


Figure 3.9: Assignment of target gene to functionally identified human accelerated enhancer hs304 through comparative genomics

Enhancer hs304 lies in intergenic interval between TMEM261 and PTPRD. The systematic comparative analysis of genic environment surrounding the enhancer hs304 across species also suggests the conservation of both the genes around the element. However, endogenous expression profiles suggest PTPRD gene to be the target of enhancer hs304.

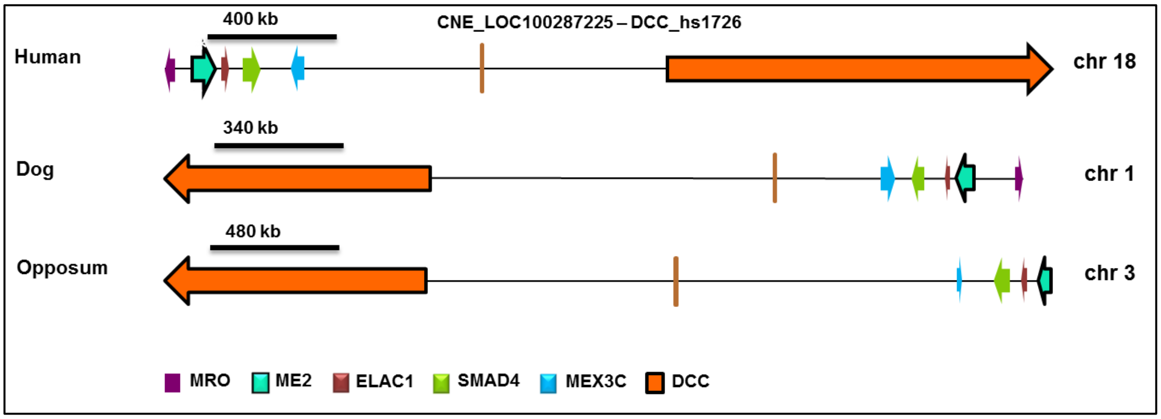


Figure 3.10: Syntenic evidence of associating target genes to a positively selected enhancer hs1726

Conserved intergenomic region across species and endogenous expression profiles of the surrounding genes suggest that the enhancer hs1723 is associated with ME2 and DCC genes.

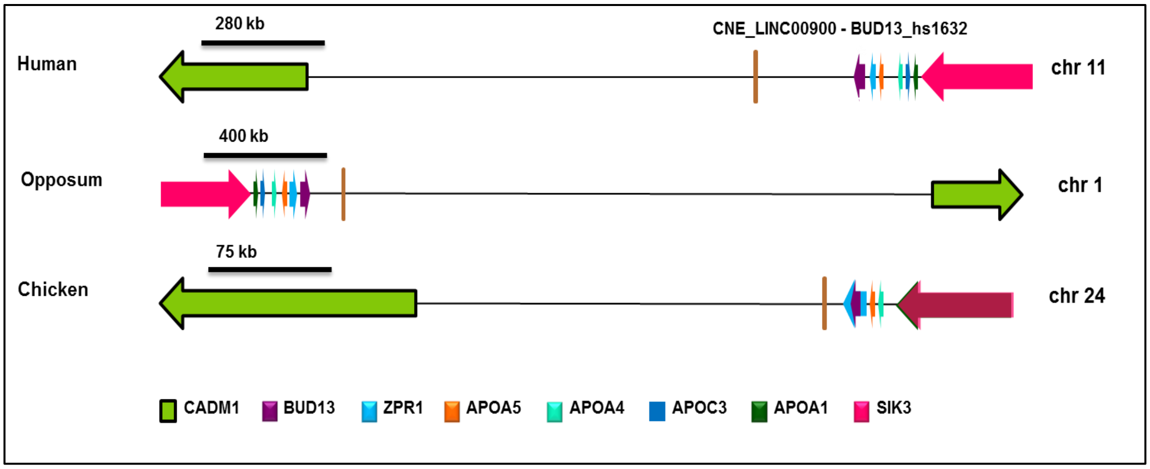


Figure 3.11: Comparison of accelerated enhancer hs1632's genic content and reporter gene expression leading to the identification of its putative target gene

Interspecies genomic conservation surrounding the element shows enhancer hs1632 to be near BUD13 gene and also suggests it to be the enhancer's probable target. However Endogenous expression profile analysis of the genes corroborates CADM1 as potential target gene of the enhancer.

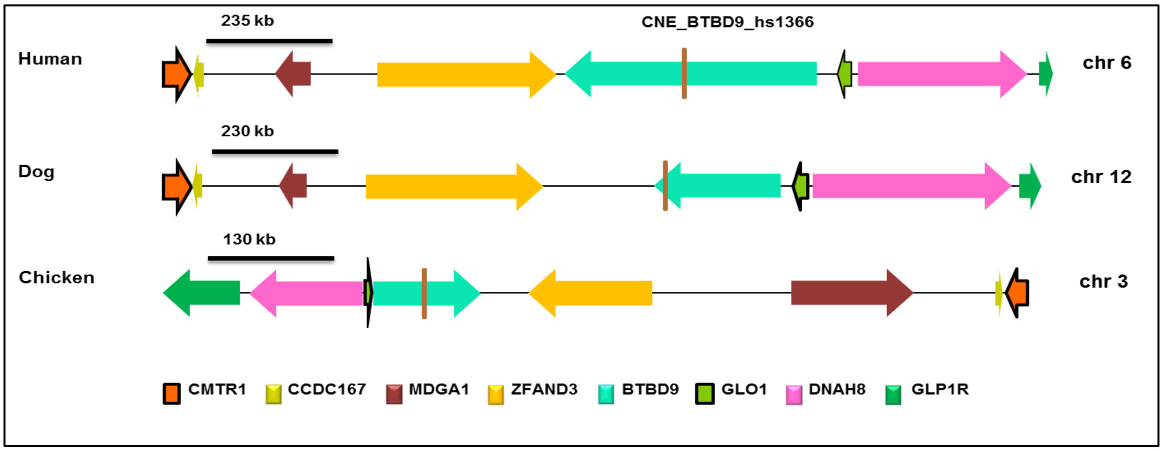


Figure 3.12: Syntenic evidence of associating target genes to a positively selected enhancer hs1366

Enhancer hs1366 lies in intron of BTBD9 gene, making it a highly likely putative target gene of the enhancer, also evidenced through interspecies syntenic conservation of the gene. However, endogenous expression profiles indicate an opposite scenario where genes such as CMTR1 and GLO1 are expressing in the corresponding brain domains. Along with the syntenic conservation, these two genes can be termed as the potential target genes of the enhancer hs1366.

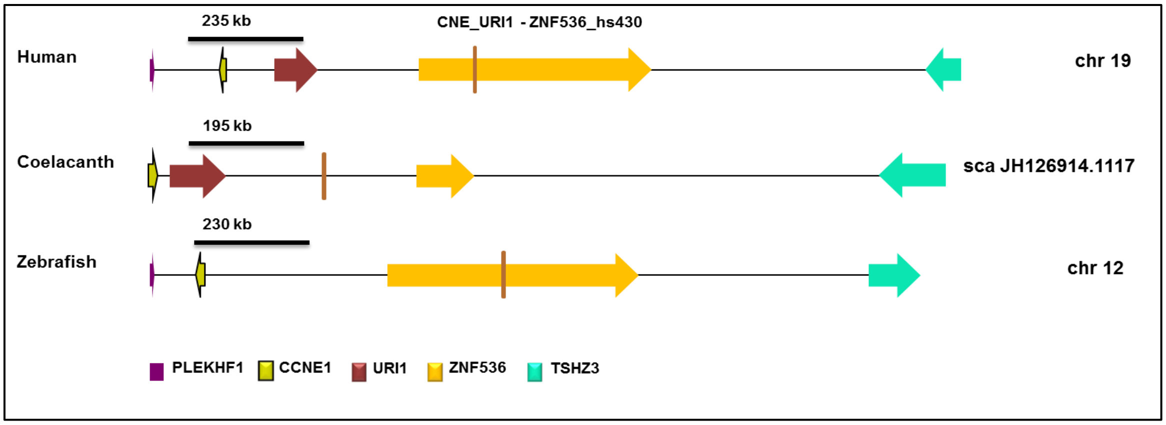


Figure 3.13: Target gene identification of human accelerated enhancer hs430 by tracing the genic context of its orthologous copies in teleost fish lineage

Enhancer hs430 lies in intron of ZNF536 suggesting it to be the target of the regulatory element. However, surrounding syntenic conservation and endogenous expression profiles of the neighboring genes, point out CCNE1 as the potential target gene of the enhancer.

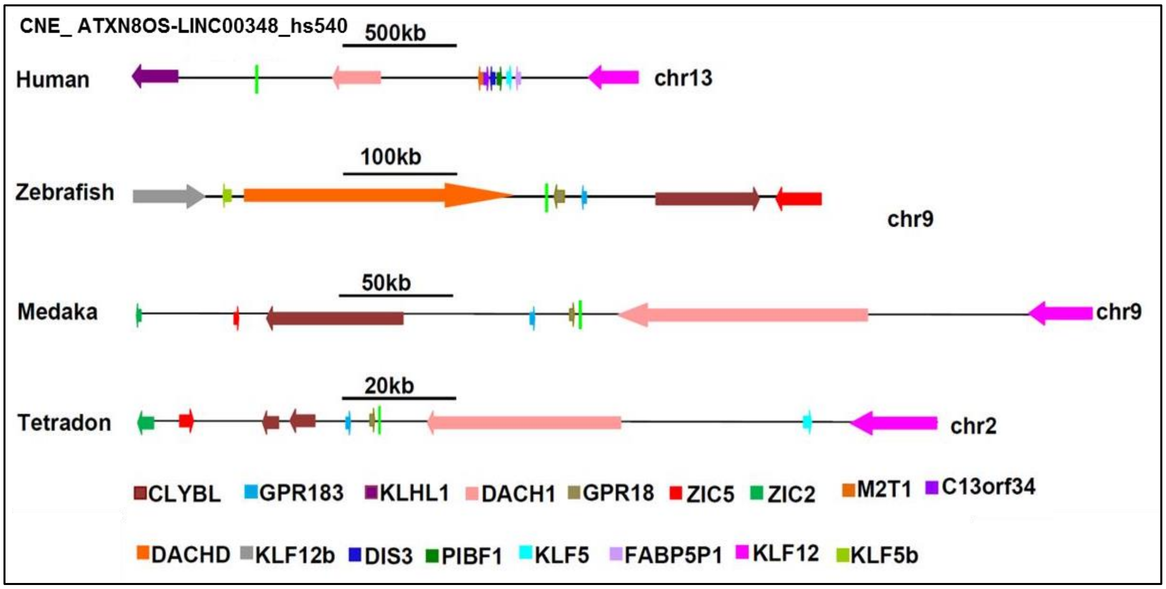


Figure 3.14: Comparison of bird-human and teleost-human genic content and reporter gene expression leading to identification of hs540's target gene

The above figure shows that analysis of human enhancer and its orthologous genomic interval in teleost fish lineage helps in associating the enhancer to its target genes. ***.*** *Analysis of expression pattern and conservation in tetrapod and fish lineages show that the enhancer hs540 is regulating the expression of DACH1 gene.*

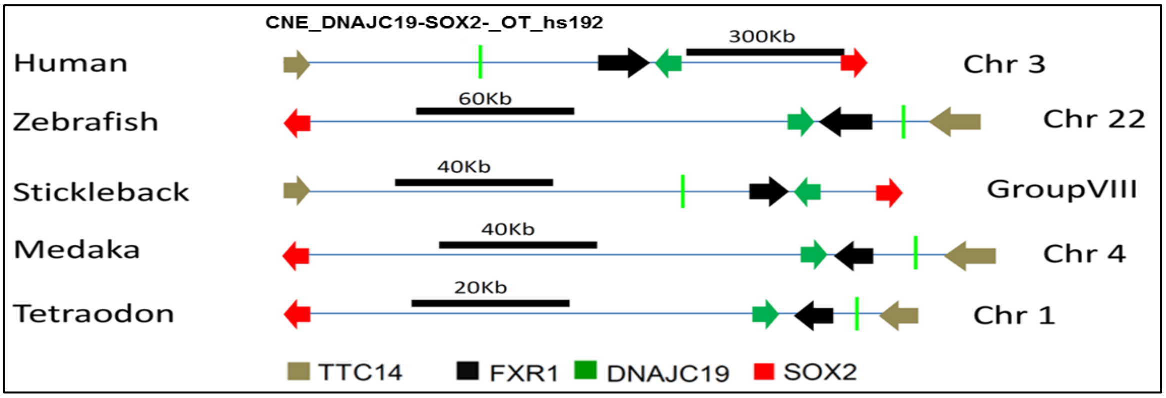


Figure 3.15: Association of accelerated enhancer hs192 by tracing the genic context of its orthologous copies in teleost fish lineage

The above figure shows that positioning of a human intragenic enhancer hs192 is conserved across human and teleost fish lineages; however, expression pattern indicates DNAJC19 and SOX2 to be the target genes of enhancer hs192.

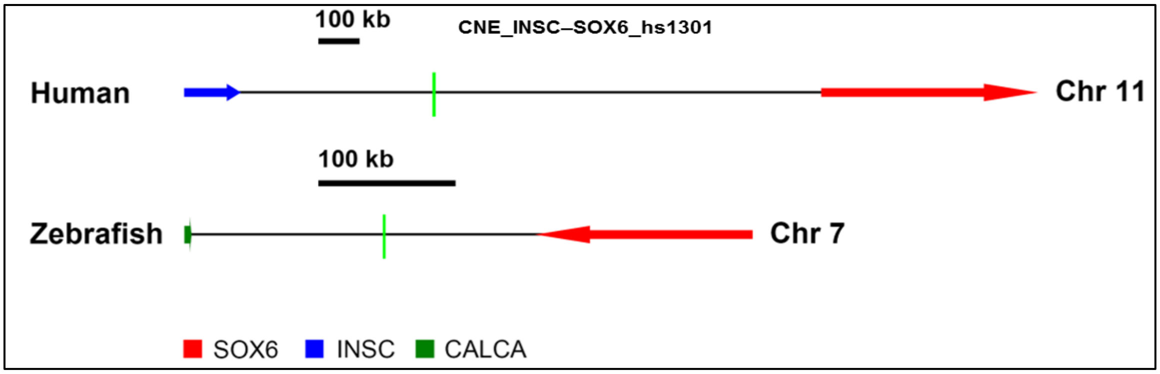


Figure 3.16: Target gene identification of human accelerated enhancer hs1301 through orthology mapping

The above figure shows that enhancer hs1301 positioned in the intergenic space between human INSC and SOX6 genes is similarly associated with the zebrafish SOX6 but not with INSC, resultantly suggesting that the human SOX6 is under the regulatory control of this intergenic enhancer.

3.1.2 TFBS analysis on BE-HAEs

Finding active regions on the annotated regulatory portion of the genome is a crucial step towards finding exquisite combinations of transcription factors that in cooperation with the enhancer regions instruct the expression of many developmental genes. In order to see whether there lies a set of transcription factors occupying the 15 previously gathered accelerated enhancers (BE-HAEs), TRANSFAC was made use of to collect motifs of 142 TFs. These TFs were carefully collected via literature survey. In order to further confirm the expressional presence of these TFs in the brain regions, MGI and Human Protein Atlas were also employed to get the validation. In our quest to find human unique binding sites within the 15 accelerated enhancers (BE-HAEs), we chose four non-human primate orthologous enhancer sequences (chimp, gorilla, macaque and orangutan) in comparison with the human counterpart. Upon careful inspection, 14 binding sites of TFs in 9/15 BE-HAEs were noted in which a unique sequence variant was present that made the site for respective TFs exclusive to human enhancer sequence. Details of all 15 BE-HAEs with their corresponding human unique TFBSs can be viewed in Table 3.2. Among the shortlisted 14 human TFBSs, we then set out to see which sites have originated solely in the modern human lineage. For this we gathered the archaic human orthologous sequences from Neanderthals and Denisovans. Upon close comparison among the three species of genus Homo, we came to find three sites of TFs SOX2, RUNX1/3 and FOS/JUND corresponding to BE-HAEs hs1210, hs563, and hs304 that seemed to have evolved in the present-day human lineage. It was however observed that the three *Homo sapien* unique sites within three BE-HAEs contained single nucleotide variants (SNVs) that made them unique to present-day human lineage. Pictorial depiction of the three TFBs of SOX2, RUNX1/3 and FOS/JUND can be seen in Figure 3.17. The remaining 11 sites can be viewed in Appendix -III: Figures A1-A11.

Table 3.2: Human unique transcription factor binding sites in a set of 15 brain exclusive enhancers with positive selection signals

SN	ID	GRCh37/hg19	Brain Domain	TF	TFBS
1	hs37	chr16:54650598-54651882	Forebrain	PEA3	ACWTCCK
2	hs1210	chr2:66762515-66765088	Forebrain	SOX2**	NNNANAACAAWGRNN
3	hs526	chr4:1613479-1614106	Forebrain	NF1B	CTGGCASGV
				POU3F2	NWAAYAAW
4	hs563	chr6:98491829-98493238	Hindbrain	RUNX1/3**	TGTGGT
5	hs1366	chr6:38358690-38360084	Midbrain	TCFAP2B	CCCCAGGC
6	hs1632	chr11:116521882-116522627	Midbrain	ZIC1	VGGGGAGS
7	hs1726	chr18:49279374-49281480	Hindbrain	-	-
8	hs1526	chr2:104353933-104357342	Forebrain	SOX9	RNACAAAGGVN
				PBX1	NYAYMCATCAAWNWNNN
9	hs847	chr4:42150091-42151064	Forebrain	LEF1	NWTCAAAGNN
				MEF2A	TATTTWWANM
10	hs540	chr13:71358093-71359507	Forebrain	-	-
11	hs1019	chr7:20838843-20840395	Forebrain	-	-
12	hs192	chr3:180773639-180775802	Forebrain	-	-
13	hs1301	chr11:16423269-16426037	Forebrain	-	-
14	hs430	chr19:30840299-30843536	Midbrain	-	-
15	hs304	chr9:8095553-8096166	Mid/Fore	FOS/JUND**	TGACTCA/TGACTCAN
				NR2F1	TGACCTY
				NURR1	YRRCCTT

TF: Transcription Factor
TFBS: Transcription Factor Binding Site
*** Modern Human specific TFBSs*

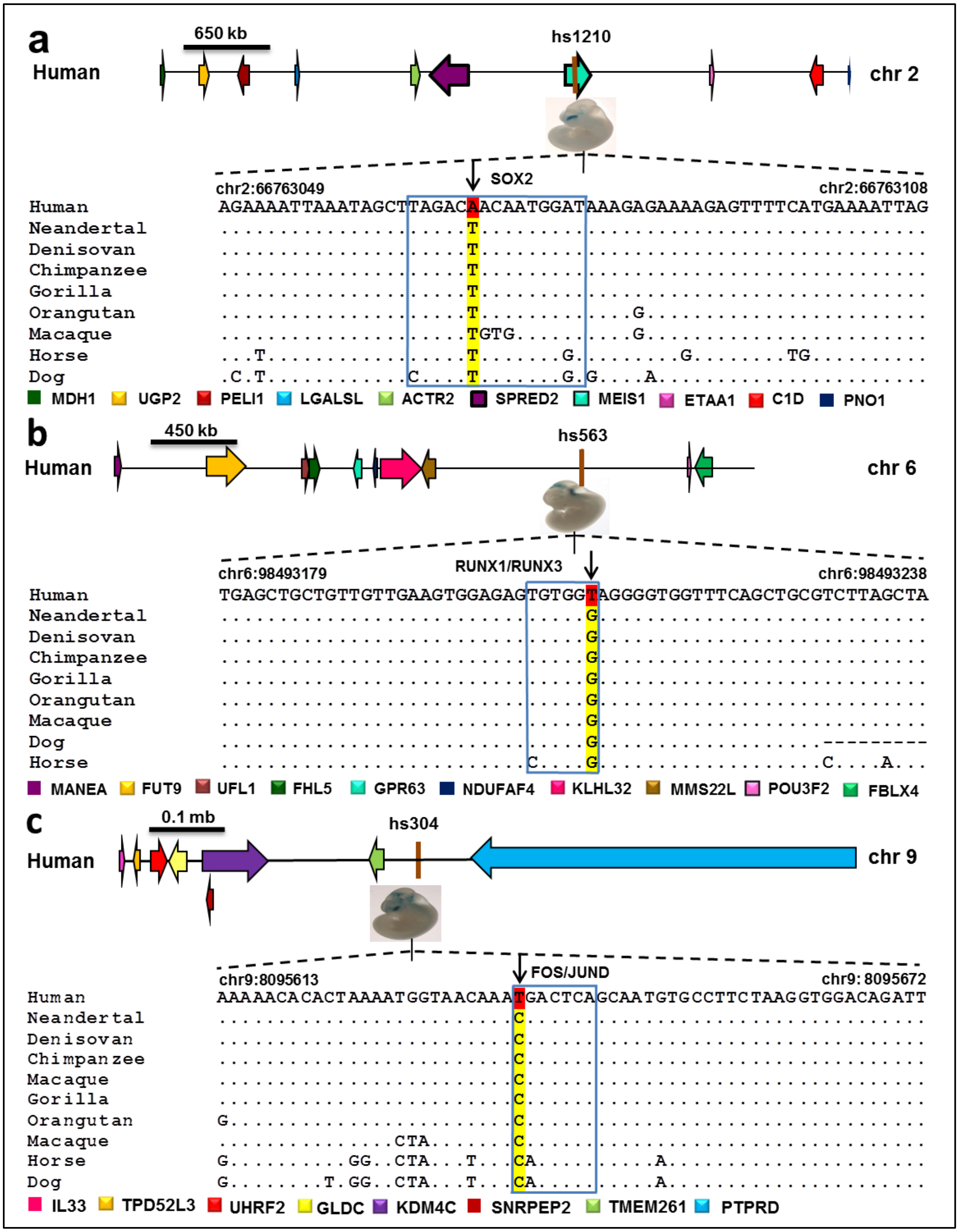


Figure 3.17: Human accelerated enhancers with *Homo sapiens* unique transcription

(a) *Human enhancer hs1210 (shown in brown) was shortlisted to be an enhancer under positive selection when compared with MEIS1 introns with a resultant P-value of 0.03. In this figure, an aligned patch within human forebrain enhancer hs1210 has been shown with an existing transcription factor binding site of SOX2. The region also showed a novel substitution within the* *binding site of SOX2 (TAGACA*ACAATGGAT) in the modern human lineage, unlike the consistent nucleotide observed for archaic humans, primates and non-primate mammals (TAGACT*ACAATGGAT).* ***(b)*** *Human enhancer hs563 (shown in brown) was shortlisted to be under positive selection when compared with a non-coding non repetitive sequence with a resultant P-value of 0.03. In this figure, an aligned patch within human hindbrain enhancer hs563 has been shown with the existing transcription factor binding motif of RUNX1/RUNX3. The region also showed a novel substitution within the binding site of RUNX1/RUNX3 (TGTGGT*) in the modern human lineage, unlike the consistent nucleotide observed for archaic humans, primates and non-primate mammals (TGTGGG*)* ***(c)*** *Human enhancer hs304 (shown in brown) was shortlisted to be under positive selection when compared with a non-coding non repetitive sequence with a resultant P-value of 0.04. In this figure, an aligned patch has been shown with the existing transcription factor binding site of FOS/JUND. The region also showed a novel substitution within the binding site of FOS/JUND (T*GACTCA) in the modern human lineage, unlike the consistent nucleotide observed for archaic humans, primates and non-primate mammals (C*GACTCA).*

3.1.3 Population Genetics

The three identified *Homo sapiens*-unique single nucleotide variants (SNVs) modifying the binding motifs of SOX2, RUNX1/3 and FOS/JUND were further substantianted as single nucleotide polymorphisms (SNPs), the difference lies in SNPs being at a >1% frequency in a population (Karki, et al., 2015). These SNPs corresponding to BE-HAEs hs1210, hs563 and hs304 have dbSNP IDs as rs11897580, rs2498442 and rs6477258, respectively (Sherry et al., 2001). It is understood that a SNP inhabiting a functional domain such as a TFBS can modify the enhancer sequence. The two or more sites that are created as a result might offer variable binding properties to the TFs (original or new TF), eventually creating activity bias for the enhancer they are occupying. However, some plausible outcomes can be expected about TFBS sequence structures that two variants of a SNP are creating, such as

1) the two variable TFBSs can retain the original TF binding property, may be through possible differential affinity,

2) the modified TFBS is impaired enough not to bind the original TF,

3) the altered TFBS can bind both original and new TFs,

4) the altered TFBS can bind only the new TFs, or

5) the altered TFBS altogether loses the ability to bind any TF (Heckmann et al., 2010).

As per conclusions, it is established that regulatory control over the genes has a major leverage in human evolution. Moreover, positive selection on such genomic regions that may influence a functional structure is another mainstream driving force to have revamped the current human status (Barreiro, Laval, Quach, Patin, & Quintana-Murci, 2008; Hussin, Nadeau, Lefebvre, & Labuda, 2010). To establish selection regime on such SNPs, we referred to 1000 Genomes Project Phase3 data and found derived alleles (TFBS modifying variants in *Homo sapien* lineage) of all three SNPs (rs11897580, rs2498442 and rs6477258) to be occurring near or below the intermediary frequency i.e. 0.5 and hence not fixed in the modern day human populations (Table 3.3).

Table 3.3: Derived allele frequencies and Weir and Cockerham Fst values of SNPs within enhancers hs1210, hs304 and hs563

Enhancer	SNP	TFBS	D/A*	Derived Allele Frequency					Weir and Cockerham F_{st} **				
				afr	amr	eur	sa	ea	afr	amr	eur	sa	ea
hs1210	rs4452126	-	T/C	0.075	0.005	0.001	0	0	0.1	0.006	-	-	-
	rs550939004	-	A/T	0.09	0.0014	0	0	0	0.15	0.013	-	-	-
	rs11897580	SOX2	A/T	0.13	0.006	0.001	0	0	0.2	0.01	-	-	-
hs304	rs6477258	FOS/JUND	C/T	0.28	0.25	0.29	0.34	0.32	0.0009	0.007	-0.0003	0.006	0.001
hs563	rs2498442	RUNX1/3	G/T	0.52	0.45	0.44	0.62	0.4	0.027	0.003	0.006	0.048	0.024

**D: Derived , A: Ancestral allele*
***Weir and Cockerham F_{st} calculated between one population and rest*
afr: Africa, amr: America, eur: Europe, sa: South Asia, ea: East Asia

3.1.3.1 Selection Signals gauged on binding site variants within three BE-HAEs

Exploiting the frequency and length of the haplotype with the variant at hand is resourceful in knowing the ongoing selection pattern on that variant and consequently its role in functional adaptation (Nielsen, 2005; Sabeti, et al., 2002; Voight, Kudaravalli, Wen, & Pritchard, 2006). In order to see whether the derived alleles of all three SNPs lie in a putatively selected haplotype, we investigated them based upon the work of Sabeti and co-workers and employed EHH, rEHH and haplotype bifurcation diagrams. (Sabeti, et al., 2002; Sabeti, et al., 2007).

3.1.3.1.1 SOX2 binding site modifying SNP rs11897580 within BE-HAE hs1210

We observed a modern human specific mutation at position Hsa2:66763070 in intronic region of gene *MEIS1*. This position falls within the predicted 15bp long binding motif of the transcription factor SOX2 residing in a 2.5kb VISTA annotated enhancer hs1210. Enhancer Sequence alignment reveals sequence identitiy with its ortholog in distant teleost fish medaka to be approx. 88%. Position 66763070 carries thymine residue in all of eutherian animals and in archaic orthologous DNA fragments. However, in current day humans, the position carrying thymine residue is replaced by adenine residue. Thus, it is estimated that the position 66763070 has been evolutionarily conserved in all of mammals (including monotremes) for approx. more than 180 Myrs and has recently been changed in modern humans (Luo, Yuan, Meng, & Ji, 2011). According to 1000 Genomes SNP data, the position 66763070 has not yet reached fixation among the human population, the ancestral state (residue thymine) still remains dominant in most current day humans than the derived state (residue adenine) except in African population where the derived state has reached a reasonable frequency (Table 3.3).

Elucidating BE-HAE hs1210, we observed core haplotype 4 (CH4) to be selected with the highest upstream rEHH value carrying the derived allele of the SNP rs11897580 (T>A) for a 2.5kb region in Africa (Table 3.4). In the same positively selected haplotype we observed another derived allele of the SNP (dbSNP ID: rs4452126:C>T) inhabiting the same HAE to be co-occuring or hitchhiking with our derived allele of interest. Hitchhiking has a typical signature of linkage disequiblirum with it i.e. the non-random association between the beneficial allele under positive selection and the neighboring alleles increases, giving less time to recombination to break the association (Hussin, et al., 2010). Hitchhiking effect has been limited to a region as low as 1kb and less for regions where recombination is high and variation is more (Fay & Wu, 2000). Noticeably, both derived alleles exist in more than 5% of Africans and absent/nearly absent elsewhere (Table 3.3). This makes the speculation that the derived alleles of the SNPs rs11897580 and rs4452126 are hitchhiking in African haplotypes, or have been positiveley co-selected for, implying sweep is underway in this region.

Furthermore, EHH plots and bifurctaion diagrams constructed for both SNPs indicated that the derived alleles are segregating under the clear influence of positive selection than their respective ancestral counterparts for a region as long as 10.8kb in Africans (Figure 3.18). To further confirm,Weir and Cockerham F_{st} test undertaken indicated that the two SNPs have statistically significant population differentiation between Africans and other samples implying that our allele of interest (SOX2 TFBS modifying allele) is segregating under the influence of positive selection in Africa (Table 3.3).

Table 3.4: Core haplotypes with SNP rs11897580 within enhancer hs1210 with each haplotype's rEHH score in African population

Core Haplotype (CH)		Hap Freq	rEHH (u, d)	rEHH P-value (u,d)
CH1	C **C** T **T** A G	370 (0.56)	0.04, 0.19	0.98, 0.56
CH2	T **C** T **T** A A	106 (0.16)	1.05, 1.12	0.59, 0.55
CH3	C **C** A **T** A A	59 (0.09)	10.17, 8.76	0.13, 0.16
CH4	C **T*** T **A*** A A	53 (0.08)	**48.51**, 11.95	**0.006**, 0.1
CH5	C **C** T **T** G A	40 (0.06)	1.62, 0.56	0.69, 0.92
CH6	C **C** T **A** A A	33 (0.05)	4.19, 2.39	0.2, 0.35
		Total=661		

Abbreviations: Hap Freq: Haplotype Frequency, u, d: upstream, downstream
**: unique derived variants of SNPs rs4452126 (T) and rs11897580 (A) in CH4*
The table enlists SNPs rs5006732, rs4452126, rs550939004, rs11897580, rs11681729 and rs10865355 in core haplotypes in a region of 2.5kb.

