# Molecular analysis of the mechanisms involved in *THBS4* differential gene-expression in the human brain.

Análisis molecular de los mecanismos implicados en las diferencias de expresión del gen *THBS4* en el cerebro humano.

DOCTORAL THESIS

Raquel Rubio Acero

Universitat Autònoma de Barcelona
Facultat de Biociències
Departament de Genètica i de Microbiologia
Bellaterra, 2013

Memoria presentada por la Licenciada en Biología

Raquel Rubio Acero para optar al grado de Doctora en Biología.

Raquel Rubio Acero

Bellaterra, 16 de octubre de 2013

El Doctor Mario Cáceres Aguilar, Profesor de Investigación ICREA del Institut de Biotecnologia i de Biomedicina de la Universitat Autònoma de Barcelona,

CERTIFICA que Raquel Rubio Acero ha llevado a cabo bajo su dirección el trabajo de investigación realizado en el Departamento de Genética y de Microbiología de la Facultad de Biociencias de la Universitat Autònoma de Barcelona y que ha dado lugar a la elaboración de esta Tesis Doctoral titulada "Molecular analysis of the mechanisms involved in *THBS4* differential gene-expression in the human brain".

Y para que conste a los efectos oportunos, firma el presente comunicado en Bellaterra, a 16 de octubre de 2013.

Dr. Mario Cáceres Aguilar

# TABLE OF CONTENTS

A mi hermano, por considerarme tu mejor regalo.

To my brother, for considering me your best gift.

# ABSTRACT

The last decades have seen a growing interest in what makes us humans and how the human brain differs from that of our closest relatives at the molecular level. Hundreds of genes with expression differences between human and non-human primates have been identified. However, it is important to study these genes in more detail to see if they are really involved in human brain characteristics. Thrombospondins are multimeric extracellular glycoproteins that modulate cell-cell and extracellular matrix interactions and have been implicated in synaptogenesis. Within the thrombospondin family, thrombospondin-2 (*THBS2)* and thrombospondin-4 (*THBS4)* show, respectively, a ~2-fold and ~6-fold upregulation in human cerebral cortex compared to chimpanzees and macaques. To analyze the causes of these expression differences, we have carried out a comparative and functional analysis of the *THBS4* promoter region in humans and chimpanzees. We have identified and validated an alternative transcription start site (TSS) for *THBS4* that is located ~44 kb upstream from the known TSS and generates a new mRNA isoform that might have appeared later that the reference one during evolution. To compare expression levels of both mRNAs, we performed quantitative RT-PCR in different human tissues and cortical regions of 11 humans, 11 chimpanzees and 8 macaques. Interestingly, the new isoform of *THBS4* is expressed mainly in brain tissues. Moreover, expression differences between human and non-human primate cortex for this alternative isoform are consistent with those shown for total *THBS4* expression. Increased *THBS4* expression in the human brain therefore appears to be related to higher transcription from the alternative promoter. To evaluate the activity of both *THBS4* promoter sequences we performed reporter assays from humans and chimpanzees in different human cell lines. We have found significant differences between both promoters, but not between species, in the neuroblastoma cell lines assayed. This result is consistent with the search for transcription factor binding sites (TFBS) in the alternative promoter region, which only detected three putative TFBS differentially predicted between both species. We also compared the DNA methylation in a CpG island upstream the new isoform in 5 humans and 5 chimpanzees

detecting similar low methylation levels in all of them. Based in the computational predictions available online and ChIP-Seq pilot experiments, we searched for a putative enhancer region controlling the *THBS4* alternative promoter without finding any reliable candidate. Finally, in humans, *THBS4* has been associated to two different haplotypes that are maintained by balancing selection, but experimental analysis did not show any effect of the genotype over *THBS4* gene expression. Although we have not been able to identify the ultimate cause of the increased *THBS4* levels in humans, based on all our results we suggest that the differential gene expression might be related to a brain-specific enhancer sequence that has so far escaped our scrutiny. Understanding its regulation could be relevant to the functional consequences of *THBS4* expression differences during human brain evolution and ultimately could give us clues of how we became humans.