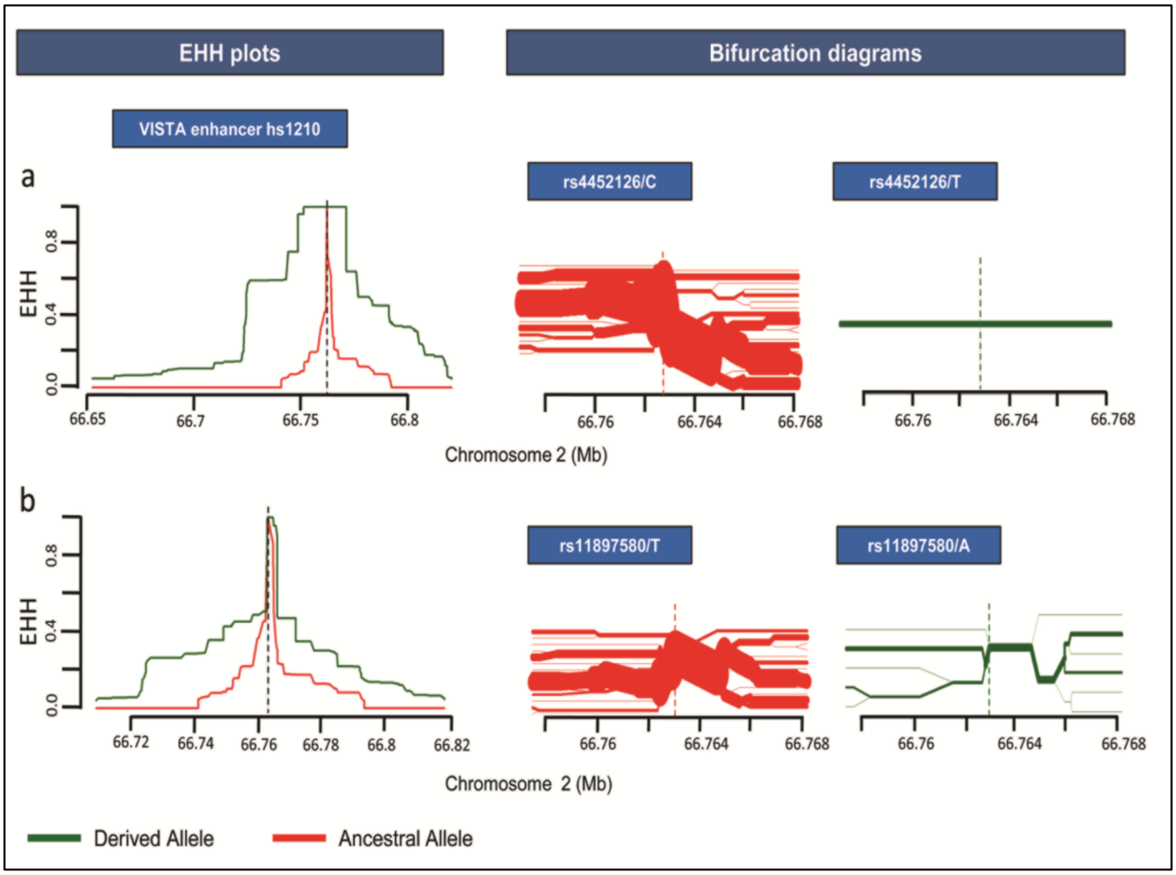


Figure 3.18: EHH plots and bifurcation diagrams of SNPs rs4452126 and rs11897580 belonging to forebrain expressing VISTA enhancer hs1210 in the African population

***(a)** EHH plot for SNP rs4452126 has a clear demarcation for derived allele T in terms of positive selection. EHH=1 indicates all haplotypes carrying either ancestral or derived state of the allele are matching upto this point. Bifurcation diagram of the derived variant of the allele confirms the deduction with a clearly long haplotype and absolutely no branching at the nodes upto 10.8kb region. **(b)** EHH plot for SOX2 TFBS modifying allele A of SNP rs11897580 also harbors evidence to be selected under positive selection compared to the ancestral allele T for a 10.8kb region. Bifurcation diagram uncovers little branching at the nodes interpreting for lesser recombination events and hence longer haplotypes for the derived allele compared to the ancestral variant T, especially for a 2.5kb region [chr2: 66762480-66764997] containing 6 SNPs (Table 3.4).*

Role of SOX2 in brain developemnt: SOX2 is a high mobility group (HMG) box TF characterized to be widely expressed in whole of neural tube, known to keep the progenitor chracateristic of the neural progenitor cell in both mature and developing CNS of humans (Beccari, Conte, Cisneros, & Bovolenta, 2012; Hutton & Pevny, 2011). Given the syntenic gene conservation around the enhancer region, *MEIS1* and *SPRED2* were assigned as target genes of VISTA enhancer hs1210. In a recent study, sproutly related protein 2 or *SPRED2* downregulation in adult zebrafish brain has been related to cell proliferation at the site of injury for neuronal repair. Myeloid Ectopic viral Integration Site 1 or *MEIS1*, is actively transcribed in developmental stages of the forebrain along with other *TALE* genes that are known to have distinguishing roles in cell differentiation and organogenesis (Barber et al., 2013). Thus, it is reasonable to speculate that SOX2 regulates the expression of *MEIS1* and *SPRED2* in developing and mature CNS.

3.1.3.1.2 RUNX1/3 binding site modifying SNP rs2498442 within BE-HAE hs563

We obeserved mutation at position Hsa6:98493210 within enhancer hs563 that falls within transcription factor binding motif of RUNX1 and RUNX3. At the sixth position of a 6bp binding motif, position 98493210 possesses a guanine residue till reptilian tetrapods and archaic humans and is replaced by thymine residue in modern human lineage. Thus, the ancestral allele (guanine residue) at this position has been evolutionarily conserved for more than 340 Myrs (Blair & Hedges, 2005). Mutation data from 1000 Genomes SNP data reveals that the derived allele (thymine residue) frequency is higher in African and South Asian populations (Table 3.3).

To assess for SNP rs2498442 (G>T) lying in BE-HAE hs563, haplotype construction revealed significant downstream rEHH P-value for core haplotype 1 (CH1) containing the derived state of the SNP again in Africans (Table 3.5). EHH plots constructed in a region wise manner, also depict positive selection in Africa in terms of greater area coverage indicating longer haplotypes and strong linkage disequiblrium with the derived state when compared to the rest of the regional plots (Figure 3.19a and 3.20). Global trend however indicates overall positive selection on downstream region for derived allele (Figure 3.22a).

Table 3.5: Core haplotypes with RUNX1/RUNX3 binding site modifying SNP rs2498442 within VISTA enhancer hs563 with each haplotype's rEHH score

Core Haplotypes (CH)			**Haplotype frequency**					**rEHH (u, d)**				
		Total	**America**	**Europe**	**South Asia**	**Africa**	**East Asia**	**America**	**Europe**	**South Asia**	**Africa**	**East Asia**
CH1	C G G **T*** T C T	1232	0.45 (156)	0.45 (227)	0.62 (303)	0.52 (344)	0.4 (202)	0.4, 0.5	0.76, 0.54	0.12, 0.32	0.3, **1.89**	0.23, 0.63
CH2	C A G **G** A T C	852	0.34 (118)	0.44 (221)	0.25 (122)	0.27 (179)	0.42 (212)	1.63, 1.7	0.76, 1.05	5.5, 2.13	2.03, 0.28	2.07, 0.92
CH3	T G T **G** A C C	344	0.2 (69)	0.11 (55)	0.13 (64)	0.1 (66)	0.18 (90)	1.87, 1.31	5.98, 6.07	5.46, 4.44	3.02, 2.27	2.5, 2.8
CH4	C G G **G** T C C	44	0.01 (4)	0	0	0.06 (40)	0	-	-	-	6.44, 0.64	-
CH5	C A T **G** A C C	13	0	0	0	0.02 (13)	0	-	-	-	5.57, 3.64	-
	Total	2492	347	503	489	649	504					

The table enlists SNPs rs62420423, rs9388046, rs4499937, rs2498442, rs2498443, rs13194250, rs2503789 in core haplotypes covering a 3.7kb region.
u,d: upstream, downstream
**: Derived allele T of SNP rs2498442 (T)*

Role of RUNX1/3 in brain developement: Runt related genes (RUNX) comprise of evolutionarliry conserved group of TFs that are highly responsible for maintaining lineage unique expression of the genes (Stifani & Ma, 2009). In mouse CNS, RUNX1 is produced in cholinergic branchial and visceral motor neurons of the hindbrain, whereas RUNX3 expression is confined to peripheral nervous system (Inoue, Shiga, & Ito, 2008). Synteny analysis of the enhancer reveals gene for CNS exclusive TF POU3F2 to be the associated target gene of the enhancer (Maricic, et al., 2013). Thus, time specific, well coordinated binding of POU3F2 TF alongwith other developmental factors on the nestin enhancer drives nestin gene expression in mouse, nestin protein being an adequate marker of neural progenitor cells in mammals that gives rise to neurons in neural tube in a developed nervous system (Jin et al., 2009).

3.1.3.1.3 FOS/JUND binding site modifying SNP rs6477258 within BE-HAE hs304

We observed a modern human specific mutation at position Hsa9:8095638 within VISTA annotated enhancer hs304 that falls at the first position of a 7bp long transcription factor binding motif of two transcription factors, FOS and JUND. The position carries an evolutionarily conserved cytosine residue in reptiles and higher animals including archaic humans for more than 340 Myrs and has recently changed to thymine residue in modern humans (Blair & Hedges, 2005). It is also known that FOS/JUN complex together with the help of other activating transcription factors, forms an activating protein 1 (AP-1) complex that binds to a plaindromic sequence of **5’-TGAC/GTCA-3’** (modern human specififc site of FOS/JUN in our study is **5’-TGACTCA-3’**), also known as TRE (TPA response element), majorly taking up the regulatory domains of many target genes (Cole & Josselyn, 2008).

For SNP rs6477258 (C>T) inhabiting BE-HAE hs304, no haplotype for any region was reported to have a significant rEHH with either the ancestral or derived state of the SNP. African population showed marked deviation in the EHH graph pattern from rest of the populations as well as the global trend, as prominent greater coverage under the curve on both sides of the graph and lesser branching with the derived allele in bifurcation diagram were observed than the counterpart ancestral allele upto a 4 mb region (Figure 19b, Figure 3.22b). EHH plots created for American, East Asian and South Asian populations with the SNP rs6477258 were in congruence with the global trend indicating downstream region with the derived state to have greater area under the curve except for European population (Figure 3.21).

Role of FOS/JUND in brain developement: Through synteny analysis, *PTPRD* gene was assigned as the putative target gene that the enhancer hs304 is potentially regulating the control of. *PTPRD* gene encodes for transmembrane protein, receptor-type IIa protein tyrosine phosphatase (PTPδ) that contains tandem repeat units of PTP domain in the intracellular region and is reported to have a role in tumor suppression in neuroblastoma, synapse formation and cell adhesion (Shishikura et al., 2016; Uetani et al., 2000). PTPδ expressed in the hippocampus region of the forebrain on deletion resulted in impaired learning capabilities because of the loss of synaptic plasticity that could in turn affect memory and learning in mice (Uetani, et al., 2000). Identification of a TRE element that is also a unique binding motif in modern humans to bind factors FOS and JUND in the regulatory element of *PTPRD* gene thus makes a safe assumption that FOS and JUND are together controlling the expression of *PTPRD* gene.

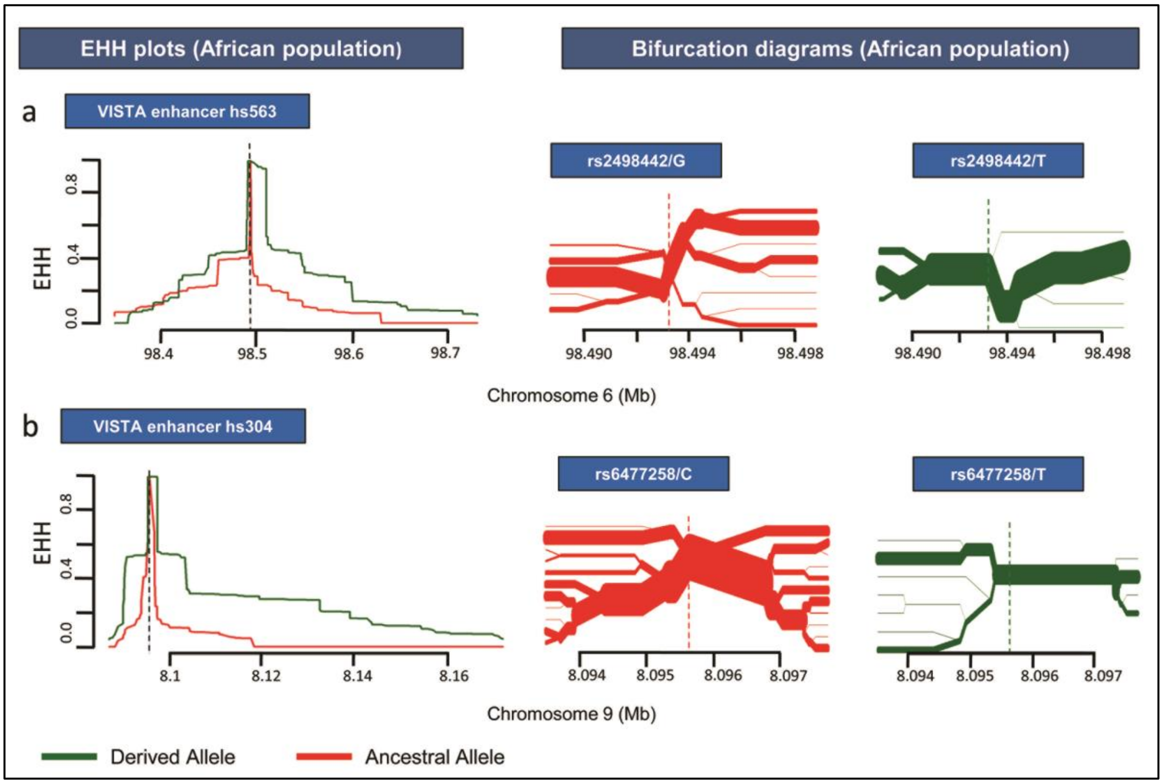


Figure 3.19: EHH plots and bifurcation diagrams for African population depicting SNPs rs2498442 and rs6477258 within VISTA enhancers hs563 and hs304 respectively

(a) *SNP rs2498442 within enhancer hs563 expressing in the hindbrain tissue. African Population shows a more pronounced EHH plot with the RUNX1/RUNX3 TFBS modifying derived allele T (shown in green) covering more area under the curve in the downstream region than the ancestral allele G (shown in red). Bifurcation diagram spanning a 10.25kb region (shown in green) has lesser branching showing lesser recombination events and making of longer haplotypes with the derived allele whereas ancestral allele has relatively more branching and shorter haplotypes in the same region.* ***(b)*** *SNP rs6477258 within enhancer hs304 expressing in the midbrain/forebrain tissue. EHH plot for FOS/JUND TFBS modifying derived allele T (shown in green) indicates greater area coverage in Africa on both sides when compared to the ancestral allele C (shown in red). Corresponding bifurcation diagram for Africa also reveal longer haplotype with lesser recombination events shown as branching at the nodes for TFBS modifying allele T than the ancestral allele C for a 4kb region.*

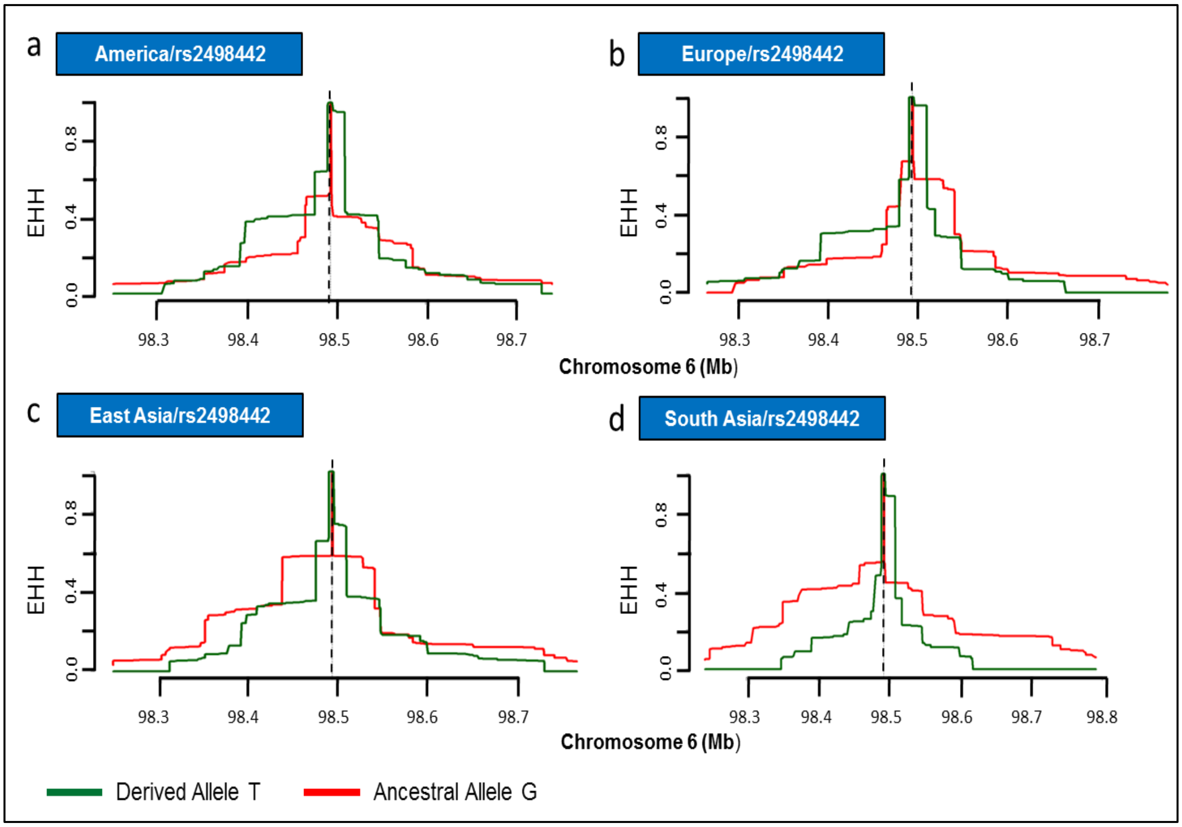


Figure 3.20: Comparative EHH plots for derived/ancestral variants of SNP rs2498442 within VISTA enhancer hs563

Figure S3 represents a comparative picture of all continental regions for ancestral and derived allele of SNP rs2498442 that lies within the TFBS of RUNX1/3 inhabiting hindbrain expressing VISTA enhancer hs563. ***(a)*** *America,* ***(b)*** *Europe,* ***(c)*** *East Asia and* ***(d)*** *South Asia depict a downward region with derived version of the SNP dominant except in East Asia. This downward trend of positive selection for all three populations (America, Europe and South Asia) is less prominent than can be seen as more pronounced in Africa (Figure 3.19a) on both sides of the EHH plot.*

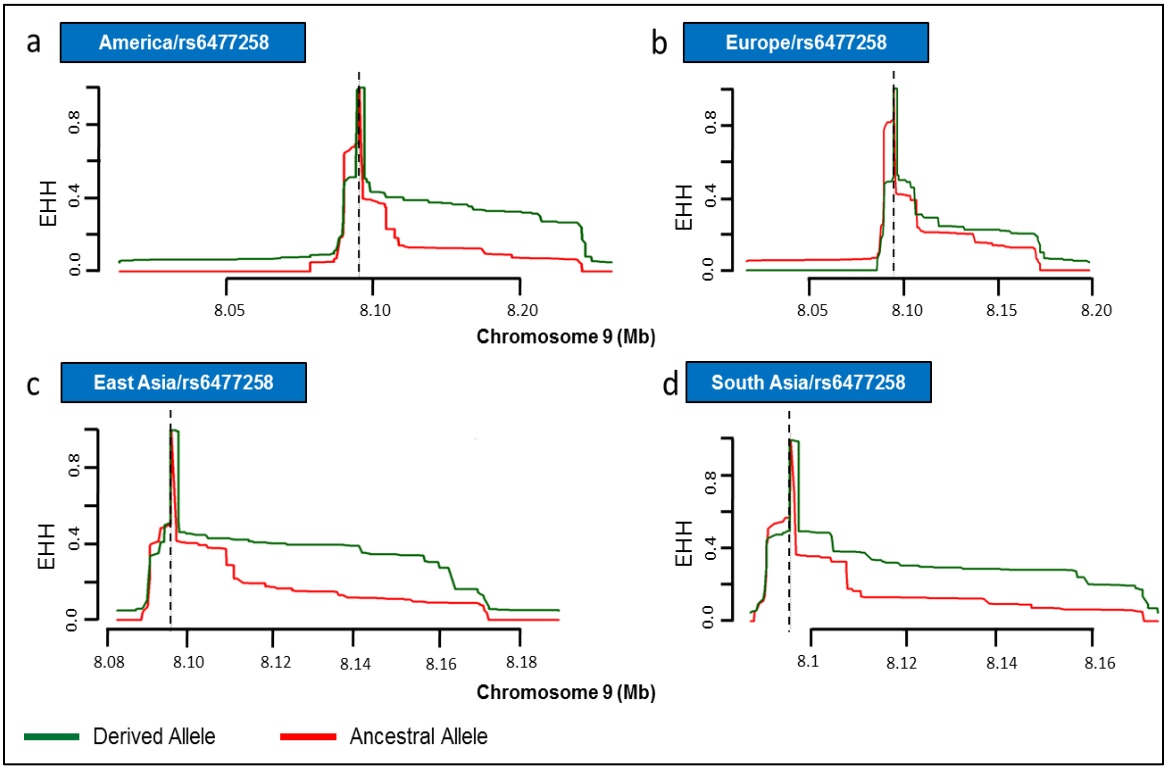


Figure 3.21: Comparative EHH plots for derived/ancestral variants of SNP rs6477258 within VISTA enhancer hs304

The above figure explains comparative picture of all 4 super populations for variants of SNP rs6477258 within FOS/JUND TFBS in VISTA enhancer hs304 ***(a)*** *America,* ***(b)*** *Europe,* ***(c)*** *East Asia and* ***(d)*** *South Asia depict a downward region with derived version of the SNP dominant except in Europe. The plots can be seen in comparison with the African EHH plot (Figure 3.19b) that has pronounced signal of positive selection for the derived allele on both sides of the graphs.*

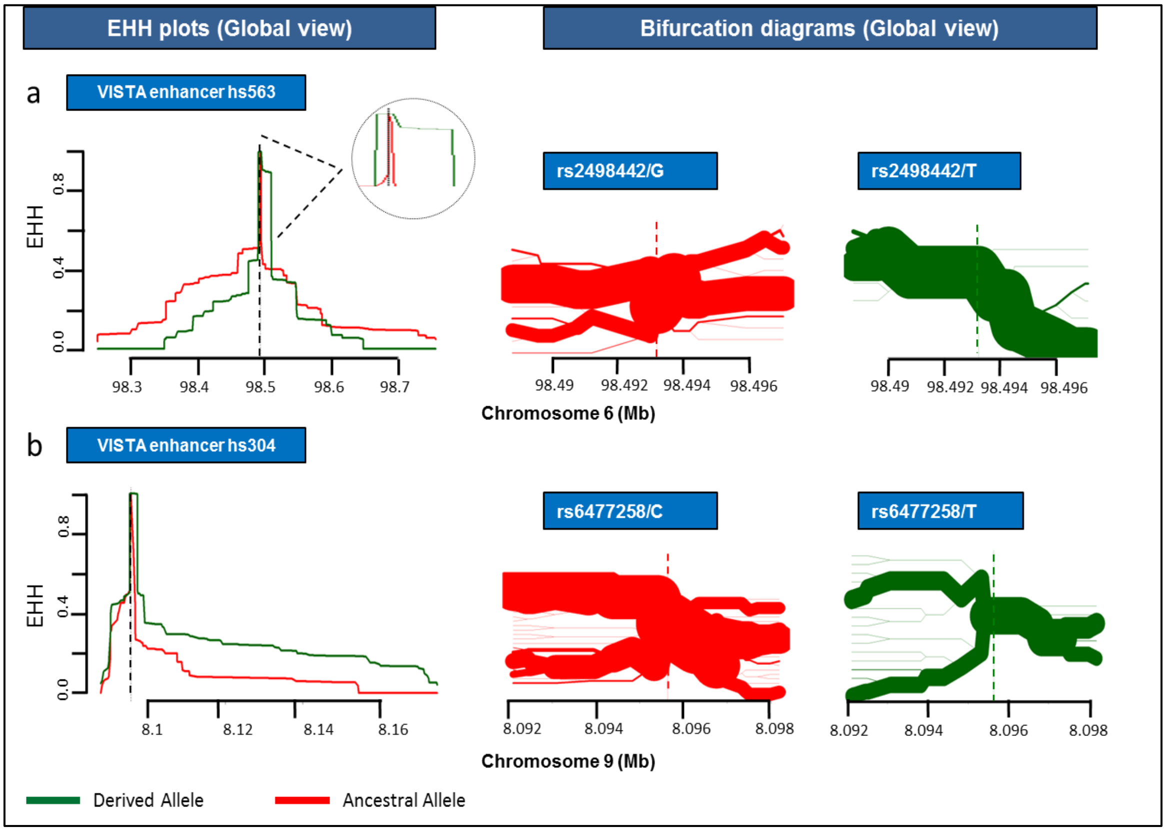


Figure 3.22: Global EHH plots and Bifurcation diagrams of SNPs rs2498442 and rs6477258 residing within VISTA enhancers hs563 and hs304 respectively.

Figure S5 narrates ***(a)*** *SNP rs2498442 in VISTA enhancer hs563. The figure represents an overall worldwide analysis of the SNP rs2498442 with 5 super populations' data available at 1000 genomes. In bifurcation diagram on the right side, 7 SNPs on either direction of the core SNP rs2498442 depict a genomic region of 10.25kb. Lesser branching refers to lesser recombination events. Longer haplotypes can be noted with the derived allele T (shown in green). The dotted lines on the EHH plot refer to the zoomed portion of the graph peak showing longest haplotype homozygosity with derived allele T (shown in green) within the mentioned region when compared to the ancestral allele G (shown in red).* ***(b)*** *SNP rs6477258 in VISTA enhancer hs304. Global trend of the segregating alleles of the SNP rs6477258 with 5 super populations data from the 1000 genomes is shown with derived allele possessing larger area coverage (shown in green) compared to the counterpart ancestral allele C (shown in red) in the downstream region. Bifurcation diagram on the right also show longer haplotypes with the derived allele T (shown in green) taking 10 SNPs on either side of the core SNP, a total region of 4kb.*

3.2 SECTION 2: Nasal Morphology

Human nasal morphology being at the core of craniofacial adaptation serves as a center stage to probe into highly variable facial features among modern humans. A large onus is set onto climatic extremes that people living in different parts of the world face (Evteev, et al., 2014; Zaidi, et al., 2017). Human nasal architecture because of its role in direct conditioning of inhaled air makes up a highly significant sub-domain of mid-facial morphology. As has been established that air conditioning of inhaled air in aquiline noses and therefore narrower, taller and more pointed nasal passages ensures cold-dry air in colder habitats to be first moistened and warmed before reaching lungs (Noback, Harvati, & Spoor, 2011; Yokley, 2009; Zaidi, et al., 2017). An opposite scenario prevails in hot-humid regions where large and bulbous noses perform otherwise. By keeping ancient migrations in mind that support out-of-Africa hypothesis, a drastic climatic shift was faced by early modern humans from hot and humid tropical environments to temperate and much colder climatic exposures (Nielsen et al., 2017). Thus, it is more likely that natural selection played a defining role in modifying nasal architecture so as to avoid complications that are inevitable in climatically challenged areas (Young & Mäkinen, 2010).

As markedly visible traits, facial features present as complex traits controlled by the net effect of epigenetics, environment (both non-genetic factors) and genome level variations (genetic factors) (Fagertun et al., 2015). In order to analyze the evolutionary trend of genetic factors that largely shaped up the basis for nasal variation in humans, we referred to all genome wide studies till date that associated SNPs with variable nasal morphologies (Adhikari, et al., 2016; Lee, et al., 2017; F. Liu, et al., 2012; Paternoster, et al., 2012; Pickrell, et al., 2016; Shaffer, et al., 2016). Among SNPs belonging to various nasal traits, we shortlisted 25 such SNPs that exceeded the conventional threshold of significance (P-value < 5×10^{-8}) (Figure 3.23, Table 3.6). Some of these traits are graphically shown in Figure 3.23b. In our investigation, by including orthologous sequences from Neanderthals and Denisovans, we found out 22 of the derived variants of the shortlisted SNPs to have arisen in modern *Homo sapiens* i.e. after their split from archaic humans (Table 3.6) (Meyer, et al., 2012; Prüfer, et al., 2014). Based upon visible differences in nasal measurements, two categories for two climatic extremes (Europe and Africa) were established in which larger nasal size, nasal width and nasal wing breadth belonged to Africa (Figure 3.23c) whereas traits such as greater nasal protrusion and midfacial height belonged to Europe (Figure 3.23d). To limit the number of analyzed SNPs in these two categories that posed as stronger candidates for probable contrasting nasal measurements in climatically extreme areas, we relied on the instance of derived allele frequency (Figure 3.23c & 3.23d). We enlisted 10/25 such SNPs that showed marked differences in their derived allele frequency, belonging to two categories of aforementioned climatic extremes (Figure 3.23c & 3.23d). These SNPs along with all four SNPs belonging to nasal traits that do not fall under the defined categories of two climatic extremes, such as columella inclination, nasal bridge breadth and nasion position, were considered for further analysis, hence, totaling the number upto 14.

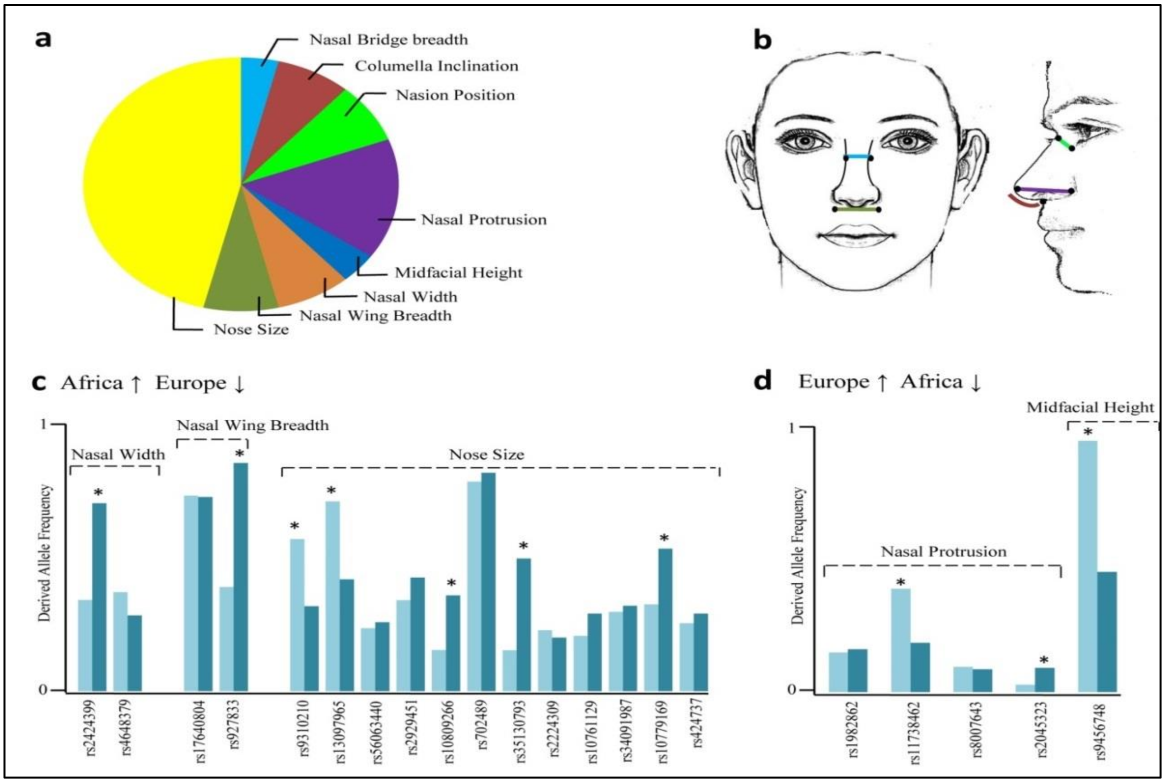


Figure 3.23: Shortlisting of SNPs associated with nasal traits based upon derived allele frequency

(a) *The pie chart shows the total number of 25 SNPs included in this study. The 8 nasal traits with which the SNPs are associated are given in differently colored sections such as nasal bridge* *breadth (in sky blue), columella inclination (in maroon), nasion position (in parrot green), nasal protrusion (in purple), mid-facial height (in dark blue), nasal width (in orange), nasal wing breadth (in olive green) and nose size (in yellow).* ***(b)*** *The color coded traits such as nasal bridge breadth (in sky blue), nasal wing breadth (in olive green), nasion position (in parrot green) nasal protrusion (in purple), columella inclination (in maroon) are graphically shown.* ***(c)*** *In bar chart, SNPs of traits such as nasal width (in orange), nasal wing breadth (in olive green) and nose size (in yellow), known to have greater measurements in Africa than the opposite climatic extreme of Europe (shown as upward and downward arrows) are grouped. SNPs were shortlisted based upon marked differences between their ancestral and derived allele frequencies. This difference is shown as largely varied bar heights representing the two allele frequencies* ***(d)*** *In bar chart, SNPs belonging to traits such as nasal protrusion (in purple) and mid-facial height (in dark blue), reportedly higher in measurements in Europe and lesser in Africa (shown as upward and downward arrows) are grouped. These SNPs were further screened for clear frequency differences in their ancestral and derived variants shown as largely varied bar heights representing the two allele frequencies. The asterisk symbol in both* ***(c)*** *and* ***(d)*** *shows 10 SNPs that were shortlisted for further analysis based upon marked allele frequency difference.*

Table 3.6: Genome wide significantly associated SNPs with nasal traits exceeding conventional threshold of P-value < 5×10-8 in six previously reported studies

		SNPs	Genic Context	Location	AA>DA	Nasal Trait	P-value	D/N	Derived Allele Frequency				
									Afr	Amr	Ea	Eur	Sa
Liu et al. (2012)	1	rs4648379	PRD1M6 (Intron)	1:3261516	C>T	Nasal width/length	1.13×10^{-08}	C/C	0.3986	0.2983	0.5526	0.3052	0.2607
	2	rs6555969	FGF18-SMIM23 (Intergenic)	5:171128464	C>T	Nasion position	1.17×10^{-09}	C/C	0.0734	0.3631	0.3403	0.3121	0.362
Paternoster et al. (2012)	3	rs7559271	PAX3 (Intron)	2:223068286	A>G	Nasion position	2.2×10^{-10}	A/A	0.553	0.5764	0.6111	0.3976	0.5399
	4	rs1982862	CACNA2D3 (Intron)	3:55064740	C>A	Nasal protrusion	1.8×10^{-08}	C/C	0.1528	0.2219	0.0754	0.165	0.1626
	5	rs11738462	C5ORF64 (Intron)	5:61013776	G>A	Nasal protrusion	1.8×10^{-08}	G/G	0.3979	0.1772	0.1637	0.1899	0.2055
Adhikari et al. (2016)	6	rs1852985	RUNX2 (Intron)	6:45329656	C>T	Nasal bridge width	6.0×10^{-10}	C/C	0.1694	0.2666	0.2411	0.1292	0.1145
	7	rs2045323	SFRP2-DCHS2 (Intergenic)	4:154831899	G>A	Nasal protrusion	1.0×10^{-08}	G/G	0.0287	0.268	0.2044	0.0934	0.2311
						Nasal tip angle	2.0×10^{-08}						
						Columella Inclination	3.0×10^{-09}						
	8	rs12644248	DCHS2 (Intron)	4:155235392	A>G	Columella Inclination	7.0×10^{-09}	A/A	0.0204	0.2176	0.1062	0.001	0.0716
	9	rs17640804	GLI3 (Intron)	7:42131390	C>T	Nasal wing breadth	9.0×10^{-09}	C/C	0.7867	0.572	0.9216	0.7813	0.816
	10	rs927833	PAX1 (Intergenic)	20:22041577	T>C	Nasal wing breadth	1.0×10^{-09}	T/T	0.4191	0.7867	0.9702	0.9185	0.9223
		rs7559271	PAX3 (Intron)	2:223068286	A>G	Nasion position	4.0×10^{-11}	A/A	0.447	0.4236	0.3889	0.6024	0.4601
Shaffer et al. (2016)	11	rs2424399	PAX1-NKX2-2 (Intergenic)	20:21632545	C>A	Nasal width	2.62×10^{-08}	C/C	0.3669	0.585	0.5982	0.7565	0.6636
	12	rs8007643	RNASE3-	14:2136	C>T	Nasal ala	3.36×10^{-08}	C/C	0.0976	0.0447	0.1458	0.0885	0.1155

			RNASE2 (Intergenic)	5801		Length						
Pickrell et al. (2016)	13	rs9310210	FOXP1(Intron)	3:71227306	T>A	Nose Size	8.2×10^{-26}	T/T	0.612	0.3991	0.3224	0.3419	0.2362
	14	rs13097965	EPHB3-MAGEF1 (Intergenic)	3:184339757	T>C	Nose Size	1.7×10^{-15}	C/C	0.7632	0.7075	0.5476	0.4503	0.7055
	15	rs56063440	CACNA2D3 (Intron)	3:54731374	G>C	Nose Size	5.5×10^{-11}	G/G	0.2534	0.1556	0.0109	0.2773	0.1646
	16	rs2929451	PPP1R3B-TNKS (Intergnic)	8:9085295	T>A	Nose Size	6.4×10^{-11}	T/T	0.3661	0.6542	0.9673	0.4573	0.729
	17	rs10809266	DMRT2-SMARCA2 (Intergnic)	9:1106093	A>G	Nose Size	1.1×10^{-09}	A/A	0.1657	0.4323	0.6875	0.3857	0.4121
	18	rs702489	GLIS1 (Intron)	1:54197688	A>G *	Nose Size	2.5×10^{-09}	G/G	0.8434	0.7723	0.8194	0.8787	0.7853
	19	rs35130793	BMP7 (Intron)	20:55792165	C>A	Nose Size	2.9×10^{-09}	C/C	0.1649	0.3919	0.0605	0.5338	0.4387
	20	rs2224309	GSC-DICER1 (Intergenic)	14:95333678	C>A	Nose Size	3.8×10^{-09}	C/C	0.2451	0.1297	0.1319	0.2157	0.1626
	21	rs10761129	ROR2 (Missense)	9:94486321	T>C	Nose Size	6.9×10^{-09}	T/T	0.2224	0.2277	0.0863	0.3121	0.4182
	22	rs34091987	SOX9 (Intergenic)	17:70025587	C>T	Nose Size	3.1×10^{-08}	C/C	0.3192	0.3934	0.0476	0.3439	0.137
	23	rs10779169	ALX1-RASSF9 (Intergenic)	12:85967804	G>A	Nose Size	3.6×10^{-08}	G/G	0.3495	0.6138	0.4792	0.5736	0.5828
	24	rs424737	ROBO1 (Intron)	3:78815906	G>A	Nose Size	4.6×10^{-08}	G/G	0.2731	0.2795	0.497	0.3121	0.2914
Lee et al. (2017)	25	rs9456748	PARK2 (Intron)	6:162590018	A>G *	Midfacial height	4.99×10^{-08}	G/G	0.9682	0.4769	0.7421	0.4632	0.6094

*AA: Ancestral Allele, DA: Derived Allele; D/N: Denisovan/Neanderthal; *Derived allele is shared with archaic humans*

3.2.1 Nasal SNPs with positive selection on either ancestral or derived variant

In order to establish plausible reasons for intra-human nasal morphological variation among African, American, European, East Asian and South Asian populations for the aforementioned 14 shortlisted SNPs, we collected 1000 Genomes Phase III SNP data (G. P. Consortium, 2015). By employing tests such as bifurcation diagrams and EHH plots (Gautier & Vitalis, 2012; Sabeti, et al., 2007), our results indicated 9 SNPs that displayed unique patterns of selection in one of the populations (Table 3.7). Among these analyzed SNPs, we observed 5/9 SNPs that stipulated contrasting patterns of selection for their ancestral and derived alleles particularly between African and rest of the populations for traits like nasal bridge breadth, nasal protrusion, nasal width, and nasal height (Table 3.7). All the 9 SNPs with results of positive selection on either ancestral or derived allele in one or more than one population are discussed in the following sections.

3.2.1.1 Differentially evolving Nasal SNPs in Africa

Out of 9 SNPs that showed signals of positive selection on either of its variants, we gained 5 SNPs that depicted results in Africa/non-Africa contrast. Each of the 5 SNPs is explained in the following sections.

SNP rs9456748-Mid-facial Height

Our results indicated an intriguing case of *PARK2* gene associated SNP rs9456748 (G>A) (Lee, et al., 2017) affecting mid-facial height was observed, in which derived allele has reached fixation in Africa with allele frequency of 0.968. In contrast, the ancestral allele has undergone positive selection in rest of the four non-African populations (Figure 3.24). This fixation of derived allele in Africans is in contrast with an opposite scenario of positive selection on ancestral allele in non-Africans. Given the climatic differences between Africa and that of much colder Europe, these results superimpose the contrasting nasal architecture (broad and aquiline), belonging to these two climatically challenged regions. It is also intriguing to note that the SNP rs9456748 is one of the initially collected 25 significant SNPs whose derived allele in modern *Homo sapiens* is also shared with Neanderthals and Denisovans, depicting a sequence level change prior to the evolution of archaic humans (Table 3.6).

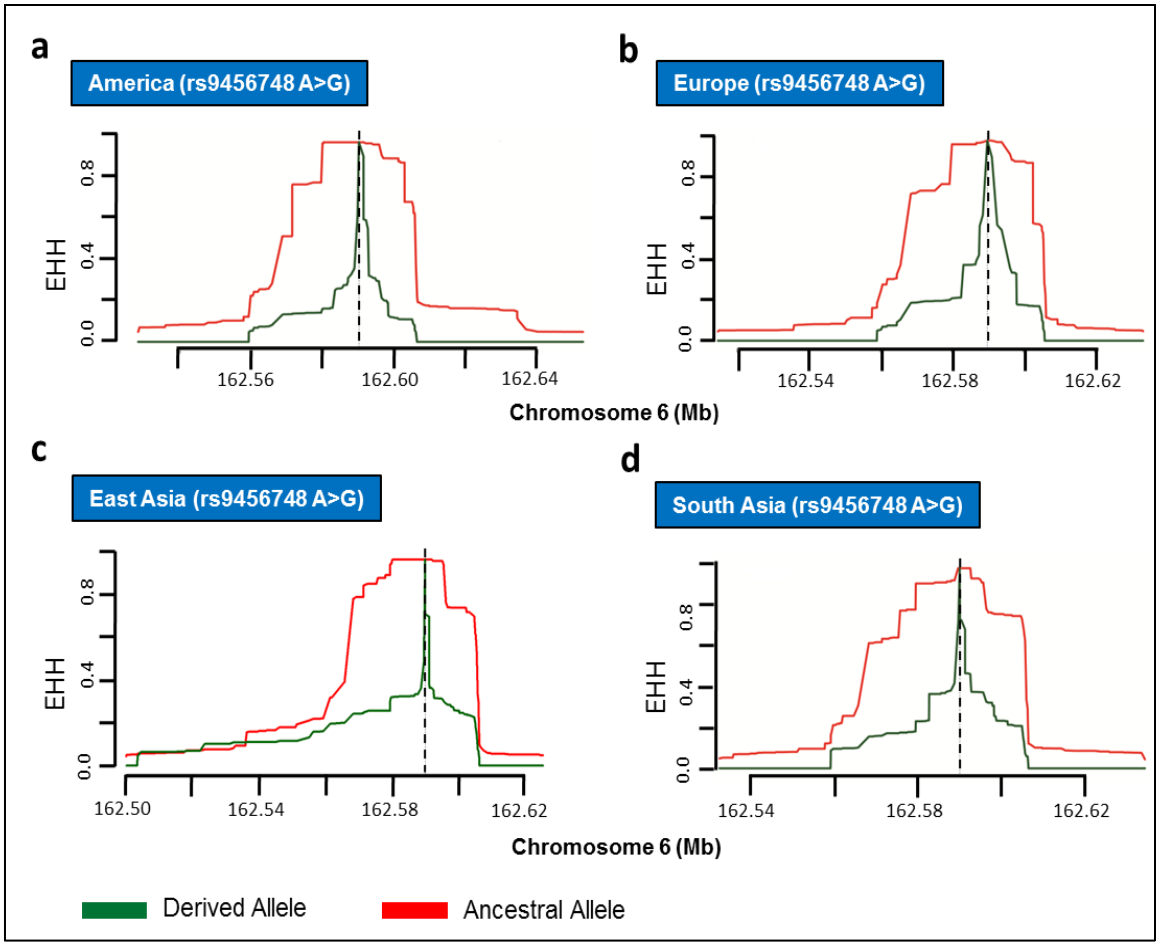


Figure 3.24: EHH plots show positive selection on ancestral allele of *PARK2* associated SNP rs9456748 for mid-facial height in non-African populations

*EHH=1 on Y-axis indicates all haplotypes carrying either ancestral or derived state of the allele are matching upto this point. X-axis contains coordinates for human chromosome 6. Ancestral allele is shown before the derived allele, separated by a “>” symbol. The four non-African populations in **(a)**,**(b)**, **(c)** and **(d)** depict ancestral allele A (in red) of SNP rs9456748 to be under positive selection. However, derived allele G has been fixed in Africa, rendering the frequency of ancestral allele to be <0.05 (not shown in Figure).*

SNP rs2045323- Nasal Tip Protrusion/Tip angle/Columella Inclination

Another interesting instance of non-African exclusive case of nasal variation was observed for *DCHS2* gene associated SNP rs2045323 (G>A) (Adhikari, et al., 2016) responsible for affecting nasal tip protrusion, tip angle and columella inclination. Nasal protrusion is termed as one probable adaptation for cold in present-day Europeans and also in Neanderthals where the entire anterior nasal cavity is prominent and much more likely associated with their mid-facial prognathism (Bastir & Rosas, 2016). Northern Asians, on the other hand, in extreme cold climate regions do not present elevated nasal protrusion when compared to other temperate East Asians (Evteev, et al., 2014). Whereas, our results indicate that derived variant of SNP rs2045323 is subjected to strong signal of positive selection in non-African populations. In Africa the derived allele frequency is still regressed i.e. <0.05, implicating the role of derived allele in non-African populations alone for the particular protruded nasal phenotype. Bifurcation diagrams reveal the longest haplotypes of 18.3kb and 10.5kb for American and European populations respectively. However, a relatively shorter unbranched haplotype of 5kb in South Asia was also observed (Figure 3.25). Lower instance of derived allele in Africa (DAF<0.05) and a branched haplotype in East Asia are in congruence with smaller nasal projections previously reported in these two regions (Zaidi, et al., 2017). Thus, we cannot fully reject the role of nasal projection in terms of adaptation in Africa (hot-humid) and non-Africa where temperate (South Asia, East Asia, America) to much colder (European) climates exist.

SNP rs1852985-Nasal Bridge Breadth

Sub features such as nasal root, nasal bridge and nasal wing collectively controlling nasal width are reported to be strongly correlated, however, their negative correlation has been reported with nasal protrusion (Adhikari, et al., 2016). Our results indicated that variants of SNPs controlling modern-day human nasal bridge breadth present significant views on adaptation commensurate with extreme climates in mind. Derived variant of SNP rs1852985 (C>T) (Adhikari, et al., 2016) controlling nasal bridge breadth, is also under positive selection in all regions except Africa where we see a much branched haplotype (Figure 3.26). Strongest results in terms of longest unbranched haplotypes were observed for Asian populations, 31.2kb in South Asia and 18.7kb in East Asia (Figure 3.26). If strict threshold for unbranched haplotype length is taken into account, we see longest unbranched haplotypes for nasal protrusion previously observed in American and European populations to be in contrast with their relatively branched haplotypes for nasal bridge (Figure 3.25 and Figure 3.26). Hence, the negative correlation between nasal protrusion and nasal width is corroborated (Zaidi, et al., 2017).

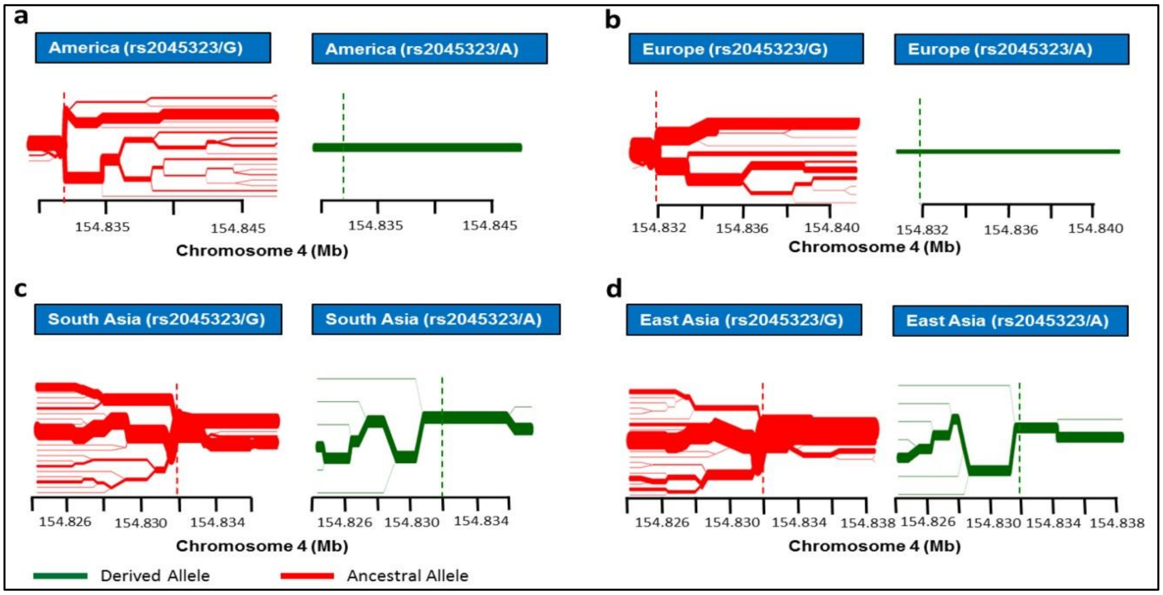


Figure 3.25: Bifurcation diagrams show positive selection on derived allele of *DCHS2* associated SNP rs2045323 for nasal protrusion/ tip angle/ columella inclination in non-African populations

*Bifurcation diagram shows little branching at the nodes interpreting for lesser recombination events and hence longer haplotypes with derived allele A compared to haplotypes with ancestral alelle G in all four non-African populations. All four populations **(a)** America **(b)** Europe **(c)** South Asia **(d)** East Asia depict derived allele A (in green) of SNP rs2045323 to be under positive selection, making unbranched or lesser branched haplotypes when seen with much branched ancestral allele haplotypes. Longest derived allele unbranched haplotypes (in green) of 18.3 and 10.5kb are observed for American **(a)** and European populations **(b)** respectively. Derived allele frequency (allele A) is however regressed in Africa i.e. < 0.05 (not shown in Figure).*

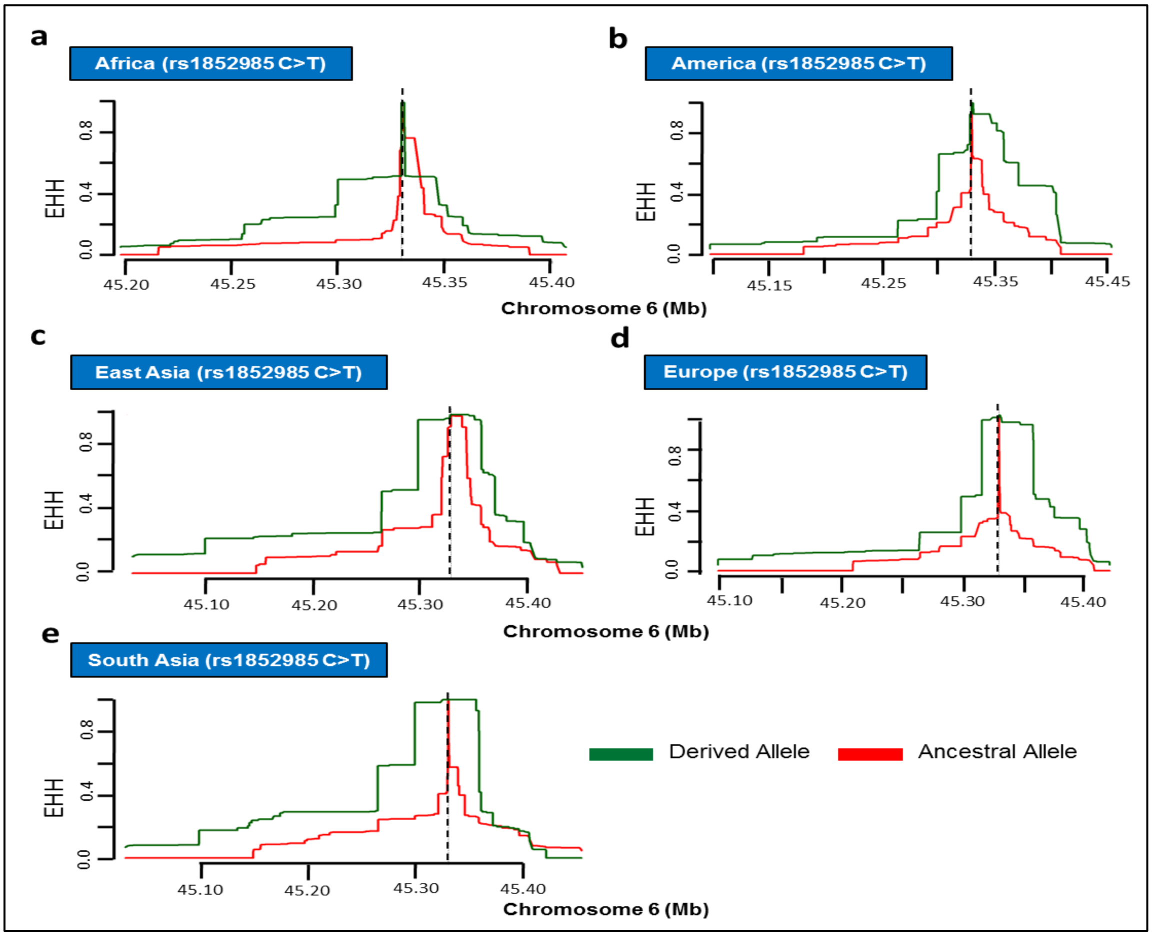


Figure 3.26: EHH plots show positive selection on derived allele of *RUNX2* associated SNP rs1852985 for nasal bridge breadth in non-African populations

*EHH=1 on Y-axis indicates all haplotypes carrying either ancestral or derived state of the allele are matching upto this point. X-axis contains coordinates for human chromosome 6. Ancestral allele is shown before the derived allele, separated by a ">" symbol **(a)** African population does not show positive selection on derived allele T (in green) of SNP rs1852985. The four non-African populations in **(b), (c), (d)** and **(e)** depict derived allele T (in green) to be under positive selection making longer unbroken haplotypes. EHH plots also show longest haplotypes of 31.2 and 18.7 for South Asian **(e)** and East Asian **(c)** populations respectively with derived allele T (in green).*

rs2424399-Nasal Width and rs11738462-Nasal Protrusion

Nasal width and nasal protrusion are two contrasting nasal traits that happen to distinguish nasal types of the two climatic extremes of African and European regions. The SNP rs2424399 associated with nasal width indicates that no selection regime is operative on either of the alleles in the African population, whereas, in rest of the populations, clear picture of positive selection can be tracked on the derived allele (Figure 3.27). The SNP rs11738462 is associated with nasal protrusion also shows a likewise contrasting trend of selection in non-African populations with respect to a much branched haplotype bifurcation diagram of both alleles in the African population (Figure 3.27).

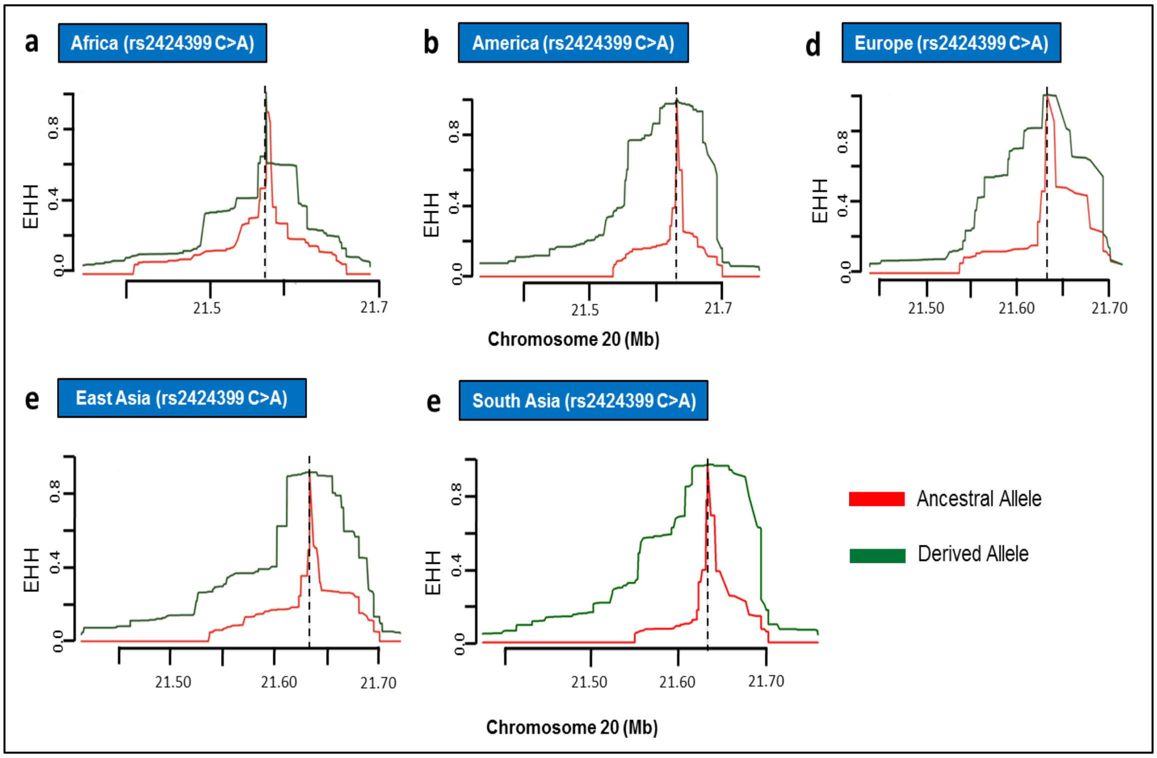


Figure 3.27: Positive selection on derived variant of SNP rs2424399 for nasal width in non-African populations

The above figure depicts EHH plots for SNP rs2424399. EHH=1 on Y-axis indicates all haplotypes carrying either ancestral or derived state of the allele are matching upto this point. X-axis contains coordinates for human chromosome 20. Ancestral allele is shown before the derived allele, separated by a ">" symbol. The figure shows positive selection on derived allele A (in green) in non-African populations, whereas no signal of positive selection can be tracked on either of the alleles in African population.

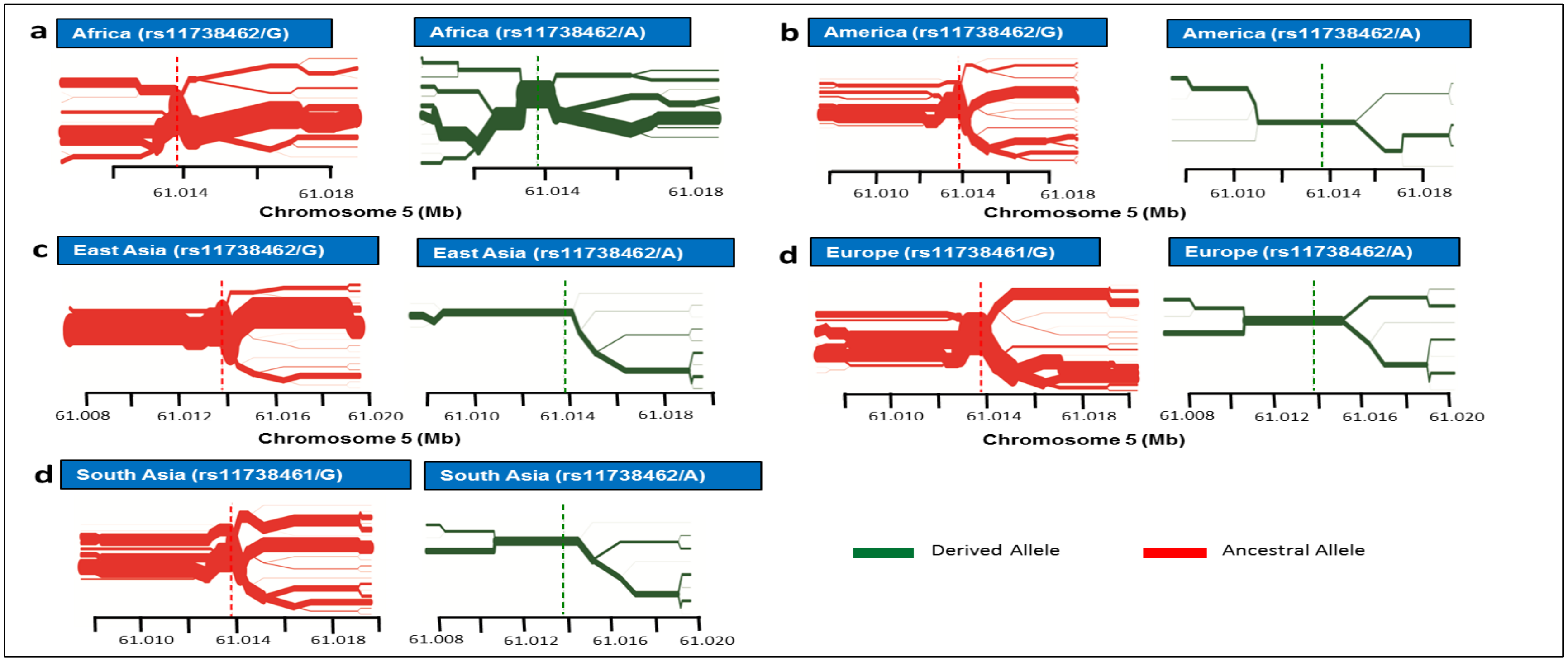


Figure 3.28: Bifurcation diagrams show positive selection on derived allele of *C5ORF64* associated SNP rs11738462 for nasal protrusion in non-African populations

The figure explains signals of positive selection on all four non- African populations for derived allele A (in green), as can be seen that longer unbranched haplotypes are formed with derived allele A in comparison with much branched haplotypes with ancestral allele G (in red). However, no evidence of positive selection is gauged on derived allele in African population because of a much branched bifurcation diagram that show greater number of recombination events and hence more branching at the nodes with derived as well as ancestral allele.