# RESUMEN

Durante las últimas décadas ha crecido el interés en cuestiones como qué nos hace humanos o cómo difiere a nivel molecular el cerebro humano del de nuestros parientes más cercanos. Se han podido identificar cientos de genes con diferencias de expresión entre el ser humano y otros primates no humanos. Sin embargo, es importante estudiar más en detalle estos genes para comprobar si realmente están involucrados en las características específicas de nuestro cerebro. Las trombospondinas son glicoproteínas extracelulares multiméricas que modulan las interacciones entre células y con la matriz extracelular y que se han implicado en la sinaptogénesis. Dentro de la familia de las trombospondinas, los genes de la trombospondina-2 (*THBS2*) y trombospondina-4 (*THBS4*) se expresan, respectivamente, alrededor de 2 y 6 veces más en la corteza cerebral humana en comparación con la de chimpancés o macacos. Para conocer las causas de estas diferencias de expresión, hemos llevado a cabo un análisis comparativo y funcional de la región promotora de *THBS4* en humanos y en chimpancés. Hemos identificado y validado un sitio de inicio de transcripción (TSS) alternativo para *THBS4* que se encuentra 44 kb aguas arriba del TSS de referencia y que genera una nueva isoforma de ARNm que podría haber aparecido más tarde durante la evolución. Para comparar los niveles de expresión de ambos transcritos se realizó RT-PCR cuantitativa en diferentes tejidos humanos y en muestras de corteza cerebral de 11 humanos, 11 chimpancés y 8 macacos. Curiosamente, la nueva isoforma de *THBS4* se expresa principalmente en los tejidos cerebrales. Por otra parte, la diferencia de expresión entre humanos y primates no humanos para la isoforma alternativa son consistentes con la encontrada al analizar la expresión total de *THBS4*. Por tanto, el aumento de la expresión de *THBS4* en el cerebro humano parece estar relacionado con una mayor transcripción a partir del promotor alternativo. Para evaluar la actividad de ambas secuencias promotoras en humanos y en chimpancés, hemos llevado a cabo ensayos con un gen indicador en diferentes líneas celulares humanas. Se han encontrado diferencias significativas entre los dos promotores, aunque en las líneas de neuroblastoma que utilizamos no existen diferencias significativas entre especies. Este

resultado es consistente con la búsqueda de sitios de unión de factores de transcripción en la región del promotor alternativo, ya que sólo se detectaron tres posibles sitios de unión diferentes entre ambas especies. Se ha comparado también la metilación del ADN en una isla CpG situada aguas arriba de la nueva isoforma de *THBS4* en 5 humanos y en 5 chimpancés, detectando niveles bajos de metilación en ambas especies. Basándonos en las predicciones informáticas disponibles y en experimentos piloto de ChIP-Seq, hemos buscado posibles enhancers (secuencias potenciadoras) que estén controlando el promotor alternativo de *THBS4,* pero no hemos encontrado ningún candidato fiable. Por último, en humanos, se ha visto que hay dos haplotipos de *THBS4* diferentes que se encuentran mantenidos mediante selección equilibradora. Sin embargo, la comparación experimental de ambos no muestra ningún efecto del genotipo sobre la expresión de *THBS4.* Aunque no se ha conseguido identificar la causa concreta del incremento en los niveles de *THBS4* en humanos, en base a nuestros resultados sugerimos que la expresión diferencial del gen podría estar relacionada con un enhancer específico del cerebro que no hemos conseguido localizar. Comprender como se encuentra regulado este posible enhancer podría ser relevante para entender cuales son las consecuencias funcionales de las diferencias de expresión de *THBS4* sobre la evolución del cerebro y, en última instancia, darnos pistas sobre como nos convertimos en humanos.

# 1

# INTRODUCTION

*"Just play. Have fun. Enjoy the game."*

– MICHAEL JORDAN –

# 1

# INTRODUCTION

## 1.1. What makes us humans?

Can we imagine if more than one species of human were walking around today? In the animal kingdom it is really common to find tens or even hundreds of species from the same genus. It may seem shocking, but if there are around a hundred of species of honeybees, at least twenty kinds of *Xenopus*, or seven living *Canis*, why are there not different living species of humans, like Neanderthal or Denisovans? Another question that has intrigued people in general and scientists particularly for many years is what distinguishes us from the chimpanzee, our closest primate relatives, or what makes human beings unique? These questions bring an apparently endless list of attributes and abilities to mind, both positive and negative, which have evolved in the human lineage after it separated from the common lineage leading also to chimpanzees and bonobos (TABLE 1). Bipedalism, relative brain size and brain topology, self-awareness, complex speech and symbolic cognition, opposable thumbs, hairless sweaty skin and increased longevity are just a few of the traits that distinguish us from other species. Humans also have a disease profile that differs from that of other primates, maybe caused by the fact that humans have the longest lifespan potential of any primate. In particular, we seem especially vulnerable to neurodegenerative diseases and cancer.

To begin with the study of human specializations, it seems reasonable to search for characteristics of the brain, behavior and cognition, as most people would regard humans as being highly specialized in these domains. In the beginning, it was thought that chimpanzees do not have many cognitive functions that are found in humans, including altruism, understanding another individual's cognitive state, social cooperation, use of tools, or cultural transmission. However, several studies in non-human primates have

subsequently described these abilities (WHITEN *et al.* 1999, PRUETZ and BERTOLANI 2007, WHITEN *et al.* 2009, HORNER *et al.* 2011, ROTH and DICKE 2012), which has reopened the debate about the uniqueness of human cognitive traits (BOESCH 2007, WHITEN and ERDAL 2012).

TABLE 1. Some phenotypic human traits. A given "difference" listed here could be a suggested gain or loss in humans, with respect to the great apes. Table redrawn from VARKI and ALTHEIDE (2005).