3.2.1.2 Differentially evolving nasal SNPs in South Asia

rs12644248 - Columella Inclination

In the present study, we also observed that allelic variants of *DCHS2* gene associated SNP rs12644248 controlling the trait of columella inclination (Adhikari, et al., 2016), are not evolving in a congruent pattern with regions of extreme climates. For derived allele we observe a 16.5kb haplotype in South Asia, whereas, relatively lesser branched haplotypes are observed for East Asian and American population with their counterpart ancestral allele (Figure 3.29). The derived allele frequency is regressed i.e. <0.05 in both European and African population.

3.2.1.3 Differentially evolving nasal SNPs in East Asia

rs755927 – Nasion Position and rs10761129 - Nose Size

SNPs rs7559271 and rs10761129 were also observed that showed prominent contrasting results between their derived and ancestral alleles in East Asian population compared to rest of the populations for traits like nasion position (Figure 3.30) and nose size respectively (Figure 3.31) (Adhikari, et al., 2016; Paternoster, et al., 2012; Pickrell, et al., 2016).

3.2.1.4 Differentially evolving nasal SNPs in Europe

SNP rs9310210 – Nose Size

Positive selection on derived allele of SNP rs9310210 was also observed in Europe for nose size (Figure 3.32) (Pickrell, et al., 2016). This result is significant in terms of contrasting nose size measurements observed for European populations compared to those of others (Zaidi, et al., 2017).

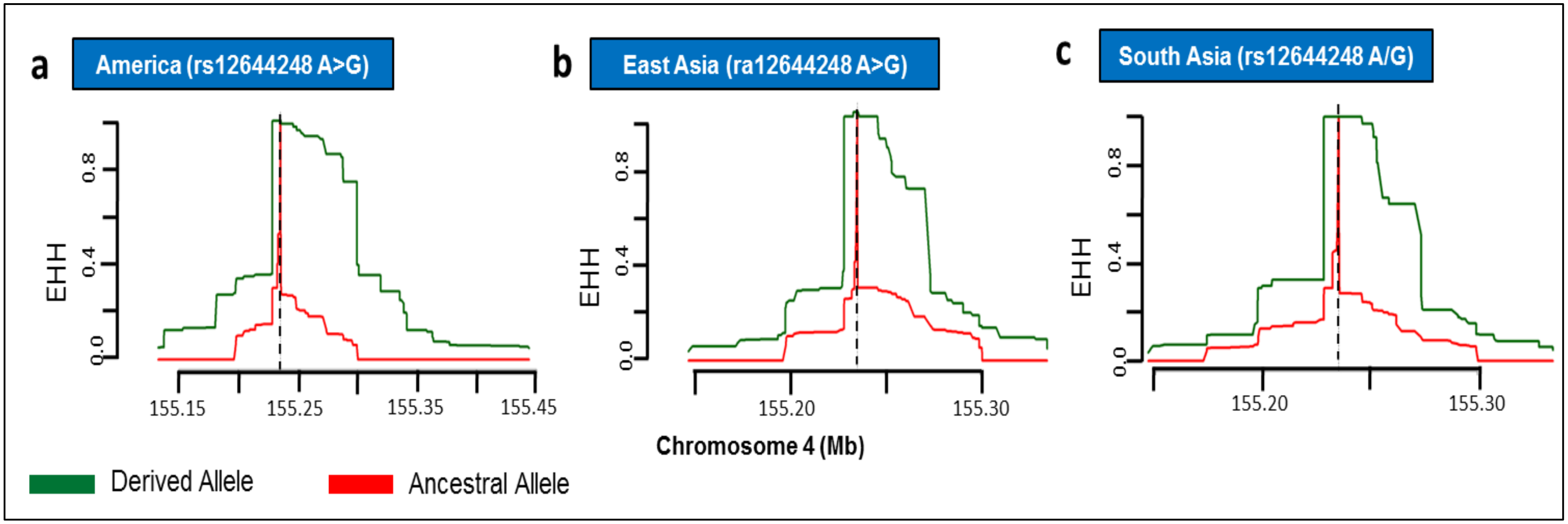


Figure 3.29: Positive selection on derived variant of *DCHS2* associated SNP rs12644248 for columella inclination in East Asian, South Asian and American populations

The above represents EHH plots for SNP rs12644248. EHH=1 on Y-axis indicates all haplotypes carrying either ancestral or derived state of the allele are matching upto this point. X-axis contains coordinates for human chromosome 4. Ancestral allele is shown before the derived allele, separated by a ">" symbol. The figure shows derived allele G (in green) of SNP rs12644248 associated with columella inclination to be under positive selection in Asian and American samples, with the longest haplotype of 16.5kb formed for South Asia. However, derived allele frequency is regressed in both climatically extreme regions of Europe and Africa i.e. <0.05.

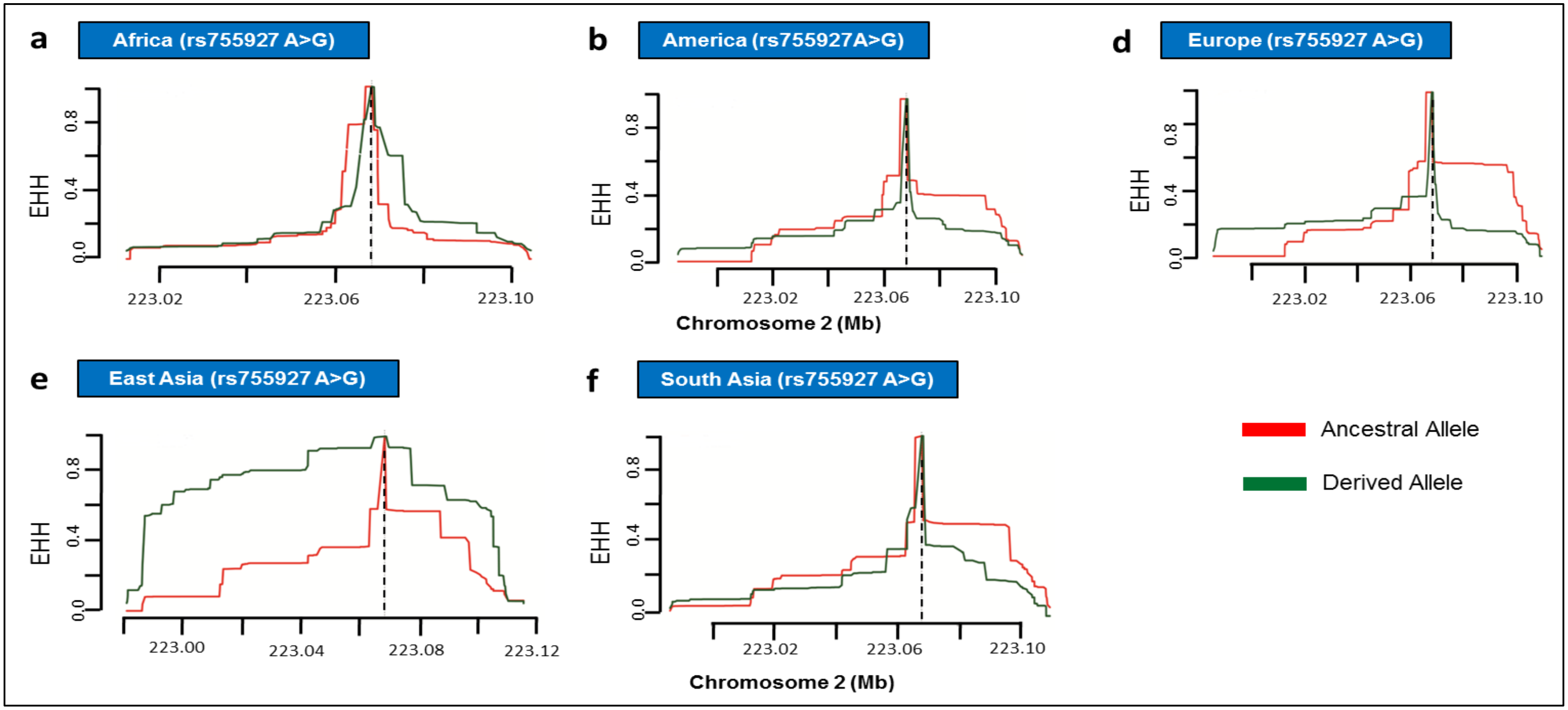


Figure 3.30: EHH plots show positive selection on derived variant of *PAX3* associated SNP rs7559272 for nasion position in East Asian population

The above depicts EHH plots for SNP rs755927. EHH=1 on Y-axis indicates all haplotypes carrying either ancestral or derived state of the allele are matching upto this point. X-axis contains coordinates for human chromosome 2. Ancestral allele is shown before the derived allele, separated by a ">" symbol. The figure shows positive selection on derived allele G (in green) in East Asia, whereas no other signal of positive selection can be tracked on either of the alleles in rest of the four populations.

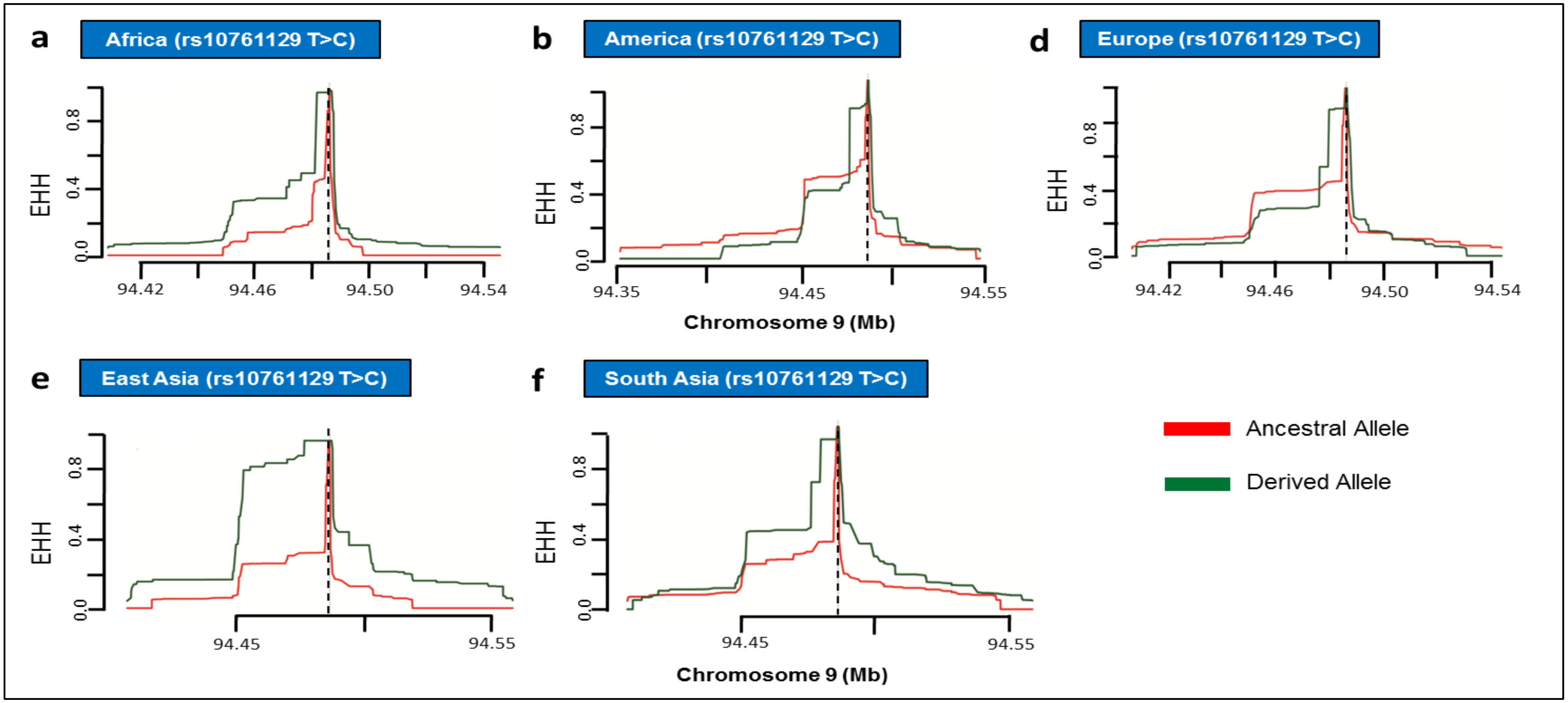


Figure 3.31: Positive selection on derived variant of *ROR2* associated SNP rs10761129 for nose size in East Asian population

The above figure depicts EHH plots for SNP rs755927. EHH=1 on Y-axis indicates all haplotypes carrying either ancestral or derived state of the allele are matching upto this point. X-axis contains coordinates for human chromosome 2. Ancestral allele is shown before the derived allele, separated by a ">" symbol. The figure shows positive selection on derived allele G (in green) in East Asia, whereas no other signal of positive selection can be tracked on either of the alleles in rest of the four populations.

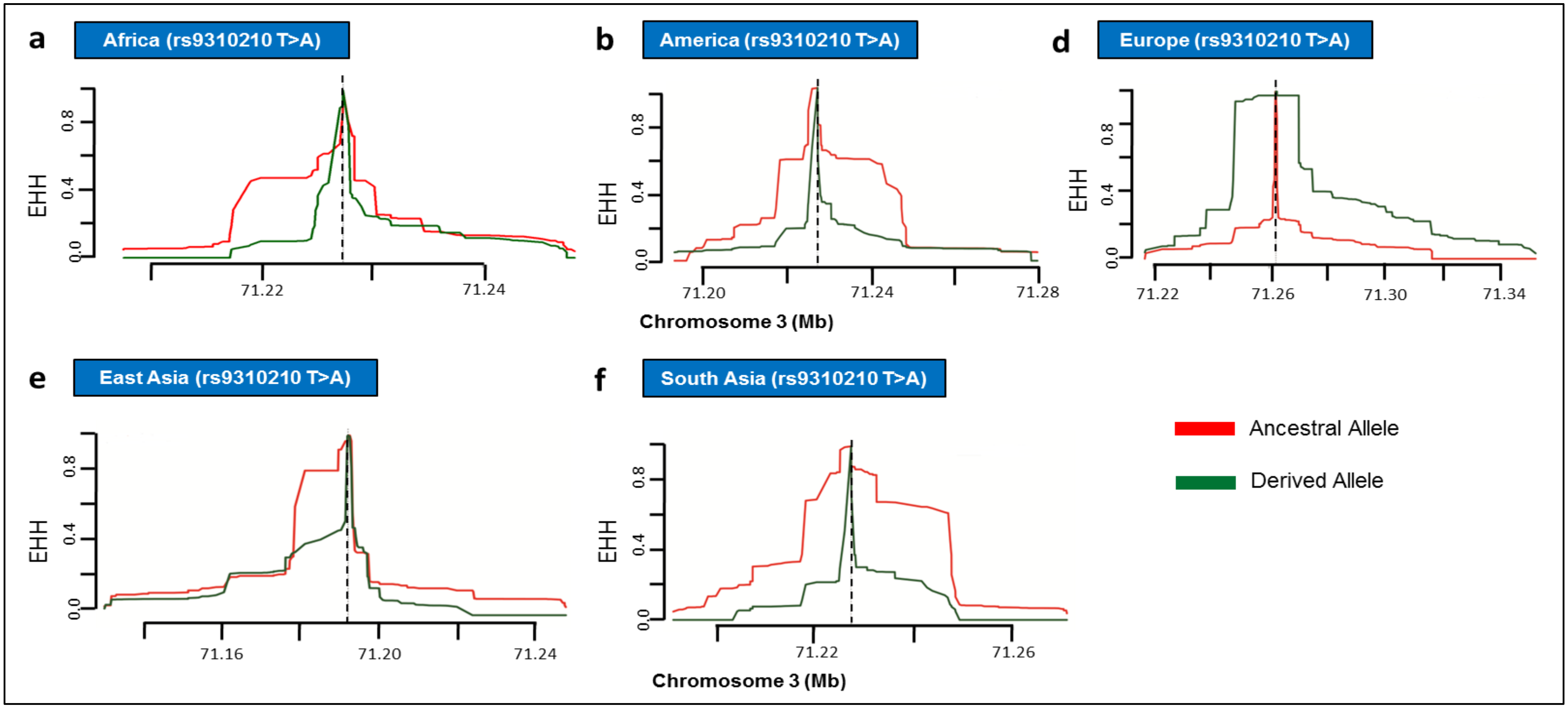


Figure 3.32: Positive selection on derived variant of *FOXP1* associated SNP rs9310210 for nose size in European population

The above figure illustrates EHH plots for SNP rs9310210. EHH=1 on Y-axis indicates all haplotypes carrying either ancestral or derived state of the allele are matching upto this point. X-axis contains coordinates for human chromosome 3. Ancestral allele is shown before the derived allele, separated by a ">" symbol. The figure shows positive selection on derived allele A (in green) in Europe, whereas no other signal of positive selection can be tracked on either of the alleles in rest of the four populations.

DISCUSSION

It has been made a point some 40 years ago that mutations are the common source of evolutionary change potentially affecting the regulation of genes (Britten & Davidson, 1969; King & Wilson, 1975; Zuckerkandl & Pauling, 1965). It is only in the past few years, empirical assessments have consolidated profound actuality of this deduction. It has now been grounded as a fundamental concept in evolutionary studies that mutations lying in cis-regulatory regions are more prone to contributing towards disease and phenotypic diversification among and between the species than those lying in trans-regulatory regions (Carroll, 2008; Stern & Orgogozo, 2008). These cis-regulatory regions are more conveniently bifurcated into promoters and enhancers, also known by the term cis-regulatory elements.

With the advent of sophisticated empirical advancements, there now exist several methodologies that can predict a regulatory active DNA across all mammalian genomes (Villar et al., 2015). Phenotypic differences across mammalian species are reported to be a consequence of innovations lying in these regulatory regions and not due to changes in the coding part of the genome (Wray, 2007). As promoters are associated with only the basal level of mRNA produced during transcription and are engaged with a highly conservative core set of transcription factors, there has been reported less evidence for their engagement in cis-regulatory divergence (Brown & Feder, 2005). They are however categorized as a critical cause of human disease if mutations occur (Savinkova et al., 2009). On the contrary, for the lack of universal transcriptional code in case of enhancers and their highly variable nature between species, species-specific sequence level variation owes much of its uniqueness to single nucleotide variants that happen to lie in these enhancers (Wray, 2007). A large onus is therefore set onto fast evolving enhancers that in combination with slowly evolving promoters comprise a distinguishing feature of all mammalian species separated by 180 Myrs (Villar, et al., 2015). Although, work has been done to interpret the evolutionarily compelling results on the 5'- flanking promoters of the neural genes, it is believed that enhancers have a much more impactful role to play in the decisive nature of the human trait advancement within the boundaries of of cis-regulatory innovations (Haygood, et al., 2007). In essence, sequence level changes in enhancers came out to be a big reason these phenotypic changes exist among species (Ludwig et al., 2005).

4.1 Human Genome Enhancements

Taking humans as the most unique of all primates and the most advanced in their physiological and anatomical characteristics, two important aspects of their genome enhancements came to notice over the past few years, i.e. human accelerated regions (HARs) and species-specific genome level reorganizations such as segmental duplications, deletions and insertions (Hubisz & Pollard, 2014; Sassa, 2013).

4.1.1 INDELS

INDELS (INsertions and DELetions) engaging regulatory elements in many studies were reported to play a significant role in species specific gain/loss of some traits. One study reported that more than five hundred regions, highly conserved between mammals and chimpanzee, are absent in the human genome, suggesting a substantive deletion of putative CREs from the human genome (McLean et al., 2011). These deletions can be important in also deleting sites for repressors and bringing together sequences for novel activator sites, hence increasing the overall cis-regulatory function instead of decreasing it (Shirangi, Dufour, Williams, & Carroll, 2009). Analysis of two of the mentioned putative regulatory regions indicated that their absence caused an observed loss of penile spines, sensory vibrissae and also a part of the brain was not expanded (McLean, et al., 2011). Insertions being another potential source of driving cis-regulatory divergence among species also create novel sites for either repressors or activators and in turn either disrupting or amplifying the regulatory function respectively (Williams et al., 2008).

4.1.2 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 frgaments 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). It is to this speculation that *in vivo* analysis of such human accelerated non-coding regions attributed to the presence of cis-regulatory transcriptional enhancers controlling the expression of many developmental genes (Prabhakar, et al., 2008). In meta analysis of five reported studies (Bird et al., 2007; Bush & Lahn, 2008; Pollard et al., 2006; Prabhakar, Noonan, Pääbo, & Rubin, 2006; Zuckerkandl & Pauling, 1965) that predicted HARs in the human genome, 2649 non-coding HARs were categorized upon exclusion of the protein coding regions, of which majority lied in the intronic and intergenic regions (Capra, Erwin, McKinsey, Rubenstein, & Pollard, 2013). Interestingly, studies also claimed to have categorized a large part of these accelerated regions of the human genome to be neuronal enhancers (Doan et al., 2016). One such evolutionary study also endorsed acceleration in enhancer sequences compared to coding and non-coding/non-enhancer genomic blocks in vertebrates during land adaptation (Yousaf, et al., 2015)..

It is also important to note that enhancer sequences found to be either conserved or recently evolved are both correlated with many phenotypic effects. To align regulatory evolution with the most developed and fascinating organ in human anatomy i.e. brain, several studies have been conducted. Humans by keeping the most advanced pre-frontal cortex (PFC) and a highly developed telencephalon, have a large part of their genome in sharing with that of chimpanzee (Levchenko, et al., 2017). Although, genes expressed in brain are reported to have evolved slower in mammals than in other tissues (Duret & Mouchiroud, 2000; Kuma, Iwabe, & Miyata, 1995; L. Zhang & Li, 2004), this rate of evolution has increased in the primate clade of which humans make the most cognitively intelligent offshoot (H.-Y. Wang et al., 2006). The question lies whether the evolution of brain expressed genes in humans has rapidly increased during the course of time and in turn contributed towards the enhanced cognition? The answer was cleared in a study that reported the rate of evolution in brain expressed genes to be lower or at least equal to that of chimpanzee and old world monkeys (H.-Y. Wang, et al., 2006). The only answer then lies with a plausible deduction that regulation of such genes might have contributed to enhanced brain faculties of the human. This was corroborated in a differential expression pattern of the brain expressed genes observed for human when comparison was made with a closely related chimpanzee (Cáceres, et al., 2003). This is in bigger context debated to have been a result of natural selection in the cis-regulatory regions that made *Homo sapiens* possessors of a beneficial leverage not only over non-human primates but also on their contemporaries of the genus Homo (Levchenko, et al., 2017). A recent study has therefore bolstered this view where human specific changes in a neuro-developmental enhancer of frizzled 8 (*FZD8)* gene produced immense differences in the size of the brain (Franchini & Pollard, 2015).

HARs and the putative enhancers that might comprise these regions have been studied in depth to relate the importance of these regions with human brain (Franchini & Pollard, 2015). Studies have also indicated a strong correlation between gene forms of mentally challenged diseased or to have a profound role in to that of accelerated regions. Two of the studies found three human specific variants in the introns of autism susceptibility candidate 2 (*AUTS2*), a HAR associated gene, in its rummaged structural variants in many neurological disorders (Pollard, et al., 2006; Prabhakar, et al., 2006). Other examples include cut-like homeobox 1 (*CUX1*) gene known to have a role in autism as a transcriptional repressor. This gene is associated with a HAR containing enhancer reported by Prabhakar and colleagues, that gains an additional TFBS due to a G>A substitution (Prabhakar, et al., 2006). This enhancer substitution along with the overexpression of the gene triggers the onset of autism and other intellectual disabilities (Doan, et al., 2016). An integrative study to link SNPs associated with schizophrenia with the accelerated regions has also been reported, highlighting the fact that the genes or SNPs involved in *Homo sapien*-specific ailment of schizophrenia also have a corresponding accelerated region that may as well be a regulatory region (Britten & Davidson, 1969; Levchenko, et al., 2017). Another fascinating study reported neuronal PAS domain containing protein 3 (*NPAS3*) to have contained a very high number of HARs in its introns (Kamm, Pisciottano, Kliger, & Franchini, 2013). *NPAS3*'s role in brain development and neuro-signalling has been well established in the previously reported work (Brunskill et al., 2005). Other genes to be regulated by the accelerated regulatory regions and also reported to have a role in brain development are polypyrimidine tract binding protein 2 (*PTBP2*) and glypican 4 (*GPC4*) (Bird, et al., 2007).

4.2 Enhancer Diversification

Recently evolved enhancers are also known to have been arisen due to changes in the ancestral sequence, instead of being driven via lineage-specific expansion of the repeat sequences (Villar, et al., 2015). It is nonetheless very important to see which of the human specific variants in the human accelerated regions belonged to the lineage of *Homo sapiens* alone. In order to sum up the number, Burbano and colleagues (Burbano, et al., 2012) analyzed the percentage of these modern human specific variations in four of the pioneer studies that set up the dynamics of accelerated evolution of the modern human genome in comparison to archaic humans (Bird, et al., 2007; Bush & Lahn, 2008; Pollard, et al., 2006; Prabhakar, et al., 2006). They estimated a decent percentage of 8.3% of the substitutions in the HARs to be solely modern human specific. This analysis puts to light the origination of so many traits unique to modern humans that may be speculated to have originated due to the sequence level acceleration and provided an advantageous edge to *Homo sapiens* in term of both cognitive status and other physiological adaptabilities.

4.2.1 Selection on Enhancers

Implying population genetics on all such species specific variants shed light on various evolutionary perspective in which standing genetic variants could be highlighted. Selective sweep is a phenomenon where a beneficial allele and the adjacent closely present chromosomal segment increases in frequency in a population due to positive selection. Considering single nucleotide substitutions, their role in disease, acceleration and loss/gain of traits, it is important to discuss and distinguish their role under the theory of population genetics. According to this theory, a beneficial allele swiftly sweeps to fixation in a population after its arrival. However, the standing genetic variants, those which have been segregating in a population for quite some time have also been observed to be under selection regime and contributing to several phenotypic adaptations (Przeworski, 2002). As an example, there has been observed loss of pelvis structures in freshwater stickleback populations as a consequence of mutations that inactivated the enhancer activity of the *PITX1* gene (Chan et al., 2010). The inactivated enhancers resulted in the loss of the pelvic girdle. These mutations were observed to be under positive selection in pelvic reduced stickleback populations and reappeared in the population on recurrent deletions (Wittkopp & Kalay, 2012) . As in our study, we were able to identify the standing genetic variants among *Homo sapiens*-specific TFBSs within the brain-exclusive enhancers on which positive selection across the present-day human population is operating in a region wise manner (Zehra & Abbasi, 2018).

4.2.2 Selection on Human Brain Enhancers

In this study, it was established that accelerated regions in the experimentally confirmed enhancers exist in the human lineage and this acceleration can be tracked to a lot of favorable outputs. Of which fine-tuning of regulatory elements in combination with their strict control on developmental genes by keeping human lineage in perspective is of great importance. We conducted a sequential study over 271 empirically confirmed brain-specific VISTA enhancers and prioritized sequence level acceleration over them (Visel, et al., 2007). By employing variable methodologies that resulted in robust confirmation of enhancers which "truly" depicted signals of positive selection, we set onto other plausible deductions about this accelerated set. Thus, out of our root dataset of empirically confirmed, brain specific enhancers, we isolated those enhancers that showed significant signatures of acceleration upon comparison with closest non-human primates. This set of enhancers was then evaluated for putative target gene association. This step was significant as it was equally intriguing to know what set of genes these accelerated enhancers were controlling and to what extent they played a role in human brain development.

For 15 BE-HARs, 31 genes were observed to be syntenically conserved among long distance species and many of them played a key developmental role in human brain development. Since, enhancers lack a universal code of identification due to ambiguous nature of residing either very close are very far from the gene's promoter site, this step provided useful insights into future analysis of the genes which could depict either a positive or a negative correlation with their associated enhancers.

We then set out to explore the transcriptional space of the accelerated enhancers with respect to transcription factor binding sites. From TRANSFAC and extensive literature survey, we gathered a set of 142 TFs that have a role to play in the anatomy of the human brain and their respective binding sites (Matys, et al., 2003). By running all 15 BE-HAEs through all of the collected binding sites, we obtained several sites that were human specific. The comparison made was with chimpanzee only. However, it came to our notice that by including other distantly spaced primate species, the sites reemerged in one or more of the other lineages and could however be perceived as a chimpanzee specific loss of the site instead of being categorized as a human specific gain. By including gorilla, orangutan and macaque orthologous sequences, it largely curtailed our set of initially gained human specific TFBSs. Amongst 15 BE-HAEs, we noticed 9 such accelerated enhancers hat possessed human unique transcription factor binding sites.

Setting a dynamic after finding the putative target genes and human specific binding sites for the accelerated enhancers, we then headed for the *Homo sapien*-unique sites. As for the evolutionary perspective, we believe that it is nonetheless mandatory to see sites which could have played a role in adapting the present-day brain structure from that of closely associated species of genus Homo. We incorporated orthologous sequences from Neanderthals and Denisovans for all our predicted human specific binding sites and evaluated three such sites that were not only unique in comparison with non-human primates but also with archaic humans. The sites belonged to TFs SOX2, RUNX1/3 and FOS/JUND, all of which play a crucial role in human brain development.