**LIFE HISTORY**
Helplessness of the newborn
Extended care of young
Adolescence
Longevity

**REPRODUCTIVITY**
Female pituitary menopause
Placentophagy
Baculum (Penis Bone)
Copulatory plug
Sperm count
Concealed ovulation

**PARTURITION**
Cephalo-pelvic disproportion
Duration of labor
Need for assistance with childbirth.
Pain during childbirth
Umbilical Cord length

**POSTNATAL DEVELOPMENT**
Late closure of cranial sutures
Duration of infant arousal
Maternal-infant eye-to-eye gaze

**ANATOMY**
Sagittal crest of skull
Brow ridge
Length of sphenoid sinus
Age of pelvic bone fusion
Bone cortex thickness
Laryngeal position
Ear lobes
Sexual body size dimorphism
Visible white of the eyes
Lacrimal gland structure

**NUTRITION**
Frugivory
Carnivory
Aquatic foods
Underground foods
Cooking

**BIOMECHANICS**
Bipedal gait
Adductive thumb
Skeletal muscle strength
Hand-eye coordination

**ORGAN PHYSIOLOGY**
Ability for sustained running
Voluntary control of breathing
Ability to dive underwater
Ability to float/swim
Emotional lacrimation
Olfactory sense
Diving reflex

**ANATOMIC PATHOLOGY**
Cortical neurofibrillary tangles

**CLINICAL PATHOLOGY**
Serum vitamin B12/B12 binding
Total leukocyte count
Absolute neutrophil count
Absolute Lymphocyte count

**SKIN BIOLOGY AND DISEASE**
Eyebrows
Eccrine Sweat Glands
Acne Vulgaris

**MENTAL DISEASE**
Schizophrenia
Bipolar psychosis
Autism
Suicide

**MEDICAL DISEASES**
HIV Progression to AIDS
*P. falciparum* malaria
Viral Hepatitis B/C
Complications Influenza A
Incidence of carcinomas
Hemorrhoids
Infectious sexually transmitted diseases

**NEUROANATOMY**
Relative brain size
Direct cortical projections
Relative volume of frontal cortex
Relative volume of corpus callosum
Relative volume of cerebellum
Rate of postnatal brain growth

**NEUROBIOLOGY**
Postnatal dendritic growth
Postnatal synapse formation
Cortical synapse density
Cortical neuron density
Dendrites per neuron
Synapses per neuron
Adult neurogenesis

**COMMUNICATION**
Gestural communication
Symbolic communication
Semantics
Grammar and syntax
Writing

**COGNITIVE CAPACITY**
Declarative memory
Imitative learning
Symbolic representation
Awareness of death
Awareness of the past/future
Theory of mind
Numeracy

**BEHAVIOR**
Control of facial expressions
Planning ahead
Intentional deception
Mechanical multi-tasking
Symbolic play
Use of containers
Control of fire
Food preparation
Domestication of animals/plants
Somnambulism

**CULTURE**
Composition of art/music/rhythms
Death rituals
Clothing
Competitive sports
Religion
Body adornment
Measurement of time
Weapons
Toys

### 1.1.1. The human brain and its evolution.

Studies on the differences between humans and non-human primates have provided a lot more information about human genetic specializations than apparent phenotypic human-specific characteristics of the brain. The main reason has been the lack of technical methods to explore the human brain with the same level of detail as can be done in other non-human species until the last decades. For non-human species, there are powerful invasive techniques available. However, these are considered unethical in humans, chimpanzees and rare and endangered species, and have to be replaced for noninvasive imaging techniques, such as magnetic resonance imaging (MRI) (DUONG 2010), positron emission tomography (PET) (RILLING *et al.* 2007), and most recently diffusion-tensor imaging (DTI) (ZHANG *et al.* 2012).

Humans are distinguished from other mammals by a much larger brain than expected for a primate of our body size. The range of body size of the great apes overlaps human body size, but humans have a far bigger brain: around 1300 grams compared to the 350-500 grams of great apes (PREUSS 2004). Initially, there was a standing notion that the frontal cortex, and particularly the prefrontal subregion, was dramatically and disproportionately large in humans (frequently citing the classic work of BRODMANN (1912)). This claim was based on the implication of the frontal cortex in the elaboration of cognitive capacities, which has culminated in the evolution of language and a complex social behavior. However, after the application of structural neuroimaging techniques, this concept has been called into question because it seems that, even though the human brain is larger, it is not disproportionately so (SEMENDEFERI *et al.* 2002, HERCULANO-HOUZEL 2012). Given that, are these mutually exclusive proposals? Was the size expansion global, affecting all (or nearly all) areas equally, or was there differential expansion of particular regions? PREUSS (2004) gathered data from several neuroanatomical studies on the absolute sizes of individual cortical areas of great apes and humans. According to his conclusions, the enlargement of the brain during the human evolution likely involved primarily the disproportionate expansion of the frontal cortex, but also other specifically important functional regions of

the parietal, temporal and occipital lobes, resulting in a relatively similar proportion of lobe volumes as compared to apes (PREUSS 2004).