As many prior studies debated that acceleration in the human regions is a result of genetic drift that states all these lineage specific changes are randomly evolving in the human genome under neutral evolution. To configure this, we already estimated signals of positive selection in our empirically confirmed brain specific enhancers. To further estimate the significance of these selection signals and to combine what we had gathered so far in terms of associated target genes, and *Homo sapien*-specific binding sites, we deemed it important to see if selection signals are operating within the present-day human population in a region wise manner or was this anatomical advance just confined to a distinction between modern and archaic humans. We then took to see positive selection results within the present-day human population on those sites which were previously categorized as modern-day human specific. To our surprise, what we initially suspected about the selection regime to must have been acting in a region wise manner came true. We found Africa to be truly a contrasting factor in terms of distinguishing results from rest of the human populations. Africa, being a source of human origination and spread to all parts of the world corresponded well in our results. Among the three previously collected binding sites of factors SOX2, RUNX1/3 and FOS/JUND, we were able to identify the sites of SOX2 and RUNX1/3 to be under positive selection in Africa. These findings are commensurate with data that describes greater percentage of variants within non-coding regulatory genome than coding part of the genome. This work also brought forth patterns of accelerated divergence across present-day human population for SNPs residing in *Homo sapien*-specific TFBSs, ones which are not shared among the orthologous enhancer archaic and non-human primate sequences (Zehra & Abbasi, 2018).

4.3 Selection on Facial Genetic Components

Our second round of work included evolutionary evaluation of the facial features, among which we specifically analyzed the nasal associated SNPs from various genome association studies (Adhikari, et al., 2016; Lee, et al., 2017; Paternoster, et al., 2012; Pickrell, et al., 2016; Shaffer, et al., 2016). Diversification in the human genome soon after the split from chimpanzees has rendered many human traits significant (Carroll, 2003). The genomic changes, either in the coding or non-coding parts of the genome, (F. Liu, et al., 2012) have manifested in a variety of morphological and anatomical traits that gave *Homo sapiens* a profound leverage over other hominoids. For a well-rounded perspective on how sequence acceleration can be playing a decisive role in adapting brain structure in humans, to also engaging other human unique anatomical features, we extended our work to an intra-population analysis of SNPs controlling nasal morphology. Our choice of this trait was placed in a much connected impact of brain expansion over facial features of which nasal morphology stays the most dynamic and the most variable of all among the present-day human population. Comparative neuroanatomy in modern humans has revealed the keystone of such manifestations i.e. an increased brain size that implicated special areas of the brain to develop sophisticated sensory, motor and cognitive abilities (Carroll, 2003). To gauge physical ramifications of an increased brain size during the course of primate evolution in directly impacting the related anatomical sub-structures such as face has long intrigued the scientific community. Previous comparative studies indicate the existence of a mechanistic interplay of basicranium with brain size (Jeffery, 2003). Because of the physical attachment between the cranial base and face, basicranium has played a decisive, a likely non-random adaptive role in reorienting and reducing the facial size and shape from middle to late Pleistocene humans (Bastir & Rosas, 2016). Although much work has been done on brain evolution and the way it has revolutionized the cognitive status of *Homo sapiens*, gaps exist in formulating the genetic underpinnings of a vast degree of human facial variation and its potential correlations with adaptability. In previous findings, evolution of the human facial form because of its high-level variation between intra and inter-hominins has been attributed to facial functionality (Lieberman, 2008). External climatic factors of humidity, temperature and dental load imposed mechanical demands and largely determined the masticatory and respiratory facial biomechanics of the hominins. This in turn greatly affected their mid-facial morphology (Bastir & Rosas, 2016). Facial prognathism in Neanderthals is one such trait shared with their predecessors to have resulted due to paramasticatory stress, i.e. the use of anterior dentition(Trinkaus, 1987). The lack of this function in modern humans resulted in relatively shorter faces, a condition that is said to be evolutionary derived in modern humans (Lieberman, 2008; Trinkaus, 2003).

The facial differentiation between *Homo sapiens* and sister taxon *Homo neanderthalensis* is predominantly large with plethora of climatic, anatomical and evolutionary constraints playing their role. Nasal morphological features of our sister taxon Neanderthals are debated as their wider nasal apertures do not seem to be in accordance with extremely cold glacial habitats they inhabited, as observed otherwise in narrower nasal form of current-day circumpolar European population. However, prior studies have indicated that adaptation to extremely cold and relatively lesser cold environments can vary with winter moisture playing a crucial variable(Evteev, et al., 2014). Such was the case when broad nasal apertures of Neanderthals were in congruence with wider nasal apertures observed in very cold and dry inland populations of Northern Asia (Evteev, et al., 2014).

The role of nose as human body's natural conditioning system makes it more sensitive towards climatic changes. For inhaling hot humid air in African regions to facing drier-cold air in Europe, nasal morphology has changed immensely over time and among different parts of the world. In order to cater to all such delicate divisions, we referred to all studies till date that explored SNPs associated with one or more nasal features. We gathered 25 SNPs exceeding the conventional threshold and selected amongst them those SNPs where there existed a vast difference between the frequency of ancestral and derived alleles. We shortlisted 14 SNPs for future study and via various population genetic tests, we were able to identify 9 Such SNPs belonging to different nasal traits to be under the influence of natural selection, again in a region wise manner. At this point, we hoped to have the same results as we expected for modern human unique binding site variants within he accelerated enhancers. To our deduction, we got the same distinction between alleles of 5 SNPs that showed variant patterns of selection on their ancestral and derived alleles in a contrasting manner between Africa and no-Africa. Features such as nasal width, nasal protrusion, and mid-facial height, apparently differing between the African and non-African parts of the world also responded on the same lines in our results of intra-population assessments of positive selection.

It is however very intriguing for us to observe majority of the nasal associated SNPs to be having a modern human specific variant and also to be lying in non-coding regions. In a previous study, cis-regulatory evolution was categorized as one strong evolutionary driving factor of craniofacial features (Attanasio et al., 2013). Prescott and co-workers narrate that species-biased expression of genes controlling craniofacial traits is majorly governed by species-biased distal cis-elements called enhancers (Prescott et al., 2015). These species-biased enhancers differ in their epigenomic make-up across orthologous counterpart enhancers and are attributable to biases in transcription factor and p300 binding along with enhanced chromatin accessibility. This altered dose of the genes regulated by species-biased enhancers implies not only facial divergence across species but failing to reach a certain expression threshold is attributed to disease related malformations. Moreover, cis-regulatory elements have been reported to play a keen role in the expression of genes influencing nasal morphology during development and embryogenesis (Pregizer & Mortlock, 2009). By keeping this in perspective, it is safe to speculate that nasal morphological variation can largely be associated with an evolved regulatory landscape governing facial status of modern humans. Our results also conclude that a significant portion of non-coding human genome is driven via accelerated divergence of alleles patterning nasal morphology across different climatic parameters.

4.4 Conclusion

A bonafide status to categorize enhancers as majorly occupying components of non-coding functional part of the genome has been established. These enhancers do not work in isolation but a whole different perspective is attained when other CREs and genetic components that also include trans regulatory factors are taken into account. As many human-specific changes also incorporated in these enhancers, the functional consequence of these changes or mutations remains a point of scrutiny for years to come. In our study we picked enhancers for the aforementioned significance and also for their role in advancing human brain faculties. Merging their evolutionary status with trans environment and pinpointing sites that were mutated or modified and also conferred selection among the present-day human population was the theme of our study. This was extended and tested among the SNPs that controlled human nasal morphology that is prone to various climatic shifts faced by different regionalization of the globe as well. By incorporating archaic human data, we tried to infer selection signatures in accelerated enhancers as well as on SNPs from GWAS associations controlling brain development nasal morphology. This in its own space provides an insight as to how human diverged from apes and archaic humans but also onto how environmental or adaptive dynamics are helping the present-day *Homo sapiens* to continue evolving.

4.5 Future Prospects

Biases in speciation events are driven by genomic sequences. Genomes being sequenced at a stupendous pace give us ample data to find signatures of selection onto these sequences. Although much work has been done to draw conclusions on human trait advancement, a lot of gaps still exist. Apart from brain that is the source of human evolution, efforts can be directed on other human anatomical features as well. We do believe that enhancers expressing in other important tissues such as heart, liver or limbs may be prone to variable pattern of special selection in human when compared with other primates. Even traits like bipedalism and various kinds of facial forms in humans can further be investigated within the bounds of cis-regulatory evolution by not just including the enhancers and promoters but also silencers, insulators and LCRs and therefore its role in fine-tuning the gene regulatory circuits can further be established. All this can be done on a genome level scale and among the same species or different species.

PUBLICATIONS

- 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.
- Nasal morphological variation across human population (Manuscript submitted)

REFERENCES

Abbasi, A. A., Paparidis, Z., Malik, S., Bangs, F., Schmidt, A., Koch, S., . . . Grzeschik, K.-H. (2010). Human intronic enhancers control distinct sub-domains of Gli3 expression during mouse CNS and limb development. *BMC developmental biology, 10*(1), 44.

Abbasi, A. A., Paparidis, Z., Malik, S., Goode, D. K., Callaway, H., Elgar, G., & Grzeschik, K.-H. (2007). Human GLI3 intragenic conserved non-coding sequences are tissue-specific enhancers. *PloS one, 2*(4), e366.

Adhikari, K., Fuentes-Guajardo, M., Quinto-Sánchez, M., Mendoza-Revilla, J., Chacón-Duque, J. C., Acuña-Alonzo, V., . . . Pérez, G. M. (2016). A genome-wide association scan implicates DCHS2, RUNX2, GLI3, PAX1 and EDAR in human facial variation. *Nature Communications, 7*, 11616.

Alexander, R. P., Fang, G., Rozowsky, J., Snyder, M., & Gerstein, M. B. (2010). Annotating non-coding regions of the genome. *Nature Reviews Genetics, 11*(8), 559.

Atkinson, E. G., Rogers, J., Mahaney, M. C., Cox, L. A., & Cheverud, J. M. (2015). Cortical folding of the primate brain: an interdisciplinary examination of the genetic architecture, modularity, and evolvability of a significant neurological trait in pedigreed baboons (genus Papio). *Genetics*, genetics. 114.173443.

Attanasio, C., Nord, A. S., Zhu, Y., Blow, M. J., Li, Z., Liberton, D. K., . . . Hosseini, R. (2013). Fine tuning of craniofacial morphology by distant-acting enhancers. *Science, 342*(6157), 1241006.

Banerji, J., Rusconi, S., & Schaffner, W. (1981). Expression of a β-globin gene is enhanced by remote SV40 DNA sequences. *Cell, 27*(2), 299-308.

Barber, B. A., Liyanage, V. R., Zachariah, R. M., Olson, C. O., Bailey, M. A., & Rastegar, M. (2013). Dynamic expression of MEIS1 homeoprotein in E14. 5 forebrain and differentiated forebrain-derived neural stem cells. *Annals of Anatomy-Anatomischer Anzeiger, 195*(5), 431-440.

Barreiro, L. B., Laval, G., Quach, H., Patin, E., & Quintana-Murci, L. (2008). Natural selection has driven population differentiation in modern humans. *Nature Genetics, 40*(3), 340-345.

Bastir, M., & Rosas, A. (2016). Cranial base topology and basic trends in the facial evolution of Homo. *Journal of human evolution, 91*, 26-35.

Baxevanis, A. D. (2004). An overview of gene identification: approaches, strategies, and considerations. *Current protocols in bioinformatics*, 4.1. 1-4.1. 9.

Beccari, L., Conte, I., Cisneros, E., & Bovolenta, P. (2012). Sox2-mediated differential activation of Six3. 2 contributes to forebrain patterning. *Development, 139*(1), 151-164.

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.

Blair, J. E., & Hedges, S. B. (2005). Molecular phylogeny and divergence times of deuterostome animals. *Molecular biology and evolution, 22*(11), 2275-2284.

Blake, J. A., Eppig, J. T., Kadin, J. A., Richardson, J. E., Smith, C. L., Bult, C. J., & Group, M. G. D. (2016). Mouse Genome Database (MGD)-2017: community knowledge resource for the laboratory mouse. *Nucleic Acids Research, 45*(D1), D723-D729.

Bomba, L., Nicolazzi, E. L., Milanesi, M., Negrini, R., Mancini, G., Biscarini, F., . . . Ajmone-Marsan, P. (2015). Relative extended haplotype homozygosity signals across breeds reveal dairy and beef specific signatures of selection. *Genetics Selection Evolution, 47*(1), 25.

Boyd, J. L., Skove, Stephanie L., Rouanet, Jeremy P., Pilaz, L.-J., Bepler, T., Gordân, R., . . . Silver, Debra L. (2015). Human-Chimpanzee Differences in a FZD8 Enhancer Alter Cell-Cycle Dynamics in the Developing Neocortex. *Current Biology, 25*(6), 772-779. doi: 10.1016/j.cub.2015.01.041

Brand, A. H., Breeden, L., Abraham, J., Sternglanz, R., & Nasmyth, K. (1985). Characterization of a "silencer" in yeast: a DNA sequence with properties opposite to those of a transcriptional enhancer. *Cell, 41*(1), 41-48.

Britten, R. J., & Davidson, E. H. (1969). Gene regulation for higher cells: a theory. *Science, 165*(3891), 349-357.

Britten, R. J., & Davidson, E. H. (1971). Repetitive and non-repetitive DNA sequences and a speculation on the origins of evolutionary novelty. *The Quarterly review of biology, 46*(2), 111-138.

Brown, R. P., & Feder, M. E. (2005). Reverse transcriptional profiling: non-correspondence of transcript level variation and proximal promoter polymorphism. *BMC genomics, 6*(1), 110.

Browning, S. R., & Browning, B. L. (2011). Haplotype phasing: existing methods and new developments. *Nature Reviews Genetics, 12*(10), 703.

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.

Burbano, H. A., Green, R. E., Maricic, T., Lalueza-Fox, C., de La Rasilla, M., Rosas, A., . . . Pääbo, S. (2012). Analysis of human accelerated DNA regions using archaic hominin genomes. *PloS one, 7*(3), e32877.

Bush, E. C., & Lahn, B. T. (2008). A genome-wide screen for noncoding elements important in primate evolution. *BMC evolutionary biology, 8*(1), 17.

Cáceres, M., Lachuer, J., Zapala, M. A., Redmond, J. C., Kudo, L., Geschwind, D. H., . . . Barlow, C. (2003). Elevated gene expression levels distinguish human from non-human primate brains. *Proceedings of the National Academy of Sciences, 100*(22), 13030-13035.

Cadzow, M., Boocock, J., Nguyen, H. T., Wilcox, P., Merriman, T. R., & Black, M. A. (2014). A bioinformatics workflow for detecting signatures of selection in genomic data. *Frontiers in genetics, 5*, 293.

Capra, J. A., Erwin, G. D., McKinsey, G., Rubenstein, J. L., & Pollard, K. S. (2013). Many human accelerated regions are developmental enhancers. *Phil. Trans. R. Soc. B, 368*(1632), 20130025.

Carroll, S. B. (2003). Genetics and the making of Homo sapiens. *Nature, 422*(6934), 849.

Carroll, S. B. (2008). Evo-devo and an expanding evolutionary synthesis: a genetic theory of morphological evolution. *Cell, 134*(1), 25-36.

Chan, Y. F., Marks, M. E., Jones, F. C., Villarreal, G., Shapiro, M. D., Brady, S. D., . . . Schmutz, J. (2010). Adaptive evolution of pelvic reduction in sticklebacks by recurrent deletion of a Pitx1 enhancer. *Science, 327*(5963), 302-305.

Choukrallah, M.-A., Song, S., Rolink, A. G., Burger, L., & Matthias, P. (2015). Enhancer repertoires are reshaped independently of early priming and heterochromatin dynamics during B cell differentiation. *Nature Communications, 6*, 8324. doi: 10.1038/ncomms9324

Chuang, J. H., & Li, H. (2004). Functional bias and spatial organization of genes in mutational hot and cold regions in the human genome. *PLoS Biol, 2*(2), e29.

Cole, C., & Josselyn, S. (2008). Transcription regulation of memory: CREB, CaMKIV, Fos/Jun, CBP, and SRF. *Learning and memory: a comprehensive reference. Elsevier, Oxford*, 547-566.

Consortium, G. P. (2010). A map of human genome variation from population-scale sequencing. *Nature, 467*(7319), 1061.

Consortium, G. P. (2015). A global reference for human genetic variation. *Nature, 526*(7571), 68-74.

Consortium, I. H. G. S. (2001). Initial sequencing and analysis of the human genome. *Nature, 409*(6822), 860.

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., . . . Sherry, S. T. (2011). The variant call format and VCFtools. *Bioinformatics, 27*(15), 2156-2158.

Dillon, N., & Grosveld, F. (1994). Chromatin domains as potential units of eukaryotic gene function. *Current opinion in genetics & development, 4*(2), 260-264.

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.

Dorighi, K. M., Swigut, T., Henriques, T., Bhanu, N. V., Scruggs, B. S., Nady, N., . . . Wysocka, J. (2017). Mll3 and Mll4 Facilitate Enhancer RNA Synthesis and Transcription from Promoters Independently of H3K4 Monomethylation. *Molecular Cell*.

Dunbar, R. I., & Shultz, S. (2007). Understanding primate brain evolution. *Philosophical Transactions of the Royal Society of London B: Biological Sciences, 362*(1480), 649-658.

Duret, L., & Mouchiroud, D. (2000). Determinants of substitution rates in mammalian genes: expression pattern affects selection intensity but not mutation rate. *Molecular biology and evolution, 17*(1), 68-070.

Enard, W. (2015). Human evolution: enhancing the brain. *Current Biology, 25*(10), R421-R423.

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.

Evans, D. J., & Francis-West, P. H. (2005). Craniofacial development: making faces. *Journal of anatomy, 207*(5), 435-436.

Evteev, A., Cardini, A. L., Morozova, I., & O'Higgins, P. (2014). Extreme climate, rather than population history, explains mid-facial morphology of northern asians. *American journal of physical anthropology, 153*(3), 449-462.

Fagertun, J., Wolffhechel, K., Pers, T. H., Nielsen, H. B., Gudbjartsson, D., Stefansson, H., . . . Jarmer, H. (2015). Predicting facial characteristics from complex polygenic variations. *Forensic Science International: Genetics, 19*, 263-268.

Fang, L., Ahn, J. K., Wodziak, D., & Sibley, E. (2012). The human lactase persistence-associated SNP− 13910* T enables in vivo functional persistence of lactase promoter–reporter transgene expression. *Human genetics, 131*(7), 1153-1159.

Fay, J. C., & Wu, C.-I. (2000). Hitchhiking under positive Darwinian selection. *Genetics, 155*(3), 1405-1413.

Finger, J. H., Smith, C. M., Hayamizu, T. F., McCright, I. J., Xu, J., Law, M., . . . Blodgett, O. (2016). The mouse gene expression database (GXD): 2017 update. *Nucleic Acids Research, 45*(D1), D730-D736.

Forrester, W. C., Epner, E., Driscoll, M. C., Enver, T., Brice, M., Papayannopoulou, T., & Groudine, M. (1990). A deletion of the human beta-globin locus activation region causes a major alteration in chromatin structure and replication across the entire beta-globin locus. *Genes & development, 4*(10), 1637-1649.

Franchini, L. F., & Pollard, K. S. (2015). Can a few non-coding mutations make a human brain? *BioEssays, 37*(10), 1054-1061.

Gautier, M., & Vitalis, R. (2012). rehh: an R package to detect footprints of selection in genome-wide SNP data from haplotype structure. *Bioinformatics, 28*(8), 1176-1177.

Geschwind, D. H., & Rakic, P. (2013). Cortical evolution: judge the brain by its cover. *Neuron, 80*(3), 633-647.

Goldman, N., & Yang, Z. (1994). A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Molecular biology and evolution, 11*(5), 725-736.

Grosveld, F., van Assendelft, G. B., Greaves, D. R., & Kollias, G. (1987). Position-independent, high-level expression of the human β-globin gene in transgenic mice. *Cell, 51*(6), 975-985.

Gu, J., & Gu, X. (2003). Induced gene expression in human brain after the split from chimpanzee. *Trends in Genetics, 19*(2), 63-65.

Hadley, T. J., & Peiper, S. C. (1997). From malaria to chemokine receptor: the emerging physiologic role of the Duffy blood group antigen. *Blood, 89*(9), 3077-3091.

Hamblin, M. T., & Di Rienzo, A. (2000). Detection of the signature of natural selection in humans: evidence from the Duffy blood group locus. *The American Journal of Human Genetics, 66*(5), 1669-1679.

Hare, E. E., Peterson, B. K., Iyer, V. N., Meier, R., & Eisen, M. B. (2008). Sepsid even-skipped enhancers are functionally conserved in Drosophila despite lack of sequence conservation. *PLoS genetics, 4*(6), e1000106.

Haygood, R., Fedrigo, O., Hanson, B., Yokoyama, K.-D., & Wray, G. A. (2007). Promoter regions of many neural- and nutrition-related genes have experienced positive selection during human evolution. *Nature Genetics, 39*(9), 1140-1144. doi: 10.1038/ng2104

Heckmann, J., Uwimpuhwe, H., Ballo, R., Kaur, M., Bajic, V. B., & Prince, S. (2010). A functional SNP in the regulatory region of the decay-accelerating factor gene associates with extraocular muscle pareses in myasthenia gravis. *Genes and immunity, 11*(1), 1-10.

Heintzman, N. D., Stuart, R. K., Hon, G., Fu, Y., Ching, C. W., Hawkins, R. D., . . . Ching, K. A. (2007). Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nature Genetics, 39*(3), 311.

Hindorff, L. A., Sethupathy, P., Junkins, H. A., Ramos, E. M., Mehta, J. P., Collins, F. S., & Manolio, T. A. (2009). Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proceedings of the National Academy of Sciences, 106*(23), 9362-9367.

Hubisz, M. J., & Pollard, K. S. (2014). Exploring the genesis and functions of Human Accelerated Regions sheds light on their role in human evolution. *Current opinion in genetics & development, 29*, 15-21.

Hussin, J., Nadeau, P., Lefebvre, J.-F., & Labuda, D. (2010). Haplotype allelic classes for detecting ongoing positive selection. *BMC bioinformatics, 11*(1), 1.

Hutton, S. R., & Pevny, L. H. (2011). SOX2 expression levels distinguish between neural progenitor populations of the developing dorsal telencephalon. *Developmental biology, 352*(1), 40-47.

Inoue, K.-i., Shiga, T., & Ito, Y. (2008). Runx transcription factors in neuronal development. *Neural development, 3*(1), 1.

Jacob, F., & Monod, J. (1978). Genetic regulatory mechanisms in the synthesis of proteins *Selected Papers in Molecular Biology by Jacques Monod* (pp. 433-471): Elsevier.

Jeffery, N. (2003). Brain expansion and comparative prenatal ontogeny of the non-hominoid primate cranial base. *Journal of human evolution, 45*(4), 263-284.

Jin, Z., Liu, L., Bian, W., Chen, Y., Xu, G., Cheng, L., & Jing, N. (2009). Different transcription factors regulate nestin gene expression during P19 cell neural differentiation and central nervous system development. *Journal of Biological Chemistry, 284*(12), 8160-8173.

Kamm, G. B., Pisciottano, 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.

Karki, R., Pandya, D., Elston, R. C., & Ferlini, C. (2015). Defining "mutation" and "polymorphism" in the era of personal genomics. *BMC medical genomics, 8*(1), 37.

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.

Katoh, K., Misawa, K., Kuma, K. i., & Miyata, T. (2002). MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Research, 30*(14), 3059-3066.

Kent, W. J. (2002). BLAT—the BLAST-like alignment tool. *Genome research, 12*(4), 656-664.

King, M.-C., & Wilson, A. C. (1975). Evolution at two levels in humans and chimpanzees. *Science, 188*(4184), 107-116.

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.

Kuma, K.-i., Iwabe, N., & Miyata, T. (1995). Functional constraints against variations on molecules from the tissue level: slowly evolving brain-specific genes demonstrated by protein kinase and immunoglobulin supergene families. *Molecular biology and evolution, 12*(1), 123-130.

Kvon, E. Z. (2015). Using transgenic reporter assays to functionally characterize enhancers in animals. *Genomics, 106*(3), 185-192.

Lee, M. K., Shaffer, J. R., Leslie, E. J., Orlova, E., Carlson, J. C., Feingold, E., . . . Weinberg, S. M. (2017). Genome-wide association study of facial morphology reveals novel associations with FREM1 and PARK2. *PloS one, 12*(4), e0176566.

Lemon, B., & Tjian, R. (2000). Orchestrated response: a symphony of transcription factors for gene control. *Genes & development, 14*(20), 2551-2569.

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.

Li, Q., Peterson, K. R., Fang, X., & Stamatoyannopoulos, G. (2002). Locus control regions. *Blood, 100*(9), 3077-3086.

Lieberman, D. E. (2008). Speculations about the selective basis for modern human craniofacial form. *Evolutionary Anthropology: Issues, News, and Reviews, 17*(1), 55-68.

Lieberman, D. E. (2015). Human locomotion and heat loss: an evolutionary perspective. *Comprehensive Physiology*.

Liu, F., Van Der Lijn, F., Schurmann, C., Zhu, G., Chakravarty, M. M., Hysi, P. G., . . . Ikram, M. A. (2012). A genome-wide association study identifies five loci influencing facial morphology in Europeans. *PLoS genetics, 8*(9), e1002932.

Liu, J., & Francke, U. (2006). Identification of cis-regulatory elements for MECP2 expression. *Human molecular genetics, 15*(11), 1769-1782.

Ludwig, M. Z., Palsson, A., Alekseeva, E., Bergman, C. M., Nathan, J., & Kreitman, M. (2005). Functional evolution of a cis-regulatory module. *PLoS biology, 3*(4), e93.

Luo, Z.-X., Yuan, C.-X., Meng, Q.-J., & Ji, Q. (2011). A Jurassic eutherian mammal and divergence of marsupials and placentals. *Nature, 476*(7361), 442.

Malt, A. L., Cesario, J. M., Tang, Z., Brown, S., & Jeong, J. (2014). Identification of a face enhancer reveals direct regulation of LIM homeobox 8 (Lhx8) by wingless-int (WNT)/β-catenin signaling. *Journal of Biological Chemistry, 289*(44), 30289-30301.

Maricic, T., Günther, V., Georgiev, O., Gehre, S., Ćurlin, 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.

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

Mathé, C., Sagot, M. F., Schiex, T., & Rouzé, P. (2002). Current methods of gene prediction, their strengths and weaknesses. *Nucleic Acids Research, 30*(19), 4103-4117. doi: 10.1093/nar/gkf543

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

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.

Meyer, M., Kircher, M., Gansauge, M.-T., Li, H., Racimo, F., Mallick, S., . . . De Filippo, C. (2012). A high-coverage genome sequence from an archaic Denisovan individual. *Science, 338*(6104), 222-226.

Minoux, M., & Rijli, F. M. (2010). Molecular mechanisms of cranial neural crest cell migration and patterning in craniofacial development. *Development, 137*(16), 2605-2621.

Monod, J., & Jacob, F. (1961). *General conclusions: teleonomic mechanisms in cellular metabolism, growth, and differentiation.* Paper presented at the Cold Spring Harbor symposia on quantitative biology.

Muse, S. V., & Gaut, B. S. (1994). A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. *Molecular biology and evolution, 11*(5), 715-724.

Narlikar, L., & Ovcharenko, I. (2009). Identifying regulatory elements in eukaryotic genomes. *Briefings in Functional Genomics and Proteomics, 8*(4), 215-230.

Neigenfind, J., Gyetvai, G., Basekow, R., Diehl, S., Achenbach, U., Gebhardt, C., . . . Kersten, B. (2008). Haplotype inference from unphased SNP data in heterozygous polyploids based on SAT. *BMC genomics, 9*(1), 356.

Neubauer, S., Hublin, J.-J., & Gunz, P. (2018). The evolution of modern human brain shape. *Science advances, 4*(1), eaao5961.

Nielsen, R. (2005). Molecular signatures of natural selection. *Annu. Rev. Genet., 39*, 197-218.

Nielsen, R., Akey, J. M., Jakobsson, M., Pritchard, J. K., Tishkoff, S., & Willerslev, E. (2017). Tracing the peopling of the world through genomics. *Nature, 541*(7637), 302-310.

Noback, M. L., Harvati, K., & Spoor, F. (2011). Climate-related variation of the human nasal cavity. *American journal of physical anthropology, 145*(4), 599-614.

Nolis, I. K., McKay, D. J., Mantouvalou, E., Lomvardas, S., Merika, M., & Thanos, D. (2009). Transcription factors mediate long-range enhancer–promoter interactions. *Proceedings of the National Academy of Sciences, 106*(48), 20222-20227.

Nord, A. S., Pattabiraman, K., Visel, A., & Rubenstein, J. L. (2015). Genomic perspectives of transcriptional regulation in forebrain development. *Neuron, 85*(1), 27-47.

O'Connell, J., Gurdasani, D., Delaneau, O., Pirastu, N., Ulivi, S., Cocca, M., . . . Rudan, I. (2014). A general approach for haplotype phasing across the full spectrum of relatedness. *PLoS genetics, 10*(4), e1004234.

Ogbourne, S., & Antalis, T. M. (1998). Transcriptional control and the role of silencers in transcriptional regulation in eukaryotes. *Biochemical Journal, 331*(Pt 1), 1.

Olds, L. C., & Sibley, E. (2003). Lactase persistence DNA variant enhances lactase promoter activity in vitro: functional role as a cis regulatory element. *Human molecular genetics, 12*(18), 2333-2340.

Park, S. G., Hannenhalli, S., & Choi, S. S. (2014). Conservation in first introns is positively associated with the number of exons within genes and the presence of regulatory epigenetic signals. *BMC genomics, 15*(1), 1.

Parveen, N., Masood, A., Iftikhar, N., Minhas, B. F., Minhas, R., Nawaz, U., & Abbasi, A. A. (2013). Comparative genomics using teleost fish helps to systematically identify target gene bodies of functionally defined human enhancers. *BMC genomics, 14*(1), 122.

Paternoster, L., Zhurov, A. I., Toma, A. M., Kemp, J. P., Pourcain, B. S., Timpson, N. J., . . . Smith, G. D. (2012). Genome-wide association study of three-dimensional facial morphology identifies a variant in PAX3 associated with nasion position. *The American Journal of Human Genetics, 90*(3), 478-485.

Paulson, J. R., & Laemmli, U. (1977). The structure of histone-depleted metaphase chromosomes. *Cell, 12*(3), 817-828.

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

Pennisi, E. (2012). ENCODE project writes eulogy for junk DNA: American Association for the Advancement of Science.

Pickrell, J. K., Berisa, T., Liu, J. Z., Ségurel, L., Tung, J. Y., & Hinds, D. A. (2016). Detection and interpretation of shared genetic influences on 42 human traits. *Nature Genetics, 48*(7), 709-717.

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.

Pond, S. L. K., Frost, S. D. W., & Muse, S. V. (2004). HyPhy: hypothesis testing using phylogenies. *Bioinformatics, 21*(5), 676-679. doi: 10.1093/bioinformatics/bti079

Prabhakar, S., Noonan, J. P., Pääbo, S., & Rubin, E. M. (2006). Accelerated evolution of conserved noncoding sequences in humans. *Science, 314*(5800), 786-786.

Prabhakar, S., Visel, A., Akiyama, J. A., Shoukry, M., Lewis, K. D., Holt, A., . . . Afzal, V. (2008). Human-specific gain of function in a developmental enhancer. *Science, 321*(5894), 1346-1350.

Pregizer, S., & Mortlock, D. P. (2009). Control of BMP gene expression by long-range regulatory elements. *Cytokine & growth factor reviews, 20*(5), 509-515.

Prescott, S. L., Srinivasan, R., Marchetto, M. C., Grishina, I., Narvaiza, I., Selleri, L., . . . Wysocka, J. (2015). Enhancer divergence and cis-regulatory evolution in the human and chimp neural crest. *Cell, 163*(1), 68-83.

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.

Przeworski, M. (2002). The signature of positive selection at randomly chosen loci. *Genetics, 160*(3), 1179-1189.

Qanbari, S., Pimentel, E., Tetens, J., Thaller, G., Lichtner, P., Sharifi, A., & Simianer, H. (2010). A genome-wide scan for signatures of recent selection in Holstein cattle. *Animal genetics, 41*(4), 377-389.

Rada-Iglesias, A., Bajpai, R., Prescott, S., Brugmann, S. A., Swigut, T., & Wysocka, J. (2012). Epigenomic annotation of enhancers predicts transcriptional regulators of human neural crest. *Cell stem cell, 11*(5), 633-648.

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.

Rollini, P., Namciu, S. J., Marsden, M. D., & Fournier, R. (1999). Identification and characterization of nuclear matrix-attachment regions in the human serpin gene cluster at 14q32. 1. *Nucleic Acids Research, 27*(19), 3779-3791.

Roseman, C. C. (2004). Detecting interregionally diversifying natural selection on modern human cranial form by using matched molecular and morphometric data. *Proceedings of the National Academy of Sciences of the United States of America, 101*(35), 12824-12829.

Rowe, T. B., Macrini, T. E., & Luo, Z.-X. (2011). Fossil evidence on origin of the mammalian brain. *Science, 332*(6032), 955-957.

Sabeti, P. C., Reich, D. E., Higgins, J. M., Levine, H. Z., Richter, D. J., Schaffner, S. F., . . . McDonald, G. J. (2002). Detecting recent positive selection in the human genome from haplotype structure. *Nature, 419*(6909), 832.

Sabeti, P. C., Varilly, P., Fry, B., Lohmueller, J., Hostetter, E., Cotsapas, C., . . . Gaudet, R. (2007). Genome-wide detection and characterization of positive selection in human populations. *Nature, 449*(7164), 913-918.

Sakai, T., Matsui, M., Mikami, A., Malkova, L., Hamada, Y., Tomonaga, M., . . . Makishima, H. (2013). Developmental patterns of chimpanzee cerebral tissues provide important clues for understanding the remarkable enlargement of the human brain. *Proc. R. Soc. B, 280*(1753), 20122398.

Salem, R. M., Wessel, J., & Schork, N. J. (2005). A comprehensive literature review of haplotyping software and methods for use with unrelated individuals. *Human Genomics, 2*(1), 39.

Sassa, T. (2013). The role of human-specific gene duplications during brain development and evolution. *Journal of neurogenetics, 27*(3), 86-96.

Savinkova, L., Ponomarenko, M., Ponomarenko, P., Drachkova, I., Lysova, M., Arshinova, T., & Kolchanov, N. (2009). TATA box polymorphisms in human gene promoters and associated hereditary pathologies. *Biochemistry (Moscow), 74*(2), 117-129.

Scheet, P., & Stephens, M. (2006). A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. *The American Journal of Human Genetics, 78*(4), 629-644.

Shaffer, J. R., Orlova, E., Lee, M. K., Leslie, E. J., Raffensperger, Z. D., Heike, C. L., . . . Nidey, N. L. (2016). Genome-wide association study reveals multiple loci influencing normal human facial morphology. *PLoS genetics, 12*(8), e1006149.

Sherry, S. T., Ward, M.-H., Kholodov, M., Baker, J., Phan, L., Smigielski, E. M., & Sirotkin, K. (2001). dbSNP: the NCBI database of genetic variation. *Nucleic Acids Research, 29*(1), 308-311.

Shirangi, T. R., Dufour, H. D., Williams, T. M., & Carroll, S. B. (2009). Rapid evolution of sex pheromone-producing enzyme expression in Drosophila. *PLoS biology, 7*(8), e1000168.

Shishikura, M., Nakamura, F., Yamashita, N., Uetani, N., Iwakura, Y., & Goshima, Y. (2016). Expression of receptor protein tyrosine phosphatase δ, PTPδ, in mouse central nervous system. *Brain research, 1642*, 244-254.

Smale, S. T., & Kadonaga, J. T. (2003). The RNA polymerase II core promoter. *Annual review of biochemistry, 72*(1), 449-479.

Spitz, F., & Furlong, E. E. M. (2012). Transcription factors: from enhancer binding to developmental control. *Nature Reviews Genetics, 13*(9), 613-626. doi: 10.1038/nrg3207

Stern, D. L., & Orgogozo, V. (2008). The loci of evolution: how predictable is genetic evolution? *Evolution, 62*(9), 2155-2177.

Stifani, S., & Ma, Q. (2009). 'Runxs and regulations' of sensory and motor neuron subtype differentiation: Implications for hematopoietic development. *Blood Cells, Molecules, and Diseases, 43*(1), 20-26.

Tekola-Ayele, F., Adeyemo, A., Chen, G., Hailu, E., Aseffa, A., Davey, G., . . . Rotimi, C. N. (2015). Novel genomic signals of recent selection in an Ethiopian population. *European Journal of Human Genetics, 23*(8), 1085-1092.

Thomsen, R., Nielsen, P. S., & Jensen, T. H. (2005). Dramatically improved RNA in situ hybridization signals using LNA-modified probes. *Rna, 11*(11), 1745-1748.

Tirosh, I., Barkai, N., & Verstrepen, K. J. (2009). Promoter architecture and the evolvability of gene expression. *Journal of biology, 8*(11), 95.

Tournamille, C., Colin, Y., Cartron, J. P., & Le Van Kim, C. (1995). Disruption of a GATA motif in the Duffy gene promoter abolishes erythroid gene expression in Duffy–negative individuals. *Nature Genetics, 10*(2), 224.

Trinkaus, E. (1987). The Neandertal face: evolutionary and functional perspectives on a recent hominid face. *Journal of human evolution, 16*(5), 429-443.

Trinkaus, E. (2003). Neandertal faces were not long; modern human faces are short. *Proceedings of the National Academy of Sciences, 100*(14), 8142-8145.

Uetani, N., Kato, K., Ogura, H., Mizuno, K., Kawano, K., Mikoshiba, K., . . . Iwakura, Y. (2000). Impaired learning with enhanced hippocampal long-term potentiation in PTPδ-deficient mice. *The EMBO journal, 19*(12), 2775-2785.

Uhlén, M., Fagerberg, L., Hallström, B. M., Lindskog, C., Oksvold, P., Mardinoglu, A., . . . Asplund, A. (2015). Tissue-based map of the human proteome. *Science, 347*(6220), 1260419.

Umarov, R. K., & Solovyev, V. V. (2017). Recognition of prokaryotic and eukaryotic promoters using convolutional deep learning neural networks. *PloS one, 12*(2), e0171410.

Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., . . . Holt, R. A. (2001). The sequence of the human genome. *Science, 291*(5507), 1304-1351.

Villar, D., Berthelot, C., Aldridge, S., Rayner, T. F., Lukk, M., Pignatelli, M., . . . Jasinska, A. J. (2015). Enhancer evolution across 20 mammalian species. *Cell, 160*(3), 554-566.

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

Voight, B. F., Kudaravalli, S., Wen, X., & Pritchard, J. K. (2006). A map of recent positive selection in the human genome. *PLoS Biol, 4*(3), e72.

Wang, H.-Y., Chien, H.-C., Osada, N., Hashimoto, K., Sugano, S., Gojobori, T., . . . Shen, C.-K. J. (2006). Rate of evolution in brain-expressed genes in humans and other primates. *PLoS biology, 5*(2), e13.

Wang, Z., Chen, Y., & Li, Y. (2004). A brief review of computational gene prediction methods. *Genomics, proteomics & bioinformatics, 2*(4), 216-221.

Ward, L. D., & Kellis, M. (2012). Evidence of abundant purifying selection in humans for recently acquired regulatory functions. *Science, 337*(6102), 1675-1678.

Weir, B. S., & Cockerham, C. C. (1984). Estimating F-statistics for the analysis of population structure. *Evolution, 38*(6), 1358-1370.

West, A. G., Gaszner, M., & Felsenfeld, G. (2002). Insulators: many functions, many mechanisms. *Genes & development, 16*(3), 271-288.

Wilkie, A. O., & Morriss-Kay, G. M. (2001). Genetics of craniofacial development and malformation. *Nature Reviews Genetics, 2*(6), 458.

Williams, T. M., Selegue, J. E., Werner, T., Gompel, N., Kopp, A., & Carroll, S. B. (2008). The regulation and evolution of a genetic switch controlling sexually dimorphic traits in Drosophila. *Cell, 134*(4), 610-623.

Wittkopp, P. J., & Kalay, G. (2012). Cis-regulatory elements: molecular mechanisms and evolutionary processes underlying divergence. *Nature Reviews Genetics, 13*(1), 59.

Wong, W. S., & Nielsen, R. (2004). Detecting selection in noncoding regions of nucleotide sequences. *Genetics, 167*(2), 949-958.

Wray, G. A. (2007). The evolutionary significance of cis-regulatory mutations. *Nature Reviews Genetics, 8*(3), 206.

Yang, Z. (1997). PAML: a program package for phylogenetic analysis by maximum likelihood. *Bioinformatics, 13*(5), 555-556.

Yang, Z., Nielsen, R., Goldman, N., & Pedersen, A.-M. K. (2000). Codon-substitution models for heterogeneous selection pressure at amino acid sites. *Genetics, 155*(1), 431-449.

Yean, D., & Gralla, J. D. (1999). Transcription reinitiation rate: a potential role for TATA box stabilization of the TFIID: TFIIA: DNA complex. *Nucleic Acids Research, 27*(3), 831-818.

Yokley, T. R. (2009). Ecogeographic variation in human nasal passages. *American journal of physical anthropology, 138*(1), 11-22.

Young, T. K., & Mäkinen, T. M. (2010). The health of Arctic populations: Does cold matter? *American Journal of Human Biology, 22*(1), 129-133.

Yousaf, A., Raza, M. S., & Abbasi, A. A. (2015). The evolution of bony vertebrate enhancers at odds with their coding sequence landscape. *Genome biology and evolution, 7*(8), 2333-2343.

Zaidi, A. A., Mattern, B. C., Claes, P., McEcoy, B., Hughes, C., & Shriver, M. D. (2017). Investigating the case of human nose shape and climate adaptation. *PLoS genetics, 13*(3), e1006616.

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.

Zhang, C., Li, J., Tian, L., Lu, D., Yuan, K., Yuan, Y., & Xu, S. (2015). Differential natural selection of human zinc transporter genes between African and Non-African populations. *Scientific reports, 5*, 9658.

Zhang, J. (2005). Evaluation of an Improved Branch-Site Likelihood Method for Detecting Positive Selection at the Molecular Level. *Molecular biology and evolution, 22*(12), 2472-2479. doi: 10.1093/molbev/msi237

Zhang, J., Nielsen, R., & Yang, Z. (2005). Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. *Molecular biology and evolution, 22*(12), 2472-2479.

Zhang, L., & Li, W.-H. (2004). Mammalian housekeeping genes evolve more slowly than tissue-specific genes. *Molecular biology and evolution, 21*(2), 236-239.

Zuckerkandl, E., & Pauling, L. (1965). Evolutionary divergence and convergence in proteins *Evolving genes and proteins* (pp. 97-166): Elsevier.

APPENDICES

7.1 Appendix -I: Determination of fast evolving enhancers with global proxy regions

Table A1. Results with *FHL1*-Intron 5 (GRCh37: ChrX: 135290801-135292029) global proxy region for 271 brain exclusive enhancers

	VISTA	VISTA Coordinates	Brain Domain	Alignment	Rate	P-Value	Q-Value
SN	ID	(GRCh37/hg19)		Length	Analysis		
1	hs526	chr4:1613479-1614106	Forebrain	620	H>C	0.00003	0.001444721
2	hs1019	chr7:20838843-20840395	Forebrain	471	H>C	0.00002	0.001444721
3	hs847	chr4:42150091-42151064	Forebrain	327	H>C	0.0007	0.00503715
4	hs1344	chr3:193660817-193662478	Forebrain	1651	H>C	0.001	0.005350819
5	hs1301	chr11:16423269-16426037	Forebrain	885	H>C	0.001	0.005350819
6	hs1746	chrX:150407692-150409052	Forebrain	1353	H>C	0.003	0.010319437
7	hs1393	Chr13:43167371-43169597	Forebrain	2143	H>C	0.003	0.010319437
8	hs599	chr15:37652783-37654460	Forebrain	479	H>C	0.0003	0.003996037
9	hs192	chr3:180773639-180775802	Forebrain	889	H>C	0.004	0.011674514
10.	hs37	chr16:54650598-54651882	Forebrain	601	H>C	0.004	0.011674514
11	hs123	chrX:25400224-25402334	Forebrain	604	H>C	0.009	0.01775084
12	hs540	chr13:71358093-71359507	Forebrain	452	H>C	0.0008	0.005163264
13	hs799	chr7:9271308-9272358	Forebrain	447	H>C	0.003	0.010319437

14	hs1358	chr6:163276830-163279930	Forebrain	3071	H>C	0.01	0.018522066
15	hs1383	chr16:61057518-61059625	Forebrain	2032	H<C	0.01	0.018522066
16	hs1382	chr10:11721307-11722768	Forebrain	1446	H>C	0.01	0.018522066
17	hs1546	chr1:38835996-38838106	Forebrain	644	H>C	0.01	0.018522066
18	hs1636	chr18:25500905-25504214	Forebrain	3276	H<C	0.01	0.018522066
19	hs947	chr18:63078262-63078839	Forebrain	572	H>C	0.01	0.018522066
20	hs1320	chr15:97128054-97130294	Forebrain	1040	H>C	0.007	0.015863604
21	hs112	chr9:973435-975288	Forebrain	1851	H=C	0.02	0.027917316
22	hs1013	chr18:52699870-52701226	Forebrain	1353	H>C	0.02	0.027917316
23	hs1017	chr9:128645462-128647097	Forebrain	725	H=C	0.02	0.027917316
24	hs1526	chr2:104353933-104357342	Forebrain	1381	H>C	0.0007	0.00503715
25	hs187	chr3:71290418-71292584	Forebrain	2160	H>C	0.03	0.036118028
26	hs1529	chr2:104578156-104580488	Forebrain	2326	H>C	0.03	0.036118028
27	hs590	chr18:34719386-34720720	Forebrain	1331	H>C	0.03	0.036118028
28	hs957	chr2:60761404-60763073	Forebrain	1668	H<C	0.03	0.036118028
29	hs1341	chr12:97468703-97471089	Forebrain	2369	H<C	0.04	0.044797554
30	hs1210	chr2:66762515-66765088	Forebrain	419	H>C	0.002	0.008375195
31	hs612	chr1:91305562-91307215	Forebrain	944	H>C	0.04	0.044797554
32	hs886	chr4:181201559-181202529	Forebrain	971	H>C	0.06	0.059314063
33	hs967	chr12:103484342-103485519	Forebrain	1164	H=C	0.07	0.06536596

34	hs1717	chr9:100636218-100640509	Forebrain	4126	H=C	0.06	0.059314063
35	hs1092	chr3:71153556-71155053	Forebrain	1348	H<C	0.05	0.052508036
36	hs170	chr2:164450144-164451758	Forebrain	531	H>C	0.06	0.059314063
37	hs22	chr16:72254566-72255825	Forebrain	1244	H>C	0.06	0.059314063
38	hs204	chr1:213597964-213599524	Forebrain	1556	H>C	0.07	0.06536596
39	hs969	chr2:105317580-105319856	Forebrain	2270	H<C	0.08	0.070782487
40	hs240	chr9:83727123-83728378	Forebrain	1253	H>C	0.09	0.075658712
41	hs1633	chr7:90777214-90780836	Forebrain	3590	H<C	0.09	0.075658712
42	hs840	chr4:66989480-66990366	Forebrain	729	H<C	0.09	0.075658712
43	hs266	chr5:87168414-87169433	Forebrain	1015	H=C	0.22	0.13250178
44	hs322	chr1:87821793-87822910	Forebrain	1109	H>C	0.25	0.142477428
45	hs342	chr14:29860529-29862348	Forebrain	1813	H=C	0.76	0.295295649
46	hs348	chr14:36020024-36020998	Forebrain	973	H>C	0.2	0.125287468
47	hs408	chr1:10851570-10852173	Forebrain	601	H>C	0.24	0.139250228
48	hs582	chrX:81464240-81465016	Forebrain	771	H>C	0.24	0.139250228
49	hs623	chr15:57426028-57426952	Forebrain	925	H<C	0.99	0.353105855
50	hs653	chr3:137185964-137186866	Forebrain	901	H<C	0.38	0.180293604
51	hs656	hr10:131400948-131402279	Forebrain	1314	H>C	0.18	0.117470266
52	hs702	chr2:105132815-105133830	Forebrain	1016	H=C	0.71	0.281333874
53	hs781	chr8:21907426-21908282	Forebrain	261	-	1	0.355404949

54	hs796	chr13:95313852-95315441	Forebrain	1589	H>C	0.71	0.281333874
55	hs807	chr7:22091362-22092557	Forebrain	1195	H<C	1	0.355404949
56	hs853	chr5:87083012-87084752	Forebrain	1740	H=C	0.77	0.298023159
57	hs855	chr11:31989173-31990022	Forebrain	850	H>C	0.16	0.108971304
58	hs883	chr11:16311593-16312881	Forebrain	1251	H<C	0.99	0.353105855
59	hs978	chr6:97754043-97755513	Forebrain	1470	H=C	0.6	0.248581858
60	hs987	chr9:128869446-128870934	Forebrain	1480	H<C	0.23	0.135904236
61	hs1011	chr18:76461276-76462723	Forebrain	1425	H<C	0.11	0.084084298
62	hs1024	chr5:92312840-92314645	Forebrain	1803	H<C	0.19	0.121458855
63	hs1025	chr2:73124730-73126091	Forebrain	1361	H>C	0.29	0.154743905
64	hs1128	chr6:98829860-98831049	Forebrain	1189	H<C	0.27	0.148599886
65	hs1161	chr1:88025863-88027203	Forebrain	1341	H>C	0.11	0.084084298
66	hs1166	chr14:36973775-36974585	Forebrain	810	H<C	0.77	0.298023159
67	hs1224	chr3:147651676-147653436	Forebrain	1480	H<C	0.45	0.200346534
68	hs1303	chr2:104667872-104670648	Forebrain	2754	H<C	0.26	0.14559205
69	hs1324	chr1:213498112-213501134	Forebrain	2975	H<C	0.43	0.19443821
70	hs1326	chr8:59941214-59943636	Forebrain	2415	H<C	0.46	0.20323294
71	hs1417	chr9:98274342-98275314	Forebrain	965	H<C	0.84	0.316540146
72	hs1537	chr18:53018678-53020044	Forebrain	1364	H>C	0.31	0.160957249
73	hs1538	chr14:36911162-36914360	Forebrain	3160	H>C	0.29	0.154743905

74	hs1548	chr21:34221456-34223948	Forebrain	2488	H<C	0.11	0.084084298
75	hs1566	chr18:23432723-23434825	Forebrain	2042	H<C	0.53	0.226138946
76	hs1579	chr14:57320664-57324319	Forebrain	3649	H<C	0.45	0.200346534
77	hs1588	chr10:35925382-35927242	Forebrain	1798	H<C	0.16	0.108971304
78	hs1597	chr9:100636218-100637962	Forebrain	1694	H<C	0.15	0.104437671
79	hs1334	chr10:37054745-37057224	Forebrain	2354	H<C	0.4	0.185220656
80	hs71	chr16:51671181-51672039	Forebrain	859	H=C	0.1	0.080071637
81	hs110	chr7:21003280-21004750	Forebrain	1446	H=C	0.35	0.172430483
82	hs119	chrX:24915382-24918272	Forebrain	2884	H=C	0.6	0.248581858
83	hs411	chr2:156726581-156727605	Forebrain	1019	H<C	0.2	0.125287468
84	hs692	chr11:15587041-15588314	Forebrain	1272	H=C	0.4	0.185220656
85	hs121	chrX:25007879-25009581	Forebrain	1699	H=C	1	0.355404949
86	hs818	chr9:128520992-128522653	Forebrain	1660	H=C	0.2	0.125287468
87	hs244	chr2:174988737-174990363	Forebrain	1616	H<C	0.69	0.275593023
88	hs111	chr7:42191728-42193638	Forebrain	1906	H=C	0.5	0.216104839
89	hs914	chr20:21214790-21217232	Forebrain	2221	H>C	0.3	0.157893018
90	hs675	chr2:144103882-144105644	Forebrain	1760	H<C	0.46	0.20323294
91	hs566	chr14:29684896-29686744	Forebrain	1838	H>C	0.21	0.128965541
92	hs399	chr2:60441495-60442515	Forebrain	1011	H>C	0.23	0.135904236
93	hs1316	chr3:62405817-62408099	Forebrain	2267	H=C	0.2	0.125287468

94	hs671	chr1:97610491-97611741	Forebrain	1232	H<C	0.99	0.353105855
95	hs860	chr2:175196043-175197114	Forebrain	1056	H<C	0.91	0.334105346
96	hs622	chr14:99466200-99467144	Forebrain	746	H<C	0.73	0.28698445
97	hs416	chr2:162094895-162095451	Forebrain	323	H>C	0.37	0.177737926
98	hs541	chr2:45030569-45032739	Forebrain	2158	H<C	0.28	0.151506334
99	hs775	chr18:77010009-77010795	Forebrain	779	H<C	0.72	0.284170313
100	hs218	chr7:114056847-114058647	Forebrain	1799	H<C	0.99	0.353105855
101	hs262	chr5:76940836-76941396	Forebrain	196	H<C	1	0.355404949
102	hs434	chr3:62350726-62351718	Forebrain	993	H<C	1	0.355404949
103	hs122	chrX:25017067-25018756	Forebrain	1486	H<C	0.2	0.125287468
104	hs748	chr10:78390590-78391875	Forebrain	1280	H<C	0.32	0.163939985
105	hs194	chr1:51034546-51036289	Midbrain	1742	H>C	0.04	0.044797554
106	hs430	chr19:30840299-30843536	Midbrain	632	H>C	0.00005	0.001605246
107	hs669	chr8:92824759-92826618	Midbrain	1829	H>C	0.02	0.027917316
108	hs765	chr9:81823297-81824667	Midbrain	1366	H>C	0.004	0.011674514
109	hs975	chr2:59304974-59306893	Midbrain	1917	H<C	0.02	0.027917316
110	hs1366	chr6:38358690-38360084	Midbrain	1380	H>C	0.00009	0.002167082
111	hs1180	chr18:22616831-22618682	Midbrain	1848	H>C	0.03	0.036118028
112	hs1632	chr11:116521882-116522627	Midbrain	630	H>C	0.0005	0.004671984
113	hs1734	chr5:106909516-106911012	Midbrain	1483	H>C	0.005	0.013315402

114	hs1857	chr1:44500383-44503337	Midbrain	2885	H>C	0.001	0.005350819
115	hs1575	chr12:103570982-103573398	Midbrain	2404	H>C	0.004	0.011674514
116	hs935	chr10:119310483-119311458	Midbrain	972	H>C	0.01	0.018522066
117	hs186	chr9:129198400-129200739	Midbrain	2338	H>C	0.01	0.018522066
118	hs1860	chr20:49306307-49309162	Midbrain	2837	H<C	0.02	0.027917316
119	hs661	chr12:16940708-16942322	Midbrain	778	H>C	0.02	0.027917316
120	hs712	chr4:84700011-84701265	Midbrain	732	H>C	0.02	0.027917316
121	hs830	chr15:38159507-38161007	Midbrain	1501	H=C	0.02	0.027917316
122	hs930	chr4:111669259-111671168	Midbrain	1241	H>C	0.03	0.036118028
123	hs260	chr4:105345575-105346895	Midbrain	645	H>C	0.03	0.036118028
124	hs394	chr2:59746377-59746992	Midbrain	615	H=C	0.07	0.06536596
125	hs559	chr4:112421802-112422905	Midbrain	1101	H>C	0.08	0.070782487
126	hs1227	chr5:91271776-91272886	Midbrain	1106	H>C	0.08	0.070782487
127	hs118	chrX:24894335-24896084	Midbrain	1733	H<C	0.4	0.185220656
128	hs149	chr2:45106927-45107653	Midbrain	722	H=C	0.2	0.125287468
129	hs181	chr15:37240805-37242498	Midbrain	1684	H>C	0.1	0.080071637
130	hs567	chr6:98461952-98463309	Midbrain	1353	H<C	0.34	0.169673637
131	hs573	chr2:157861101-157862409	Midbrain	1308	H>C	0.6	0.248581858
132	hs575	chr13:68429117-68430526	Midbrain	1409	H<C	0.1	0.080071637
133	hs593	chr14:37726340-37727348	Midbrain	1009	H>C	0.2	0.125287468

134	hs413	chr2:157551014-157551952	Midbrain	930	H=C	0.38	0.180293604
135	hs627	chr17:37774485-37774988	Midbrain	456	H=C	0.5	0.216104839
136	hs195	chr8:106333530-106334859	Midbrain	1320	H=C	0.5	0.216104839
137	hs209	chr3:137047544-137049271	Midbrain	1709	H<C	0.39	0.1827871
138	hs261	chr5:3511978-3513399	Midbrain	1422	H<C	0.34	0.169673637
139	hs277	chr1:44715420-44716129	Midbrain	694	H<C	0.28	0.151506334
140	hs298	chr7:96633582-96634303	Midbrain	717	-	1	0.355404949
141	hs314	chr9:126537718-126539929	Midbrain	2212	H>C	0.43	0.19443821
142	hs720	chr7:113793545-113794562	Midbrain	1018	H<C	0.86	0.321652627
143	hs793	chr7:10684318-10685359	Midbrain	1032	H<C	0.42	0.191414154
144	hs813	chr5:158143424-158144725	Midbrain	1302	H<C	0.28	0.151506334
145	hs851	chr18:36763763-36764791	Midbrain	1027	H<C	0.85	0.319105983
146	hs863	chr11:31502035-31503157	Midbrain	1077	H<C	0.4	0.185220656
147	hs690	chr2:63193855-63194929	Midbrain	1069	H<C	0.9	0.331651513
148	hs701	chr7:21084342-21085460	Midbrain	1084	H<C	0.1	0.080071637
149	hs1015	chr9:128919674-128920432	Midbrain	759	H=C	0.33	0.166844433
150	hs1093	chr2:103792328-103793819	Midbrain	1477	H=C	0.2	0.125287468
151	hs1115	chr3:148006499-148007810	Midbrain	1294	H>C	0.1	0.080071637
152	hs1218	chr14:57430887-57432346	Midbrain	1458	H<C	0.1	0.080071637
153	hs1425	chr7:69596979-69598263	Midbrain	1270	H<C	0.32	0.163939985

154	hs1648	chr3:114936330-114938229	Midbrain	1895	H<C	0.39	0.1827871
155	hs1702	chr1:18958671-18960284	Midbrain	1602	H<C	0.1	0.080071637
156	hs1791	chr14:57474144-57478090	Midbrain	3913	H<C	0.4	0.185220656
157	hs1802	chr2:145339602-145341530	Midbrain	1759	H<C	0.1	0.080071637
158	hs605	chr12:17657732-17659008	Midbrain	1275	H=C	0.1	0.080071637
159	hs1867	chr2:170869607-170871165	Midbrain	1537	H=C	0.38	0.180293604
160	hs1726	chr18:49279374-49281480	Hindbrain	1092	H>C	0.0005	0.004671984
161	hs161	chr16:52446050-52447237	Hindbrain	1163	H>C	0.001	0.005350819
162	hs101	chr16:48912816-48914144	Hindbrain	1322	H>C	0.004	0.011674514
163	hs828	chr15:36964819-36966098	Hindbrain	1272	H>C	0.007	0.015863604
164	hs563	chr6:98491829-98493238	Hindbrain	416	H>C	0.002	0.008375195
165	hs2144	chr11:19194084-19196536	Hindbrain	2408	H>C	0.008	0.016872655
166	hs628	chr9:159657-160780	Hindbrain	1105	H>C	0.01	0.018522066
167	hs1086	chr20:39334182-39335059	Hindbrain	877	H>C	0.01	0.018522066
168	hs562	chr10:131106522-131108742	Hindbrain	1666	H>C	0.02	0.027917316
169	hs327	chr1:88926796-88928508	Hindbrain	621	H>C	0.03	0.036118028
170	hs1535	chr2:60498057-60502013	Hindbrain	3924	H<C	0.06	0.059314063
171	hs529	chr9:17322200-17324371	Hindbrain	2060	H<C	0.095	0.07791886
172	hs137	chr13:72300849-72302934	Hindbrain	1870	H<C	0.11	0.084084298
173	hs401	chr2:104736518-104737365	Hindbrain	848	H=C	0.34	0.169673637

174	hs210	chr3:137067622-137068925	Hindbrain	1304	H=C	0.13	0.094735811
175	hs232	chr10:131691086-131692848	Hindbrain	1747	H<C	0.31	0.160957249
176	hs246	chr2:176940070-176941410	Hindbrain	1330	H<C	1	0.355404949
177	hs296	chr7:26728697-26729802	Hindbrain	1106	H>C	0.5	0.216104839
178	hs307	chr9:16710536-16711184	Hindbrain	649	H>C	0.27	0.148599886
179	hs155	chr16:53948201-53949846	Hindbrain	1592	H<C	0.22	0.13250178
180	hs640	chr2:164574007-164575458	Hindbrain	1417	H=C	0.2	0.125287468
181	hs662	chr2:157720628-157721586	Hindbrain	952	H=C	0.86	0.321652627
182	hs679	chr18:45087290-45088074	Hindbrain	784	H>C	0.77	0.298023159
183	hs705	chr1:3190581-3191428	Hindbrain	824	H=C	0.35	0.172430483
184	hs816	chr7:14379627-14380740	Hindbrain	1110	H=C	0.9	0.331651513
185	hs330	chr10:126905322-126906003	Hindbrain	654	H=C	0.64	0.260831451
186	hs966	chr7:114326912-114329772	Hindbrain	2825	H<C	0.35	0.172430483
187	hs993	chr12:17311784-17313759	Hindbrain	1972	H=C	0.14	0.099697329
188	hs1081	chr6:98902034-98904516	Hindbrain	2455	H<C	0.15	0.104437671
189	hs592	chr14:36814302-36815937	Hindbrain	1631	H=C	0.44	0.197415308
190	hs603	chr5:3182218-3183271	Hindbrain	1054	H=C	0.23	0.135904236
191	hs1139	chr1:39248757-39250129	Hindbrain	1367	H>C	0.13	0.094735811
192	hs1142	chr2:60855056-60856888	Hindbrain	1769	H<C	0.18	0.117470266
193	hs1235	chr1:164620038-164621164	Hindbrain	1093	H<C	0.89	0.329179528