Within the human brain, two areas in particular have received a greater attention when trying to understand the production of speech: Brodmann's Area 44 and the temporal planum (FIGURE 1). The former delineates parts of Broca's Area within the inferior frontal lobe of the neocortex. In right-handed individuals, this region is larger in the left hemisphere of the brain than in the right, an asymmetry that has been correlated with language ability, particularly in motor aspects of speech such as articulation and fluency (LURIA 1970, FOUNDAS *et al.* 1998). The second area of interest, the temporal planum, is the portion of the superior temporal lobe commonly identified with Wernicke's Area. This site is implicated in human spoken and gestural communication, which also shows dominance towards the left-hemisphere in right-handed individuals (GESCHWIND and LEVITSKY 1968). From magnetic resonance imaging technology, similar left–right asymmetry like that in humans has been found on a population basis in chimpanzees, bonobos and gorillas, indicating that the neuroanatomical substrate of left-hemisphere dominance in speech production preceded the origin of hominins (CANTALUPO and HOPKINS 2001). However, whereas an asymmetrical temporal planum pattern has been confirmed in chimpanzees (GANNON *et al.* 1998, HOPKINS *et al.* 1998), asymmetry of the Broca's Area in our closest relatives has been questioned, suggesting that Broca's Area in the left hemisphere expanded in relative size during human evolution, possibly as an adaptation for human language abilities (SCHENKER *et al.* 2010).

From a metabolic and histological point of view, only few studies have explored differences in brain cellular organization and neural connectivity between humans and apes so far. In particular, it seems that the expansion of the human brain entailed a high metabolic cost (BULLMORE and SPORNS 2012). Pyramidal neurons of humans cortex are longer and display much larger dendritic arborizations and more numerous spines than those in chimpanzees or other primates with smaller brains, contributing to the higher cost of neuronal activity (ELSTON 2003, ELSTON *et al.* 2006, BIANCHI *et al.* 2012). One way to measure this metabolic cost was based on the examination of the glia-to-neurons ratio

(SHERWOOD *et al.* 2006). Due to the crucial role of glia cells in the nurturing of neurons and emission of neurotrophic messengers associated with neural activity (TSACOPOULOS and MAGISTRETTI 1996, BARRES and RAFF 1999), their density in the normal brain provides an indication of the metabolic demand of neighboring neurons. SHERWOOD *et al.* (2006) suggested that the human frontal cortex displays a 46% higher glia-to-neuron ratio compared to other anthropoid primates, which could resemble a consequence of the elevated metabolic costs of maintaining membrane potentials in the neurons situated within an enlarged brain in the human lineage.

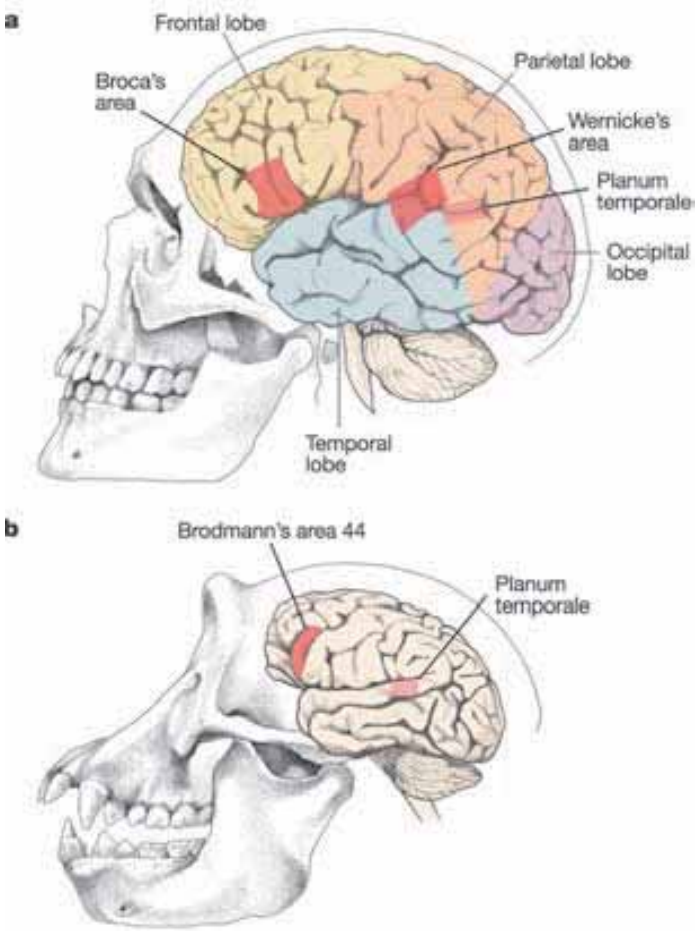



FIGURE 1. Comparative neuroanatomy of humans and chimpanzees. Lateral views of the left hemispheres of a modern human (a) and a chimpanzee (b) brain. Equivalent regions associated with communication and production of speech are shown: Brodmann's Area 44 (Broca's Area) and around the temporal planum ("Planum temporale"; Wernicke's Area), Figure extracted from CARROLL (2003).

### 1.1.2. Anatomical modifications and diseases.

In general, non-human primates are very similar to humans. However, from an anatomical point of view, there are certain differences that may be behind some common human diseases and the study of these changes in non-human primates represents a great opportunity to find an explanation. Probably, one of the most characteristic human anatomical features is the bipedal upright posture. It results in the cervical, thoracic and lumbar vertebrae of the human skeleton having developed ventral curvatures, which are not present in non-human primates. This upright walking imposes stress and strains to the spine and back muscles contributing to possible causes for the chronic back pain suffered by many humans (VARKI *et al.* 2011). The upright posture also produced a narrowing of the pelvic outlet. This explains the difficulty of passing a newborn through the maternal pelvic canal and in part the long labor of humans compared to chimpanzees (TREVATHAN 1996). Other common human-specific diseases like sinusitis occur because humans, unlike non-human primates who do not experience this ailment, have four air-containing sinus spaces in their skulls lined by epithelium that secretes mucus and swells during inflammation, thus blocking the exit channels for mucus (MARPLE *et al.* 2009).

Additionally, there are other, and probably more intriguing, biomedical differences that cannot be explained by anatomical differences. One of the best known differences is the predisposition for Alzheimer's disease (AD). This is a progressive neurodegenerative disorder the early stage of which is characterized by synaptic loss and impairment of episodic memory (BORLIKOVA *et al.* 2013). With age, all non-human primates analyzed to date develop senile (Aβ) plaques and cerebral β-amyloid angiopathy (HEUER *et al.* 2012). However, significant tauopathy is unusual in simians (HEUER *et al.* 2012) and Alzheimer's disease appears to be a uniquely human condition, which is possibly attributable to our expanded longevity and peculiar capacity for episodic memory (REID and EVANS 2013).

Another surprising difference appears to be the significantly higher incidence of the most common human cancers, such as carcinomas of the breast, ovary, lung, stomach,

colon, pancreas, and prostate (SEIBOLD and WOLF 1973, PUENTE *et al.* 2006). According to the last report of the International Agency for Research of Cancer (www.iarc.fr), these cancers caused around 13% of deaths in the human population in 2008 (FERLAY *et al.* 2010), whereas the incidence of these cancers in the non-human primates has been rated to about 2%–4% and seems to be even lower in the great apes (MCCLURE 1973, SEIBOLD and WOLF 1973, BENIASHVILI 1989, SCOTT 1992). In cancer, the apoptotic machinery is disrupted, resulting in tumor initiation, progression and metastasis (LOWE and LIN 2000). It has been shown that human cells display a significant reduction of the apoptotic function relative to the chimpanzee and macaque cells, which could suggests that it may contribute to the increased propensity of humans to develop cancer (ARORA *et al.* 2012). Additionally, it has been suggested that selection for increased cognitive ability and size in human brains, may have coincidently resulted in an increased risk for cancer and other diseases associated with reduced apoptotic function (ARORA *et al.* 2009).

### 1.1.3. Spot the difference: from phenotype to genes.

The very beginning in the study of modern human molecular evolution can be attributed to KING and WILSON (1975). They hypothesized that the difference in gene regulation, rather than changes of protein–coding sequences, is the major evolutionary force underlying the divergence of humans and chimpanzees. This idea later received additional support (GOULD 1977), emphasizing the possibility that a small number of expression changes acting early in development could have profound phenotypic consequences. A big step in human molecular evolution analysis came with the sequencing of the whole human genome (VENTER *et al.* 2001, CONSORTIUM 2004) and chimpanzee genome (CONSORTIUM 2005). From there on, it was possible to have a more or less reliable catalogue of the genetic changes and to estimate the differences between both species in around 1-4%, depending on whether only single-nucleotide divergence or also insertion and deletion (indel) events are counted (VARKI and ALTHEIDE 2005, POLAVARAPU *et al.* 2011).

The biggest challenge is then to determine which of these genetic changes are related to phenotypic characteristics.

As reviewed by O'BLENESS and colleagues, several factors, such as imprecise annotations or relying only on few sequenced individuals, can potentially impede accurate comparisons of genes and genomic sequences, making the identification of human genomic changes difficult (O'BLENESS *et al.* 2012). During the last years the number of human specific traits that have been discovered has rapidly increased; all of which –from the smallest change in the DNA sequence to complex structural variations– have been listed at the Center for Academic Research and Training in Anthropogeny webpage (CARTA, http://carta.anthropogeny.org). In addition, the Matrix of Comparative Anthropogeny (MOCA, http://carta.anthropogeny.org/moca/domains) permits to link the genetic and genomic changes to many other features of human uniqueness, in domains ranging from molecules to culture.

#### 1.1.3.1. Large-scale cytogenetic changes.

Large-scale genomic differences between humans and chimpanzees were noticed when the karyotypes of both species were analyzed by chromatin-stained banding techniques (YUNIS and PRAKASH 1982). The three most relevant changes specific to humans determined were: a fusion of two ancestral ape chromosomes, producing a variation in the haploid number (from 24 in apes to 23 in humans) and resulting in the large human chromosome 2; the addition of human-specific constitutive heterochromatic C bands on chromosomes 1, 9, 16 and Y; and the discovery of human-specific pericentric inversions on chromosomes 1 and 18 (YUNIS and PRAKASH 1982). More recently, it has been determined that in the genomic regions of some of these cytogenetic changes there are more than just human-specific modifications of the heterochromatin; these regions are frequently adjacent to regions which are greatly enriched in evolutionarily recent gene duplications (FORTNA *et al.* 2004, DUMAS *et al.* 2007). For example, the 1q21.1 region of the genome adjacent to the human-specific 1q12 C band and located within the chromosome 1 pericentric inversion specific for the human lineage, has undergone numerous copy number variations within the

region, and has been implicated in multiple recurrent human developmental and neurogenetic diseases (DUMAS and SIKELA 2009). These include microcephaly and macrocephaly (BRUNETTI-PIERRI *et al.* 2008), autism (AUTISM GENOME PROJECT *et al.* 2007, PINTO *et al.* 2010, CRESPI and CROFTS 2012), schizophrenia (INTERNATIONAL SCHIZOPHRENIA 2008, CRESPI and CROFTS 2012), mental retardation (MEFFORD *et al.* 2008) and neuroblastoma cancer (DISKIN *et al.* 2009).