194	hs363	chr17:35329349-35329944	Hindbrain	589	H=C	0.59	0.245455758
195	hs1539	chr14:29710885-29713340	Hindbrain	2412	H<C	0.21	0.128965541
196	hs191	chr5:91036888-91038899	Hindbrain	1959	H>C	0.38	0.180293604
197	hs2094	chr1:10795106-10799241	Hindbrain	4122	H<C	0.11	0.084084298
198	hs304	chr9:8095553-8096166	Mid/Fore	614	H>C	0.01	0.018522066
199	hs1346	chr21:34465959-34469066	Mid/Fore	3086	H=C	0.01	0.018522066
200	hs1391	chr6:3349397-3352257	Mid/Fore	2808	H<C	0.02	0.027917316
201	hs1563	chr3:193489359-193491333	Mid/Fore	1970	H<C	0.004	0.011674514
202	hs1638	chr5:55896173-55899069	Mid/Fore	2893	H=C	0.01	0.018522066
203	hs1724	chr16:73362809-73364292	Mid/Fore	1480	H>C	0.0008	0.005163264
204	hs1308	chr7:127174386-127177546	Mid/Fore	3092	H<C	0.02	0.027917316
205	hs1032	chr10:119309200-119310544	Mid/Fore	971	H<C	0.02	0.027917316
206	hs1545	chr4:109254340-109257033	Mid/Fore	799	H=C	0.02	0.027917316
207	hs1571	chr12:114101195-114103805	Mid/Fore	1730	H<C	0.04	0.044797554
208	hs1577	chr5:91927845-91931024	Mid/Fore	3153	H<C	0.02	0.027917316
209	hs1723	chr12:103613944-103615320	Mid/Fore	1363	H<C	0.003	0.010319437
210	hs1363	chr18:42942363-42944135	Mid/Fore	1764	H<C	0.03	0.036118028
211	hs646	chr2:172820365-172821314	Mid/Fore	945	H<C	0.99	0.353105855
212	hs654	chr3:147801015-147802169	Mid/Fore	1138	H>C	0.27	0.148599886
213	hs672	chr10:120074039-120075696	Mid/Fore	1657	H<C	0.24	0.139250228

214	hs699	chr10:130831457-130833175	Mid/Fore	1714	H<C	0.67	0.269759715
215	hs779	chr2:60352514-60353602	Mid/Fore	1089	H<C	0.49	0.212939598
216	hs841	chr10:118854124-118855243	Mid/Fore	1114	H>C	0.1	0.080071637
217	hs956	chr7:114299711-114302078	Mid/Fore	2362	H=C	0.44	0.197415308
218	hs1131	chr2:105032493-105034445	Mid/Fore	1937	H<C	0.22	0.13250178
219	hs1318	chr8:77598007-77600645	Mid/Fore	2638	H<C	0.73	0.28698445
220	hs1523	chr14:29857930-29860548	Mid/Fore	2600	H=C	0.36	0.175117711
221	hs1540	chr12:103405110-103408796	Mid/Fore	3669	H>C	0.11	0.084084298
222	hs271	chr5:93226985-93228322	Mid/Fore	1336	H=C	0.18	0.117470266
223	hs281	chr6:41523224-41523677	Mid/Fore	454	H=C	0.19	0.121458855
224	hs435	chr3:62359866-62360569	Mid/Fore	705	H>C	0.3	0.157893018
225	hs565	chr11:31622822-31624118	Mid/Fore	1247	H<C	0.26	0.14559205
226	hs619	chr13:72333516-72334988	Mid/Fore	1472	H>C	1	0.355404949
227	hs1394	chr13:78406128-78407714	Fore/Hind	1577	H<C	0.02	0.027917316
228	hs1568	chr13:28318579-28320134	Fore/Hind	588	H<C	0.03	0.036118028
229	hs754	chr5:3197865-3198942	Fore/Hind	1078	H=C	0.02	0.027917316
230	hs1060	chr5:92613862-92616844	Fore/Hind	2974	H<C	0.04	0.044797554
231	hs1202	chr1:164604141-164605474	Fore/Hind	1328	H>C	0.04	0.044797554
232	hs643	chr9:23004730-23005789	Fore/Hind	1059	H>C	0.06	0.059314063
233	hs1027	chr18:22744668-22746270	Fore/Hind	1597	H>C	0.09	0.075658712

234	hs433	chr14:30741750-30743626	Fore/Hind	1877	H<C	0.53	0.226138946
235	hs611	chr12:111495397-111496252	Fore/Hind	856	H<C	1	0.355404949
236	hs625	chr16:49735099-49736449	Fore/Hind	1347	H>C	0.15	0.104437671
237	hs12	chr16:78510608-78511944	Fore/Hind	1216	H<C	0.5	0.216104839
238	hs426	chrX:81788884-81790571	Fore/Hind	1656	H=C	0.15	0.104437671
239	hs1064	chr14:29226075-29227673	Fore/Hind	1594	H=C	0.26	0.14559205
240	hs427	chrX:139169379-139171545	Fore/Hind	2157	H>C	0.28	0.151506334
241	hs1385	chr2:3268196-3270849	Mid/Hind	2612	H>C	0.0003	0.003996037
242	hs20	chr16:72738568-72740149	Mid/Hind	1569	H>C	0.002	0.008375195
243	hs980	chr12:17848111-17849347	Mid/Hind	1231	H>C	0.0004	0.004393304
244	hs2064	chr6:52253728-52256212	Mid/Hind	2469	H>C	0.006	0.014692079
245	hs737	chr10:130366868-130368005	Mid/Hind	1138	H>C	0.03	0.036118028
246	hs568	chr2:146692288-146693283	Mid/Hind	994	H>C	0.05	0.052508036
247	hs1205	chr20:21488551-21490021	Mid/Hind	1360	H>C	0.06	0.059314063
248	hs1418	chr7:155264047-155265809	Mid/Hind	1728	H>C	0.07	0.06536596
249	hs217	chr6:51148668-51149710	Mid/Hind	1041	H<C	0.46	0.20323294
250	hs282	chr6:98116085-98116943	Mid/Hind	858	H>C	0.36	0.175117711
251	hs371	chr18:35063482-35064528	Mid/Hind	1047	H>C	0.244	0.140555005
252	hs704	chr14:36933150-36934532	Mid/Hind	1379	H<C	0.6	0.248581858
253	hs749	chr7:13450920-13451719	Mid/Hind	792	H=C	0.13	0.094735811

254	hs755	chrX:136316806-136317871	Mid/Hind	1062	H>C	0.93	0.33895935
255	hs762	chr1:163441941-163442842	Mid/Hind	902	H<C	0.99	0.353105855
256	hs865	chr6:50685244-50686237	Mid/Hind	992	H=C	0.84	0.316540146
257	hs901	chr9:37251207-37252223	Mid/Hind	1016	H>C	0.21	0.128965541
258	hs1030	chr9:128516934-128518372	Mid/Hind	1428	H<C	0.52	0.222822908
259	hs1192	chr7:114463797-114464462	Mid/Hind	659	H=C	0.23	0.135904236
260	hs1213	chr7:42252831-42254560	Fore/Mid/Hind	1724	H>C	0.003	0.010319437
261	hs1573	chr3:147563409-147566604	Fore/Mid/Hind	3188	H<C	0.001	0.005350819
262	hs2223	chr10:79935570-79940095	Fore/Mid/Hind	4488	H<C	0.003	0.010319437
263	hs1534	chr2:105044282-105047512	Fore/Mid/Hind	3224	H=C	0.06	0.059314063
264	hs1544	chr18:23044107-23046853	Fore/Mid/Hind	2726	H=C	0.06	0.059314063
265	hs1325	chr7:25791903-25794282	Fore/Mid/Hind	2376	H=C	0.07	0.06536596
266	hs269	chr5:90928612-90929226	Fore/Mid/Hind	615	H=C	0.6	0.248581858
267	hs532	chr13:28395961-28397536	Fore/Mid/Hind	1573	H=C	0.7	0.278474868
268	hs981	chr4:113442390-113443530	Fore/Mid/Hind	1133	H=C	0.14	0.099697329
269	hs1006	chr10:102244842-102246334	Fore/Mid/Hind	1492	H>C	0.14	0.099697329
270	hs1360	chr9:82276120-82278534	Fore/Mid/Hind	2377	H=C	0.33	0.166844433
271	hs1578	chr2:212254840-212257158	Fore/Mid/Hind	2312	H<C	0.17	0.11331146

Table A2. Results with *FHL1*-Intron 1 (GRCh37: Chr X: 135252140-135288565) proxy region for 86 enhancers

SN	VISTA ID	VISTA Coordinates (GRCh37/hg19)	Expression Domain	Alignment Length (bp)	P-Value
1	hs526	chr4:1,613,479-1,614,106	Forebrain	620	0.00011
2	hs1019	chr7:20,838,843-20,840,395	Forebrain	471	0.0001
3	hs847	chr4:42,150,091-42,151,064	Forebrain	327	0.009
4	hs1344	chr3:193,660,817-193,662,478	Forebrain	1651	0.01
5	hs1301	chr11:16,423,269-16,426,037	Forebrain	885	0.01
6	hs540	chr13:71,358,093-71,359,507	Forebrain	452	0.01
7	hs799	chr7:9,271,308-9,272,358	Forebrain	447	0.04
8	hs1210	chr2:66,762,515-66,765,088	Forebrain	419	0.03
9	hs1746	chrX:150,407,692-150,409,052	Forebrain	1353	0.07
10	hs1393	Chr13:43,167,371-43,169,597	Forebrain	2143	0.05
11	hs599	chr15:37,652,783-37,654,460	Forebrain	479	0.003
12	hs192	chr3:180,773,639-180,775,802	Forebrain	889	0.08
13	hs37	chr16:54,650,598-54,651,882	Forebrain	601	0.05
14	hs123	chrX:25,400,224-25,402,334	Forebrain	604	0.15
15	hs1358	chr6:163,276,830-163,279,930	Forebrain	3071	0.34
16	hs1383	chr16:61,057,518-61,059,625	Forebrain	2032	0.3
17	hs1382	chr10:11,721,307-11,722,768	Forebrain	1446	0.24
18	hs1546	chr1:38,835,996-38,838,106	Forebrain	644	0.21
19	hs1636	chr18:25,500,905-25,504,214	Forebrain	3276	0.26
20	hs947	chr18:63,078,262-63,078,839	Forebrain	572	0.23
21	hs1320	chr15:97,128,054-97,130,294	Forebrain	1040	0.11
22	hs112	chr9:973,435-975,288	Forebrain	1851	0.4
23	hs1013	chr18:52,699,870-52,701,226	Forebrain	1353	0.46
24	hs1017	chr9:128,645,462-128,647,097	Forebrain	725	0.22
25	hs1526	chr2:104,353,933-104,357,342	Forebrain	1381	0.005
26	hs187	chr3:71,290,418-71,292,584	Forebrain	2160	0.63

27	hs1529	chr2:104,578,156-104,580,488	Forebrain	2326	0.56
28	hs590	chr18:34,719,386-34,720,720	Forebrain	1331	0.51
29	hs957	chr2:60,761,404-60,763,073	Forebrain	1668	0.6
30	hs1341	chr12:97,468,703-97,471,089	Forebrain	2369	0.8
31	hs612	chr1:91,305,562-91,307,215	Forebrain	944	0.8
32	hs194	chr1:51,034,546-51,036,289	Midbrain	1742	0.25
33	hs430	chr19:30,840,299-30,843,536	Midbrain	632	0.0003
34	hs669	chr8:92,824,759-92,826,618	Midbrain	1829	0.44
35	hs765	chr9:81,823,297-81,824,667	Midbrain	1366	0.07
36	hs975	chr2:59,304,974-59,306,893	Midbrain	1917	0.4
37	hs1366	chr6:38,358,690-38,360,084	Midbrain	1380	0.0002
38	hs1180	chr18:22,616,831-22,618,682	Midbrain	1848	0.6
39	hs1632	chr11:116,521,882-116,522,627	Midbrain	630	0.006
40	hs1734	chr5:106,909,516-106,911,012	Midbrain	1483	0.09
41	hs1857	chr1:44,500,383-44,503,337	Midbrain	2885	0.01
42	hs1575	chr12:103,570,982-103,573,398	Midbrain	2404	0.05
43	hs935	chr10:119,310,483-119,311,458	Midbrain	972	0.19
44	hs186	chr9:129,198,400-129,200,739	Midbrain	2338	0.25
45	hs1860	chr20:49,306,307-49,309,162	Midbrain	2837	0.6
46	hs661	chr12:16,940,708-16,942,322	Midbrain	778	0.34
47	hs712	chr4:84,700,011-84,701,265	Midbrain	732	0.31
48	hs830	chr15:38,159,507-38,161,007	Midbrain	1501	0.46
49	hs930	chr4:111,669,259-111,671,168	Midbrain	1241	0.6
50	hs260	chr4:105,345,575-105,346,895	Midbrain	645	0.42
51	hs1726	chr18:49,279,374-49,281,480	Hindbrain	1092	0.006
52	hs161	chr16:52,446,050-52,447,237	Hindbrain	1163	0.02
53	hs101	chr16:48,912,816-48,914,144	Hindbrain	1322	0.08
54	hs828	chr15:36,964,819-36,966,098	Hindbrain	1272	0.13
55	hs563	chr6:98,491,829-98,493,238	Hindbrain	416	0.04

56	hs2144	chr11:19,194,084-19,196,536	Hindbrain	2408	0.15
57	hs628	chr9:159,657-160,780	Hindbrain	1105	0.23
58	hs1086	chr20:39,334,182-39,335,059	Hindbrain	877	0.26
59	hs562	chr10:131,106,522-131,108,742	Hindbrain	1666	
60	hs327	chr1:88,926,796-88,928,508	Hindbrain	621	0.36
61	hs304	chr9:8,095,553-8,096,166	Mid/Fore	614	0.15
62	hs1346	chr21:34,465,959-34,469,066	Mid/Fore	3086	0.21
63	hs1391	chr6:3,349,397-3,352,257	Mid/Fore	2808	0.51
64	hs1563	chr3:193,489,359-193,491,333	Mid/Fore	1970	0.06
65	hs1638	chr5:55,896,173-55,899,069	Mid/Fore	2893	0.38
66	hs1724	chr16:73,362,809-73,364,292	Mid/Fore	1480	0.007
67	hs1308	chr7:127,174,386-127,177,546	Mid/Fore	3092	0.54
68	hs1032	chr10:119,309,200-119,310,544	Mid/Fore	971	0.33
69	hs1545	chr4:109,254,340-109,257,033	Mid/Fore	799	0.38
70	hs1571	chr12:114,101,195-114,103,805	Mid/Fore	1730	0.7
71	hs1577	chr5:91,927,845-91,931,024	Mid/Fore	3153	0.5
72	hs1723	chr12:103,613,944-103,615,320	Mid/Fore	1363	0.06
73	hs1363	chr18:42,942,363-42,944,135	Mid/Fore	1764	0.8
74	hs1394	chr13:78,406,128-78,407,714	Hind/Fore	1577	0.4
75	hs1568	chr13:28,318,579-28,320,134	Hind/Fore	588	0.4
76	hs754	chr5:3,197,865-3,198,942	Hind/Fore	1078	0.44
77	hs1060	chr5:92,613,862-92,616,844	Hind/Fore	2974	1
78	hs1202	chr1:164,604,141-164,605,474	Hind/Fore	1328	0.74
79	hs1385	chr2:3,268,196-3,270,849	Hind/Mid	2612	0.0004
80	hs20	chr16:72,738,568-72,740,149	Hind/Mid	1569	0.03
81	hs980	chr12:17,848,111-17,849,347	Hind/Mid	1231	0.003
82	hs2064	chr6:52,253,728-52,256,212	Hind/Mid	2469	0.09
83	hs737	chr10:130,366,868-130,368,005	Hind/Mid	1138	0.53
84	hs1213	chr7:42,252,831-42,254,560	Hind-Mid-Fore	1724	0.04

85	hs1573	chr3:147,563,409-147,566,604	Hind-Mid-Fore	3188	0.008
86	hs2223	chr10:79,935,570-79,940,095	Hind-Mid-Fore	4488	0.03

7.2 Appendix -II: Determination of fast evolving enhancer with local intronic proxy regions

Table A3: Results with locus specific intronic/NCNRS proxy region for previously shortlisted 86 Enhancers

SN	VISTA ID	VISTA Coordinates (GRCh37/hg19)	Expression Domain	Alignment Length (bp)	Proxy Within 100kb	Proxy Coordinates	Proxy Alignment Length (bp)	Distance From Proxy (kb)	P-Value
1	hs1344	chr3:193660817-193662478	Forebrain	1651	NCNRS	chr3:193626500-193628500	1973	32.3	0.63
2	hs187	chr3:71290418-71292584	Forebrain	2160	FOXP1*	chr3:71003844-71633140	30311	Intragenic	1
3	**hs192**	**chr3:180773639-180775802**	**Forebrain**	**889**	**FXR1**	**chr3:180585929-180700541**	**21996**	**73.1**	**0.04**
4	**hs1301**	**chr11:16423269-16426037**	**Forebrain**	**885**	**SOX6***	**chr11:15987995-16761138**	**31688**	**Intragenic**	**0.02**
5	**hs526**	**chr4:1613479-1614106**	**Forebrain**	**620**	**SLBP**	**chr4:1694527-1714282**	**4644**	**80.4**	**0.03**
6	hs957	chr2:60761404-60763073	Forebrain	1668	BCL11A	chr2:60678302-60780702	4970	Intragenic	0.33
7	hs1636	chr18:25500905-25504214	Forebrain	3276	CDH2*	chr18:25530930-25757410	26489	26.7	0.8
8	hs1013	chr18:52699870-52701226	Forebrain	1353	CCDC68	chr18:52568740-52626739	17273	73.1	1
9	**hs540**	**chr13:71358093-71359507**	**Forebrain**	**452**	**NCNRS**	**chr13:71343593-71345593**	**1990**	**12.5**	**0.03**
10	**hs37**	**chr16:54650598-54651882**	**Forebrain**	**601**	**NCNRS**	**chr16:54687882-54690000**	**2111**	**36**	**0.02**
11	hs123	chrX:25400224-	Forebrain	604	NCNRS	chrX:	1982	0.9	0.07

		25402334				25403300-25405300			
12	hs599	chr15:37652783-37654460	Forebrain	479	NCNRS	chr15:37639555-37642006	2240	10.7	0.09
13	hs799	chr7:9271308-9272358	Forebrain	447	NCNRS	chr7: 9311358-9313358	1987	39	0.5
14	**hs1210**	**chr2:66762515-66765088**	**Forebrain**	**419**	**MEIS1**	**chr2:66660584-66801001**	**12410**	**Intragenic**	**0.03**
15	**hs847**	**chr4:42150091-42151064**	**Forebrain**	**327**	**BEND4**	**chr4:42112955-42154895**	**9570**	**Intragenic**	**0.03**
16	**hs1019**	**chr7:20838843-20840395**	**Forebrain**	**471**	**ABCB5**	**chr7:20654830-20816658**	**32041**	**22.2**	**0.006**
17	hs1320	chr15:97128054-97130294	Forebrain	1040	NCNRS	chr15:97143224-97144527	1301	12.9	0.6
18	hs590	chr18:34719386-34720720	Forebrain	1331	KIAA1328	chr18:34409069-34812135	3931	Intragenic	0.7
19	hs612	chr1:91305562-91307215	Forebrain	944	ZNF644	chr1:91380859-91487829	5558	73.6	0.1
20	hs1529	chr2:104578156-104580488	Forebrain	2326	NCNRS	chr2:104635606-104638038	2413	55.1	0.83
21	hs112	chr9:973435-975288	Forebrain	1851	DMRT1	chr9:841690-969090	7058	4.4	1
22	hs1393	Chr13:43167371-43169597	Forebrain	2143	TNFSF11	chr13:43136872-43182149	4940	Intragenic	1
23	hs947	chr18:63078262-63078839	Forebrain	572	NCNRS	chr18:63081862-63083862	1991	3	0.5
24	1017	chr9:128645462-128647097	Forebrain	725	PBX3 *	chr9:128509624-128729656	16350	Intragenic	0.7

25	hs1546	chr1:38835996-38838106	Forebrain	644	NCNRS	chr1:38920506-38922766	2258	82.4	0.3
26	hs1382	chr10:11721307-11722768	Forebrain	1446	ECHDC3	chr10:11784365-11806069	6661	61.6	0.9
27	hs1383	chr16:61057518-61059625	Forebrain	2032	NCNRS	chr16:61020069-61022372	2285	35.1	0.7
28	hs1358	chr6:163276830-163279930	Forebrain	3071	PACRG	chr6:163148164-163736524	10064	Intragenic	1
29	hs1341	chr12:97468703-97471089	Forebrain	2369	NCNRS	chr12:97441203-97443540	2295	25.2	1
30	hs1746	chrX:150407692-150409052	Forebrain	1353	NCNRS	chrX:150415352-150416792	1342	6.3	1
31	**hs1526**	**chr2:104353933-104357342**	**Forebrain**	**1381**	**NCNRS**	**chr2:104388797-104390900**	**2096**	**31.5**	**0.03**
32	hs194	chr1:51034546-51036289	Midbrain	1742	FAF1*	chr1:50905150-51425935	36949	Intragenic	0.4
33	**hs430**	**chr19:30840299-30843536**	**Midbrain**	**632**	**ZNF536 ***	**chr19:30719197-31204445**	**7435**	**Intragenic**	**0.0007**
34	hs669	chr8:92824759-92826618	Midbrain	1829	NCNRS	chr8:92811876-92814003	2101	10.8	1
35	hs765	chr9:81823297-81824667	Midbrain	1366	NCNRS	chr9:81837401-81839446	1997	12.7	1
36	hs975	chr2:59304974-59306893	Midbrain	1917	NCNRS	chr2:59345351-59347342	1972	38.5	1
37	**hs1366**	**chr6:38358690-38360084**	**Midbrain**	**1380**	**BTBD9**	**chr6:38136227-38607924**	**17735**	**Intragenic**	**0.03**
38	hs1180	chr18:22616831-22618682	Midbrain	1848	ZNF521	chr18:22641890-22932154	8264	23.2	1

39	**hs1632**	**chr11:116521882-116522627**	**Midbrain**	**630**	**BUD13**	**chr11:116618886-116643704**	**10609**	**96.3**	**0.04**
40	hs1734	chr5:106909516-106911012	Midbrain	1483	EFNA5*	chr5:106712590-107006596	5207	Intragenic	0.07
41	hs1857	chr1:44500383-44503337	Midbrain	2885	SLC6A9	chr1:44457172-44497139	9372	3.2	0.2
42	hs1575	chr12:103570982-103573398	Midbrain	2404	C12ORF42	chr12:103631369-103889749	9588	58	0.4
43	hs935	chr10:119310483-119311458	Midbrain	972	EMX2	chr10:119301955-119309056	2117	1.4	0.5
44	hs186	chr9:129198400-129200739	Midbrain	2338	MVB12B	chr9:129089128-129269320	19247	Intragenic	1
45	hs1860	chr20:49306307-49309162	Midbrain	2837	FAM65C	chr20:49202645-49308065	18731	Intragenic	0.36
46	hs661	chr12:16940708-16942322	Midbrain	778	NCNRS	chr12:16986322-16988079	1717	44	1
47	hs712	chr4:84700011-84701265	Midbrain	732	NCNRS	chr4:84714391-84716099	1691	13.1	1
48	hs830	chr15:38159507-38161007	Midbrain	1501	TMCO5A	chr15:38214140-38259925	4941	53.1	1
49	hs930	chr4:111669259-111671168	Midbrain	1241	PITX2	chr4:111690073-111692163	8472	18.9	0.27
50	hs260	chr4:105345575-105346895	Midbrain	645	CXXC4	chr4:105353067-105354400	2320	6.1	1
51	hs101	chr16:48912816-48914144	Hindbrain	1322	NCNRS	chr16:48931800-48933240	1435	17.6	0.4
52	hs161	chr16:52446050-52447237	Hindbrain	1163	TOX3*	chr16:52471917-52581714	10692	24.7	0.2

53	hs628	chr9:159657-160780	Hindbrain	1105	CBWD1	chr9:121041-188979	10576	Intragenic	1
54	hs828	chr15:36964819-36966098	Hindbrain	1272	C15ORF41 *	chr15:36871812-37102449	16312	Intragenic	0.5
55	**hs1726**	**chr18:49279374-49281480**	**Hindbrain**	**1092**	**NCNRS**	**chr18:49291974-49293480**	**1496**	**10.5**	**0.02**
56	**hs563**	**chr6:98491829-98493238**	**Hindbrain**	**416**	**NCNRS**	**chr6:98467400-98470038**	**2627**	**21.8**	**0.03**
57	hs2144	chr11:19194084-19196536	Hindbrain	2408	ZDHHC13	chr11:19138646-19197969	23897	Intragenic	1
58	hs1086	chr20:39334182-39335059	Hindbrain	877	NCNRS	chr20:39339590-39341077	1485	4.5	0.07
59	hs562	chr10:131106522-131108742	Hindbrain	1666	NCNRS	chr10:131126417-131128098	1440	17.7	
60	hs327	chr1:88926796-88928508	Hindbrain	621	NCNRS	chr1:88916899-88919611	2702	7.2	0.8
61	hs1346	chr21:34465959-34469066	Midbrain/Forebrain	3086	C21orf54	chr21:34537776-34542541	3400	68.7	1
62	hs1391	chr6:3349397-3352257	Midbrain/Forebrain	2808	SLC22A23	chr6:3269196-3457256	14725	Intragenic	1
63	hs1563	chr3:193489359-193491333	Midbrain/Forebrain	1970	OPA1	chr3:193310933-193415612	35411	73.7	1
64	hs1638	chr5:55896173-55899069	Midbrain/Forebrain	2893	NCNRS	chr5:55853222-55856811	3581	39.4	0.7
65	hs1724	chr16:73362809-73364292	Midbrain/Forebrain	1480	NCNRS	chr16:73381335-73383382	2040	17	0.7
66	**hs304**	**chr9:8095553-8096166**	**Midbrain/Forebrain**	**614**	**NCNRS**	**chr9:8107387 - 8108217**	**828**	**11.2**	**0.04**

67	hs1308	chr7:127174386-127177546	Midbrain/Forebrain	3092	NCNRS	chr7:127218813-127222028	3195	41.3	0.3
68	hs1032	chr10:119309200-119310544	Midbrain/Forebrain	971	EMX2	chr10:119301955-119309056	2117	144	0.7
69	hs1545	chr4:109254340-109257033	Midbrain/Forebrain	799	NCNRS	chr4:109280470-109283359	2868	23.4	1
70	hs1571[1]	chr12:114101195-114103805	Midbrain/Forebrain	1730	NCNRS	-	-	-	-
71	hs1577	chr5:91927845-91931024	Midbrain/Forebrain	3153	NCNRS	chr5:91940409-91943762	2808	9.4	0.4
72	hs1723	chr12:103613944-103615320	Midbrain/Forebrain	1363	C12ORF42	chr12:103631369-103889749	7235	16.1	1
73	hs1363	chr18:42942363-42944135	Midbrain/Forebrain	1764	SLC14A2*	chr18:42792960-43263072	25340	Intragenic	1
74	hs1394	chr13:78406128-78407714	Forebrain/Hindbrain	1577	SLAIN1	chr13:78272023-78338377	11830	67.7	0.6
75	hs1568	chr13:28318579-28320134	Forebrain/Hindbrain	588	POLR1D	chr13: 28196003-28241548	2360	77	0.4
76	hs754	chr5:3197865-3198942	Forebrain/Hindbrain	1078	NCNRS	chr5:3175617-3177319	1541	20.5	1
77	hs1060	chr5:92613862-92616844	Forebrain/Hindbrain	2974	NCNRS	chr5:92692093-92696075	3963	75.2	1
78	hs1202	chr1:164604141-164605474	Forebrain/Hindbrain	1328	PBX1*	chr1:164524821-164868533	15725	Intragenic	1
79	hs1385	chr2:3268196-3270849	Midbrain/Hindbrain	2612	TSSC1	chr2:3192696-3381653	12227	Intragenic	0.2
80	hs20	chr16:72738568-72740149	Midbrain/Hindbrain	1569	ZFHX3*	chr16:72816784-73093597	15927	76.6	0.06

81	hs980	chr12:17848111-17849347	Midbrain/Hindbrain	1231	NCNRS	chr12:17885459-17888582	3058	36.1	1
82	hs2064	chr6:52253728-52256212	Midbrain/Hindbrain	2469	PAQR3	chr4:79808281-79860592	9735	Intragenic	1
83	hs737	chr10:130366868-130368005	Midbrain/Hindbrain	1138	NCNRS	chr10:130,368,392-130,370,529	2123	387	1
84	hs1213	chr7:42252831-42254560	Fore/Mid/Hind	1724	GLI3*	chr7:42000548-42277469	20513	Intragenic	0.4
85	hs1573	chr3:147563409-147566604	Fore/Mid/Hind	3188	NCNRS	chr3:147586409-147588804	2358	19.8	0.32
86	hs2223[1]	chr10:79935570-79940095	Fore/Mid/Hind	4488	-	-	-	-	-

Rows highlighted in bold indicate enhancers with signals of positive selection.
*** Proxy coordinates are given for non-coding, non-repetitive sequences (NCNRS) and genes lying within 100kb distance from the enhancer region are obtained for genome build GRCh37/hg19 from UCSC and Ensembl respectively*
** Proxy genes harboring other VISTA elements in their introns*
FOXP1*: hs1231, hs965, hs864, hs865;* ***SOX6****: hs1720, hs883, hs236, hs518 ,hs71 ,hs1301;***CDH2****:hs1634;* ***PBX3****: hs1030, hs818, hs983, hs316, hs1099, hs1095, hs1000, hs317****;*** ***FAF1****: hs1978, hs247,hs194, hs200;* ***ZNF536****:hs384, hs821;***EFNA5****: hs1733, hs1734;***TOX3:*** *hs63, hs62, hs164, hs153****;C15ORF41****: hs828, hs812, hs1871;***SLC14A2****: hs1440, hs1464, hs1363;***PBX1****: hs203, hs1136, hs1144, hs970, hs1235;* ***ZFHX3****: hs16, hs17,hs18, hs19;***GLI3****:hs111, hs1586, hs1213*
1: Suitable proxy regions could not be found