### 1.1.3.2. Structural changes.

Alteration of the gene structure provides a major mechanism through which evolutionarily adaptive changes can be introduced (O'BLENESS *et al.* 2012). There are many genetic structural changes between several kb to Mb in size that have occurred between humans and non-human primates, such as copy number variants (PERRY *et al.* 2008, GAZAVE *et al.* 2011), inversions (FEUK *et al.* 2005), or segmental duplications (BEKPEN *et al.* 2012), identifying also potential traits specific for the human lineage that have arisen after the split from the common chimpanzee lineage. One example of this is gene conversion, in which a portion of one gene is misaligned during recombination and ends up copied onto another gene. This is more likely to happen genes that are members of the same family. The *SIGLEC11* gene, for example, underwent two consecutive human specific gene conversion events with an adjacent pseudogene *SIGLEC16*P in the hominin lineage. Gene conversion of *SIGLEC11* in humans resulted in microglia specific expression in the human brain (WANG *et al.* 2012). Microglia cells are responsible for immune defense and neuroprotection in the central nervous system. *SIGLEC11* expression alleviates neurotoxicity of microglial cells, and this can damage neurons and contribute to neurodegenerative disorders (WANG and NEUMANN 2010).

Other common ways of modifying gene function is through deletions of parts of the genome. There are thousands of well-conserved regions in non-human primates that have been removed from the human genome, and some of them might have important consequences. MCLEAN *et al.* (2011) for example tested the enhancer activity of a forebrain

specific p300 binding site located downstream near the tumor suppressor gene *GADD45G* in transgenic embryonic mice embryos (E14.5 days). This conserved chimpanzee enhancer, which represses cell cycle and can activate apoptosis, is located within a 3,181bp human-specific deletion. The authors have seen that the chimpanzee sequence drives significant gene expression in the developing ventral telencephalon and diencephalon in at least five independent transgenic embryos confirming that the ancestral sequence corresponds to a conserved forebrain-specific enhancer. In addition, the chimpanzee sequence presented significant gene expression when tested in immortalized human fetal neural progenitor cells. This suggests that the enhancer would also affect *GADD45G* transcription if it were still present in humans, thus inhibiting brain size expansion (MCLEAN *et al.* 2011).

The variability in the number of copies of a specific genomic region is one of the major components of human genomic variation and is thought to be an important contributor to phenotypic diversity and human disease (STANKIEWICZ and LUPSKI 2010). Due to the general concern about the impact of these copy number variations in disease susceptibility in the human population, the DECIPHER Consortium was created (FIRTH *et al.* 2009) to share information about different patients with any chromosomal imbalance internationally. GAZAVE *et al.* (2011) stepped forward into the human evolution to investigate these regions with variable number of copies in all great apes (bonobo, chimpanzee, gorilla and orangutan). They were able to identify genomic regions which have no or very low-frequency structural polymorphism in humans, while they present a high-frequency of copy number variants in all four great apes or in all three African great apes (excluding orangutans). Interestingly, by comparison with the DECIPHER data base, they were able to identify 285 regions with a known human pathogenic structural variant, which also present a high-frequency of copy number variation in some or all of the healthy great apes (GAZAVE *et al.* 2011).

Within the gene structure, it is possible that some portions of a protein sequence evolve and function independently from the rest of the protein. One remarkable example are the genome sequences encoding the DUF1220 protein, whose domains have undergone an increase in the copy number in the human lineage: the larger the evolutionary distance

from humans, the lower the number of copies was (POPESCO *et al.* 2006). It has been found that humans have 272 copies, more than twice the copy number of chimpanzees, which are next in number with 126 copies. Around 28 additional copies of DUF1220 domains have been added specifically to the human genome every million years since the human and Pan lineages diverged, and these species are really far from other mammals like mice, rats or squirrels that have only one copy (O'BLENESS *et al.* 2012). Interestingly, most of the DUF1220 sequences map within a copy number variable region of the chromosome 1q21.1 region that has been implicated in numerous recurrent human neurogenetic diseases, as noted above (DUMAS and SIKELA 2009). Therefore, DUF1220 domains might function as general effectors of brain size and may be largely responsible for the dramatic evolutionary expansion in brain size that occurred in the human lineage (DUMAS *et al.* 2012).

#### 1.1.3.3. Small genetic changes.

Small local alterations in the sequence of a gene can result in potential changes in the amino acid sequence of the encoded protein. The forkhead box P2 (*FOXP2*) gene provides a great example of this. It was discovered studying a family with an inherited defect of speech in both males and females of three generations (known as the KE family). They presented an arginine-to-histidine substitution at position 553 (R553H) of the FOXP2 protein, producing its loss of function (LAI *et al.* 2001). In humans, the loss of one copy of *FOXP2* leads to language impairment and abnormal cognitive activity (LAI *et al.* 2001). It was then proposed that this gene could have had an impact on human speech evolution (ENARD *et al.* 2009). The analysis of the evolution of *FOXP2* in primates revealed that the human protein sequence differs by two amino acid substitutions in exon 7 (a threonine-to-asparagine substitution at position 303 (T303N) and an asparagine-to-serine substitution at position 325 (N325S)) of the protein in chimpanzees, gorillas, and macaques, all of which share identical sequences (ENARD *et al.* 2002). The incidence of two amino acid fixations in a short amount of time in an overall much conserved protein is highly unlikely to have occurred by chance. In fact, based on the ratio of nonsynonymous to synonymous nucleotide changes (Ka/Ks) in humans, chimpanzees, and other species, it has been

reported that these *FOXP2* changes were likely the result of positive selection (ENARD *et al.* 2002, ZHANG *et al.* 2002, CLARK *et al.* 2003). ENARD and colleagues also investigated the properties of *FOXP2* in transgenic mice. They created mice with normal endogenous *FOXP2* gene carrying the two human-specific amino acid changes, and then compared them with mice with their normal domains and mice with only one functional copy of *FOXP2*. They suggested the comparison of "humanized"*FOXP2*-mice with mice with a single functional copy of *FOXP2,* since humans heterozygous for a nonfunctional *FOXP2* allele show speech and language impairments. Thus, some of the traits affected in opposite directions identified by ENARD and colleagues, such as decreased exploratory behavior, reduction of dopamine levels, increment of dendritic length, increment of long-term synaptic depression or the alteration of the vocalization are of potential interest as candidates for being involved in aspects of speech and language evolution.

A second example of a functionally important human specific amino acid change is in the cytochrome c oxidase subunit Va gene (*COX5A*). It encodes for highly conserved subunits, the majority of which have an increased rate of nonsynonymous substitutions in the anthropoid primate linage. Particularly, COX5A contains two human specific amino acid changes probably important in fat metabolism regulation due to its interaction with the thyroid hormone T2 (UDDIN *et al.* 2008, CIOFFI *et al.* 2010). These alterations in metabolism could have had important consequences in the increasing energetic demands of the brain during human evolution.

### 1.1.3.4. Changes in gene expression.

From early on, changes in gene expression on the human evolutionary lineage have been suggested to play an important role in the establishment of human-specific phenotypes (KING and WILSON 1975). During the last decade, gene-expression patterns in different tissues and stages of development have been compared between humans, chimpanzees, and other apes (especially orangutan and rhesus macaques) initially by microarrays and more recently by RNA-sequencing techniques (TABLE 2). These studies have aimed, among other

things, to determine whether primate models of human neurodegenerative diseases are valid according to genomic criteria (MARVANOVA *et al.* 2003), to identify genes with expression differences between species (ENARD *et al.* 2002, CÁCERES *et al.* 2003), to categorize shared and species-specific gene-expression profiles (UDDIN *et al.* 2004), to determine region-specific expression differences within the brain (KHAITOVICH *et al.* 2004, XU *et al.* 2010), and to evaluate the differences in their epigenetic regulation (ZENG *et al.* 2012).

In one of the first studies performed to understand the gene-expression differences between human and chimpanzees (ENARD *et al.* 2002), the authors concluded that the human lineage, in comparison to the chimpanzee lineage, underwent an accelerated rate of evolutionary changes in gene expression, specifically in the brain. It was shown that the gene-expression changes in the liver accumulated at equal rates in both species, whereas for the cerebral cortex, the change in the human lineage was more pronounced. Subsequent re-analysis of the microarray data of ENARD *et al.* has pointed out that the number of significant expression changes in the brain was indeed three times higher in humans than in chimpanzees (GU and GU 2003). An independent study using various samples of the cerebral cortex has suggested also the presence of a few hundred genes with higher expression levels in the brain of humans compared with that of chimpanzees (CÁCERES *et al.* 2003), which suggested that the human brain evolution was characterized by a global upregulation of the expression of many genes. A gene ontology analysis performed by CÁCERES *et al.* (2003) revealed that many of the genes showing an up- regulation in the human brain are involved in neuronal function. However, although the data sets from ENARD *et al.* and CÁCERES *et al.* provide evidence for predominant upregulation in human brain evolution, UDDIN and colleagues reported an increased down-regulation of genes in the human lineage, rather than an up-regulation (UDDIN *et al.* 2004). The results of this study (UDDIN *et al.* 2004) should be taken with caution, since as pointed out in a review by PREUSS *et al.* (2004), samples size at UDDIN's study was small, with only one chimpanzee and one gorilla included. Additionally, the down-regulation of expression could be mostly explained by the fact that the study's focus was on many expression changes of small magnitude that could be overestimated by a compensatory effect of array normalization. In

fact when the data of UDDIN *et al.* (2004) were re-analyzed using the same criteria employed in other studies, 59% of the genes that were differentially expressed showed higher expression levels in humans than in chimpanzees (PREUSS *et al.* 2004).