7.3 Appendix -III: Hominin specific transcription factor binding sites in positively selected enhancers

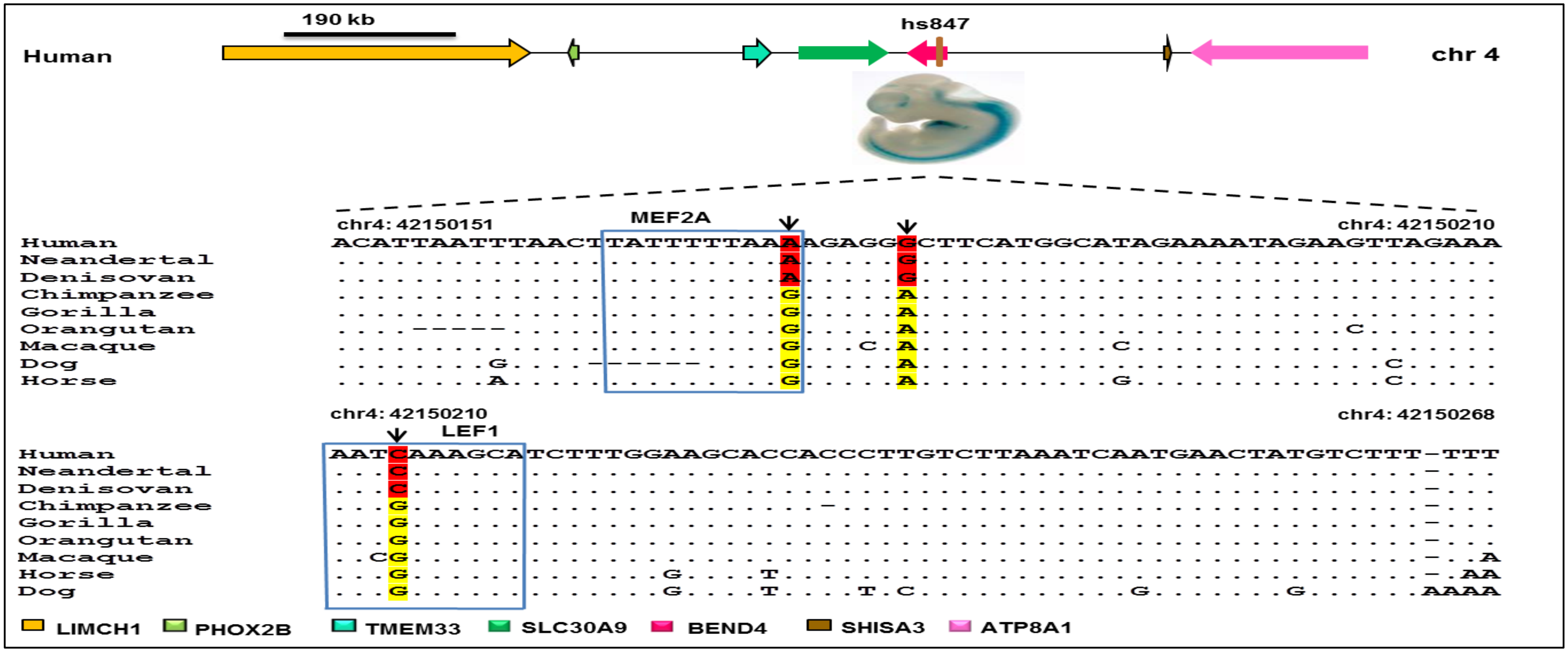


Figure A1. Hominin shared binding motifs of MEF2A and LEF1 in Forebrain exclusive VISTA enhancer hs847

Multiple sequence alignment with orthologous sequences from non-human primates and older mammalian species depicts ancestral site conservation till horse. Transcription factor (TF) MEF2A's hominin shared binding motif TATTTTTAAA is preceded by TATTTTTAAG* in non-human primates and older mammals (representative species shown in the figure). Similarly, transcription factor LEF1's binding site AATC*AAAGCA is the shared unique site among hominins is preceded by AATG*AAAGCA in non-human primates and older mammals.*

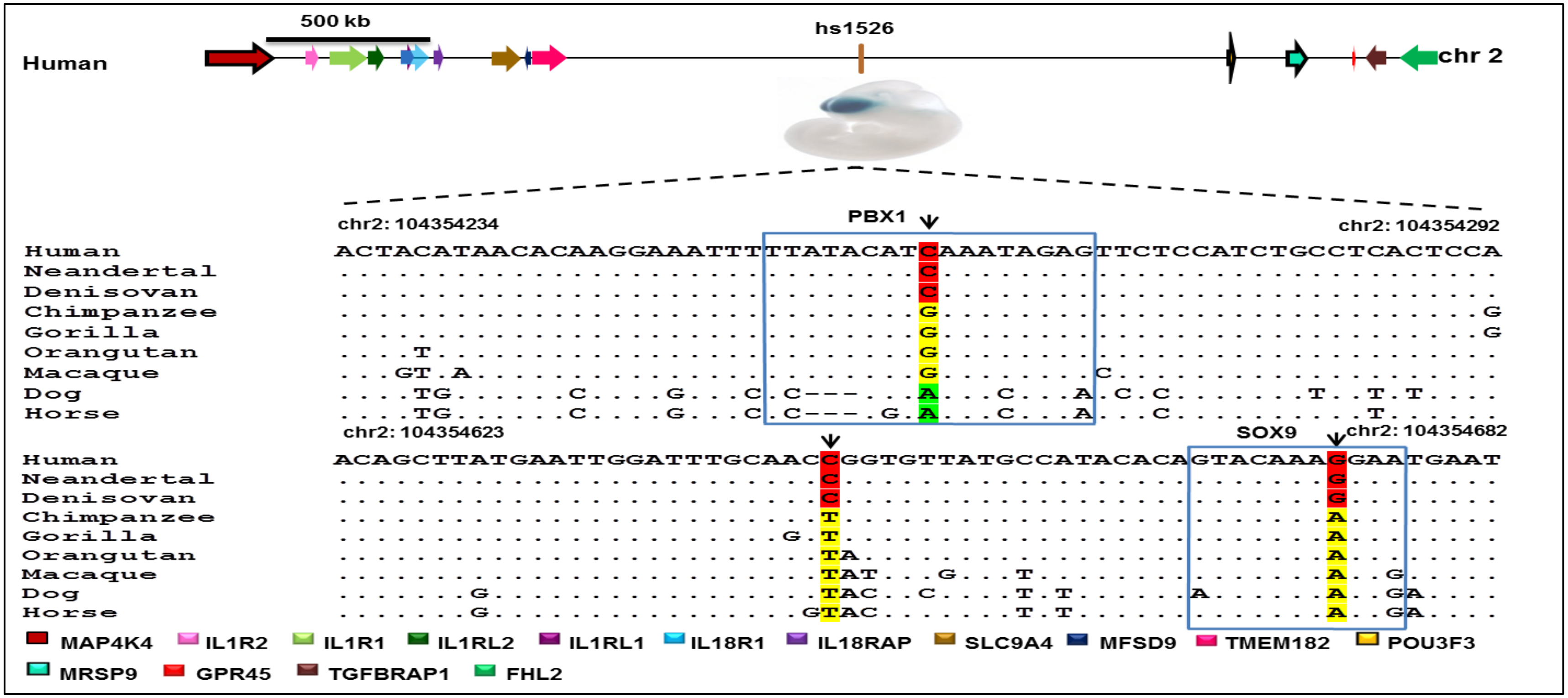


Figure A2. Hominin shared TFBSs of PBX1 and SOX9 in Forebrain exclusive VISTA enhancer hs1526

*The above figure depicts forebrain exclusive VISTA enhancer hs1526 carrying two hominin shared unique transcription factor binding sites (TFBSs) of transcription factors PBX1 and SOX9. Orthologous sequences from non-human primates and older mammalian species show ancestral sites conservation till horse for both TFs. Transcription factor PBX1 has TTATACATC*AAATAGAG as the hominin shared motif to be preceded by TTATACATG*AAATAGAG in non-human primates and TTATACATA*AAATAGAG in dog and horse. Similarly, transcription factor SOX9's TFBS GTACAAAG*GAA is the shared unique binding motif among hominins is preceded by GTACAAAA*GAA in non-human primates and older mammals.*

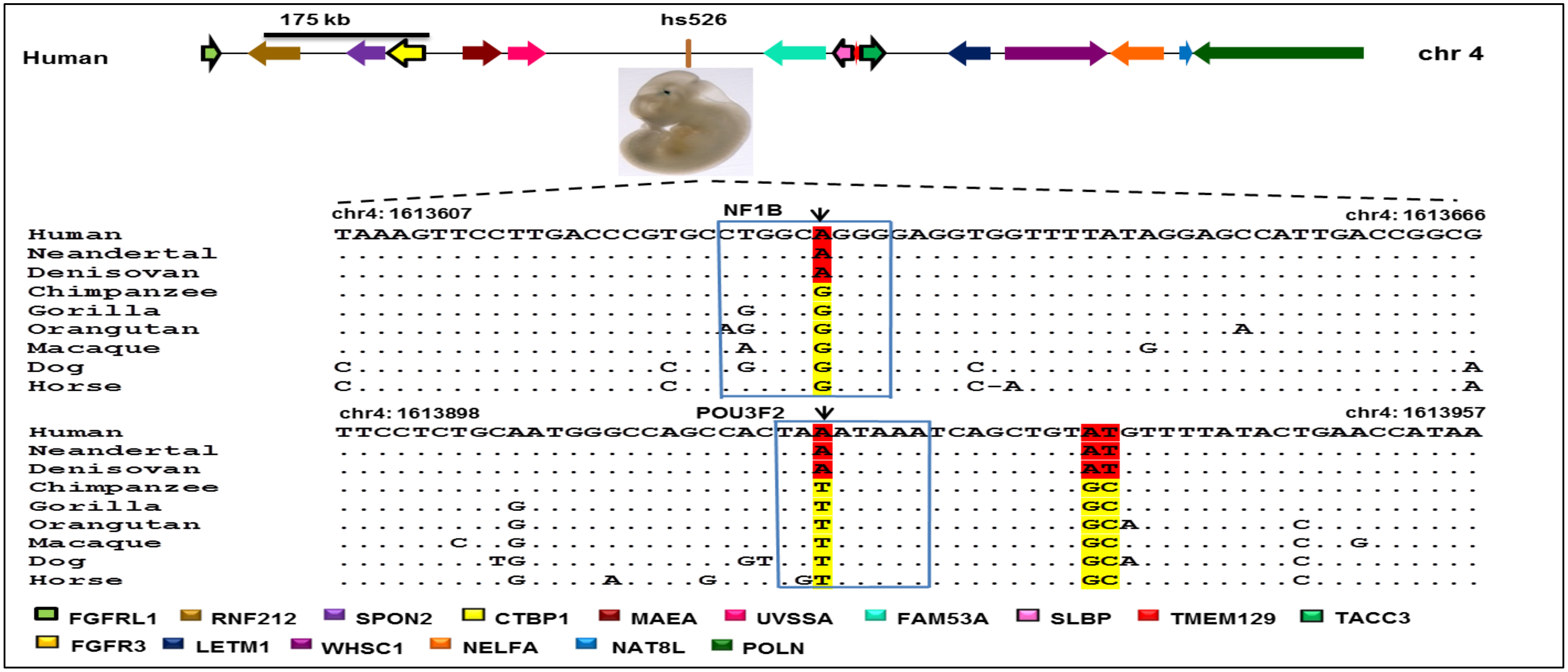
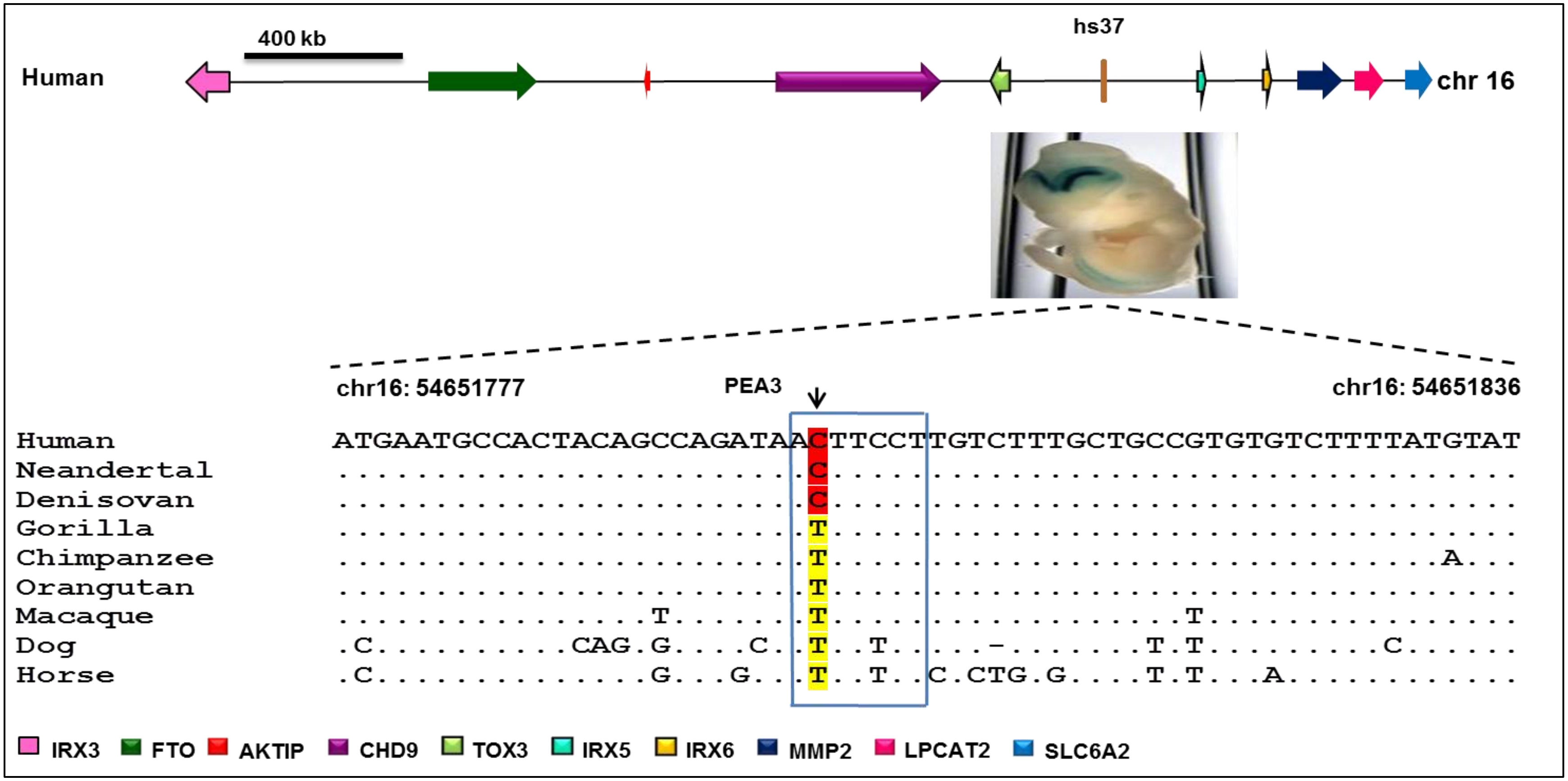


Figure A3. Hominin shared TFBSs of NF1B and POU3F2 in Forebrain exclusive VISTA enhancer hs526

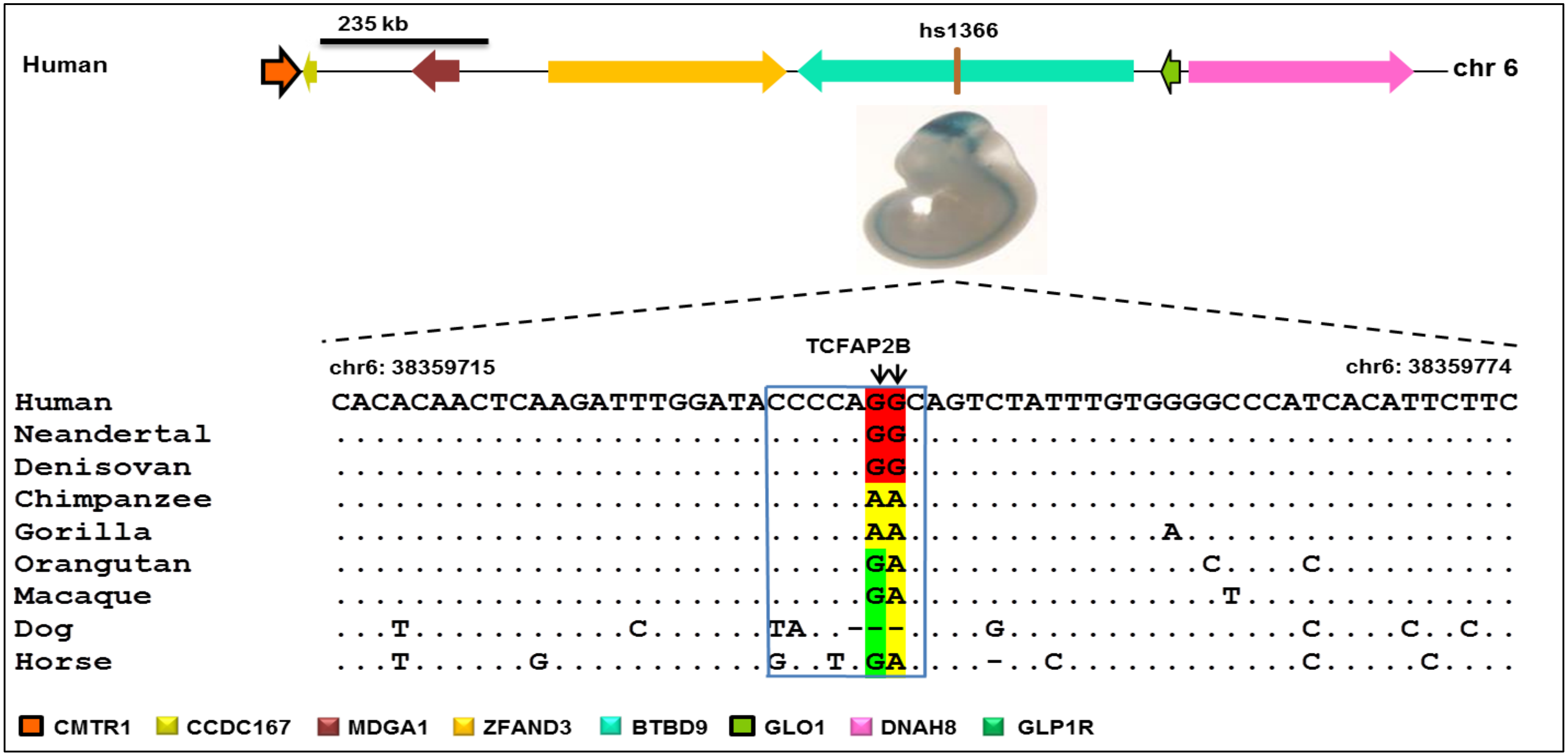
*The figure explains two hominin shared unique TFBSs of TFs POU3F2 and NF1B inhabiting forebrain exclusive VISTA enhancer hs526. Orthologous sequences from non-human primate and older mammalian species show ancestral site conservation till horse for both TFs. Transcription factor NF1B has CTGGCA*GGG as the hominin shared motif to be preceded by CTGGCG*GGG in non-human primates and older mammals (representative species shown in the figure). Similarly, transcription factor POU3F2's modified TFBS TAA*ATAAA is the shared unique site among hominins to be preceded by ancestral motif TAT*ATAAA in non-human primates and older mammals*



```
Human       ATGAATGCCACTACAGCCAGATAACTTCCTTGTCTTTGCTGCCGTGTGTCTTTTATGTAT
Neandertal  ........................C...................................
Denisovan   ........................C...................................
Gorilla     ........................T...................................
Chimpanzee  ........................T...............................A...
Orangutan   ........................T...................................
Macaque     ................T.......T..................T................
Dog         .C..........CAG.G....C..T..T.....-.......T.T.........C......
Horse       .C..............G...G...T..T..C.CTG.G....T.T...A............
```

Figure A4. Hominin shared PEA3 TFBS in Forebrain exclusive VISTA enhancer hs37

*The figure narrates for transcription factor PEA3 the unique hominin shared TFBS in the forebrain exclusive VISTA enhancer hs37. For the representative orthologous species in the figure, the newly arisen site among hominins is AC*TTCCT whereas the ancestral site is AT*TTCCT among non-human primates and older mammals.*



```
Human       CACACAACTCAAGATTTGGATACCCCAGGCAGTCTATTTGTGGGGCCCATCACATTCTTC
Neandertal  ...........................GG...............................
Denisovan   ...........................GG...............................
Chimpanzee  ...........................AA...............................
Gorilla     ...........................AA.............A.................
Orangutan   ...........................GA...............C....C..........
Macaque     ...........................GA................T..............
Dog         ...T...........C......TA..---....G...............C....C..C..
Horse       ...T......G...........G..T.GA....-..C............C.....C....
```

Figure A5. Hominin shared TCFAP2B TFBS in Midbrain exclusive VISTA enhancer hs1366

*Figure A5 represents for transcription factor TCFAP2B the unique hominin shared TFBS in the midbrain exclusive VISTA enhancer hs1366. For the representative orthologous species in the figure, the newly arisen site among hominins is CCCCAGG*C whereas the ancestral site is CCCCAGA*C among non-human primates and older mammals.*

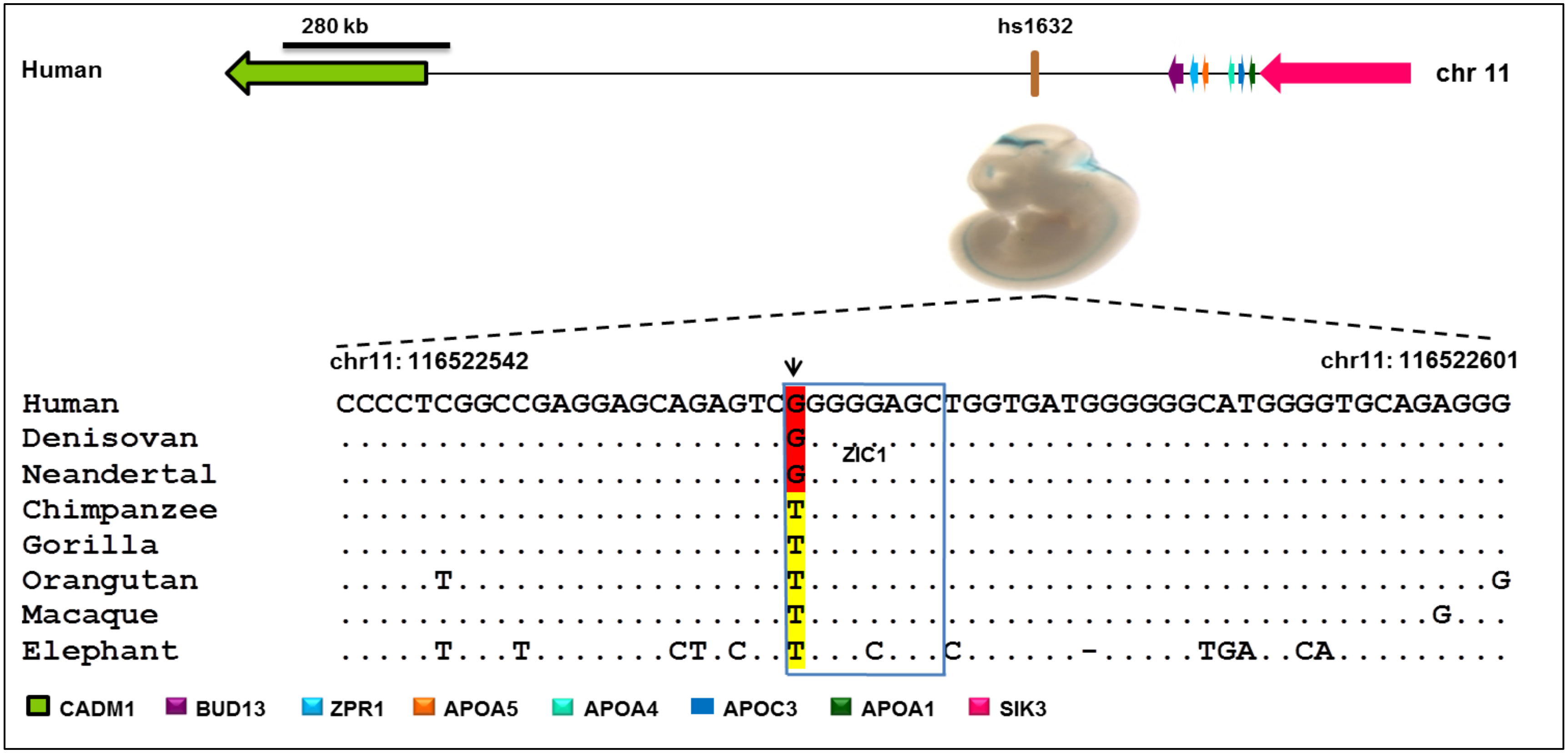

280 kb
hs1632
Human
chr 11
chr11: 116522542
chr11: 116522601
Human CCCCTCGGCCGAGGAGCAGAGTCGGGGGAGCTGGTGATGGGGGGCATGGGGTGCAGAGGG
Denisovan ....G....................................
Neandertal ....G....................................
Chimpanzee ....T....................................
Gorilla ....T....................................
Orangutan .....T.................T...................................G
Macaque .......................T................................G...
Elephant .....T...T.......CT.C..T...C...C......-.....TGA..CA.........
ZIC1
🟩 CADM1
🟪 BUD13
🟦 ZPR1
🟧 APOA5
🟦 APOA4
🟦 APOC3
🟩 APOA1
🟥 SIK3

Figure A6. Hominin shared ZIC1 TFBS in Midbrain exclusive VISTA enhancer hs1632

*Figure A6 represents for transcription factor ZIC1 the unique hominin shared TFBS in the midbrain exclusive VISTA enhancer hs1632. Orthologous sequences of horse and hog are not included because of poor sequence conservation. Instead, orthologous elephant sequence is added. The figure narrates newly formed site for ZIC1 among hominins is G*GGGGAGC whereas the ancestral site is T*GGGGAGC among non-human primates and older mammals.*

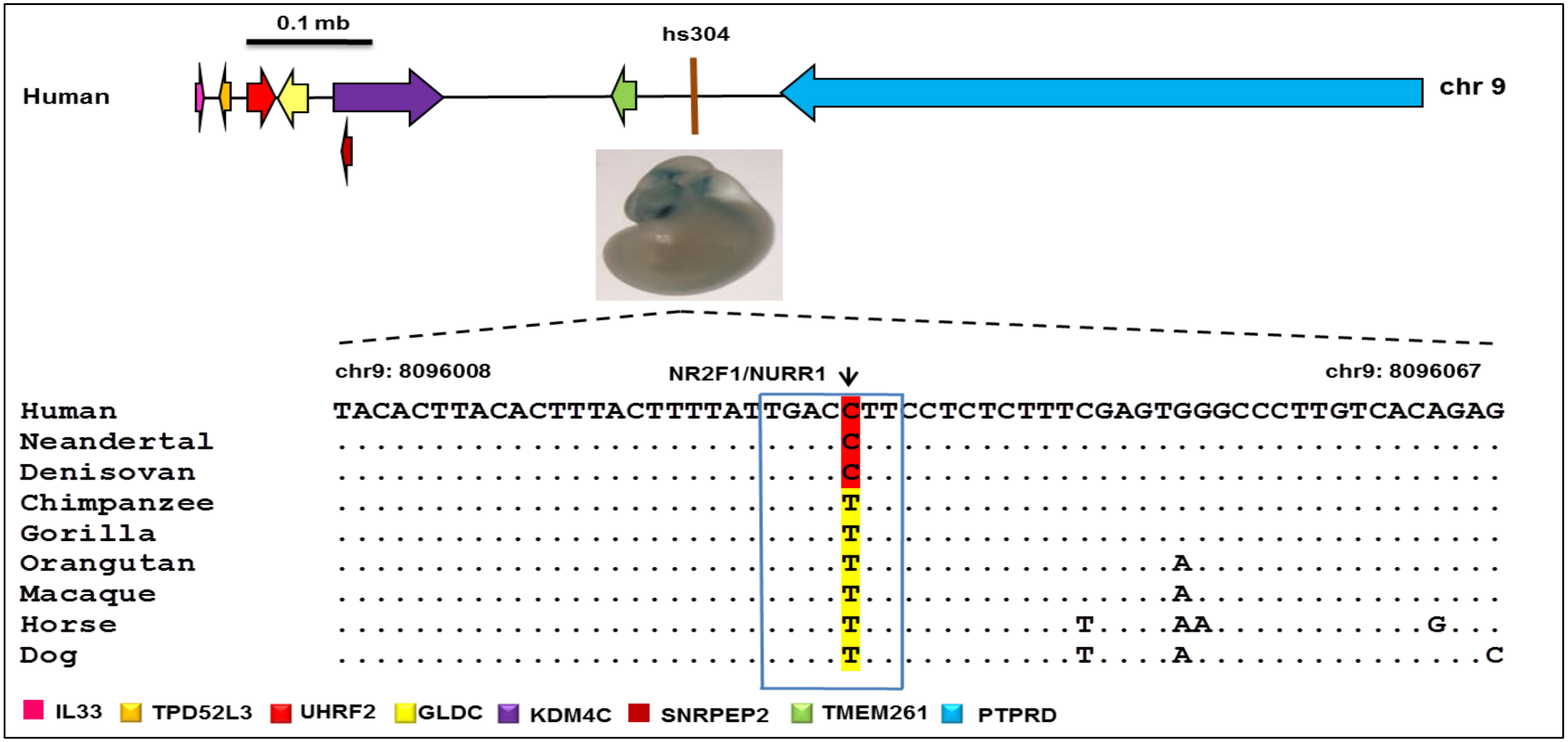

0.1 mb
hs304
Human
chr 9
chr9: 8096008
NR2F1/NURR1
chr9: 8096067
Human TACACTTACACTTTACTTTTATTGACCTTCCTCTCTTTCGAGTGGGCCCTTGTCACAGAG
Neandertal ....C.................................
Denisovan ....C.................................
Chimpanzee ....T.................................
Gorilla ....T.................................
Orangutan ..........................T................A................
Macaque ..........................T................A................
Horse ..........................T...........T....AA...........G...
Dog ..........................T...........T....A...............C
IL33
TPD52L3
UHRF2
GLDC
KDM4C
SNRPEP2
TMEM261
PTPRD

Figure A7. Hominin shared NR2F1/NURR2 TFBSs in Midbrain/Forebrain exclusive VISTA enhancer hs304

*Figure A7 represents for transcription factors NR2F1 and NURR1 the two overlapping unique hominin shared TFBSs in the midbrain/forebrain exclusive VISTA enhancer hs304. For the representative orthologous species in the figure, the newly arisen site among hominins is TGACC*TT whereas the ancestral site is TGACT*TT among non-human primates and older mammals.*