TABLE 2. Comparative studies of human and non-human primate gene expression levels and other regulatory elements in brain tissue.

| Brain tissue | Method | Species compared (number samples) | Sex | Reference |
|---|---|---|---|---|
| Prefrontal cortex (area 9) | HG_U95A arrays | Humans (3)<br>Chimpanzees (3)<br>Orangutan (1) | M<br>M<br>M | ENARD *et al.* (2002) |
| Neocortex | Membrane-based cDNA array | Human (7)<br>Chimpanzee (4)<br>Macaque (2) | -<br>-<br>- | ENARD *et al.* (2002) |
| Cerebral cortex | HG_U95Av2 arrays | Human (5)<br>Chimpanzee (4)<br>Macaque (4) | 3M/2F<br>1M/3F<br>1M/3F | CÁCERES *et al.* (2003) |
| Prefrontal cortex | HG_U95A arrays | Human (1)<br>Chimpanzee (1)<br>Macaque (2)<br>Marmoset (2) | -<br>M<br>F<br>F | MARVANOVA *et al.* (2003) |
| Anterior cingulated cortex | HG_U133A, -B arrays | Human (3)<br>Chimpanzee (1)<br>Gorilla (1)<br>Macaque (3) | F<br>F<br>M<br>2M/1F | UDDIN *et al.* (2004) |
| Several areas | HG_U95Av2, -B, -C, -D, -E arrays | Human (3)<br>Chimpanzee (3) | M<br>M | KHAITOVICH *et al.* (2004) |
| Prefrontal cortex (area 9) | HG_U133plus2 | Human (6)<br>Chimpanzee (5) | M<br>M | KHAITOVICH *et al.* (2005) |
| Cerebellar cortex | RNA-sequencing | Human (10)<br>Chimpanzee (4)<br>Macaque (5) | M<br>M<br>M | XU *et al.* (2010) |
| Prefrontal cortex and cerebellum | RNA-sequencing and microarrays | Human (33)<br>Chimpanzee (16)<br>Macaque (44) | 22M/11F<br>8M/8F<br>40M/4F | LIU *et al.* (2012) |
| Telencephalon | Digital gene expression and microarrays | Human (12)<br>Chimpanzee (8)<br>Macaque (4) | 9M/3F<br>6M/2F<br>3M/1F | KONOPKA *et al.* (2012) |
| Frontal cortex | Methyl-C-Seq | Human (3)<br>Chimpanzee (3) | 2M/1F<br>M | ZENG *et al.* (2012) |

One of the most complete microarray studies (KHAITOVICH *et al.* 2005) was performed in five different tissues (heart, kidney, liver, prefrontal cortex and testis) to correlate the sequence differences between humans and chimpanzees with expression differences. The

authors showed that, even though the patterns of evolutionary change in gene expression were compatible with a neutral model, testes and brain tissue stood out as putative targets of positive selection. These tissues presented an increase of both expression and amino acid changes on the human lineage, suggesting that their regulation could have played crucial roles in the evolution of some organ systems, such as those involved in cognition or male reproduction (KHAITOVICH *et al.* 2005).

Although all of these microarray studies were important first steps in uncovering human-specific patterns of gene expression in the brain, array technology has several limitations, which are especially relevant to evolutionary comparisons. For example, microarray technology relies on a priori knowledge of the sequence of the gene being measured, which excludes identifying transcripts not annotated (KONOPKA *et al.* 2012). Background levels of hybridization also limit the accuracy of expression measurements, particularly for transcripts present in low abundance (MARIONI *et al.* 2008). Finally, –and perhaps more importantly– probes on the microarray are human-specific, and sequence mismatches between human probes and non-human primate mRNAs can give false indications of lower gene-expression levels in the non-human primates compared to humans (PREUSS *et al.* 2004). To overcome these limitations, new sequence-based approaches have been used in the last years to measure gene-expression levels, the so-called Next-Generation Sequencing technology (NGS). These technologies produce millions of short sequence reads and are routinely being applied to the study of genomes, epigenomes and transcriptomes (OSHLACK *et al.* 2010). Some of these studies have revealed for example that 40-48% of the total brain transcriptome corresponds to transcripts originating within intronic and intergenic repetitive sequences (XU *et al.* 2010) and they have identified human-specific transcriptional networks that may provide particular insight into human brain evolution (KONOPKA *et al.* 2012). Interestingly, a comparative postnatal brain ontogenesis study in humans, chimpanzees, and rhesus macaques, using both microarray and sequencing platforms (LIU *et al.* 2012), found that the most relevant human-specific expression changes affect genes associated with synaptic functions and represent an extreme shift in the timing of synaptic development in the prefrontal cortex, but not in the cerebellum. The authors suggest that this delay in cortical synaptogenesis, extending the period of synapse formation

to over 5 years of age in humans compared to a few months in chimpanzees and macaques, could be a potential mechanism contributing to the emergence of human-specific cognitive skills (LIU *et al.* 2012). Finally, following a similar methodology to the sequence-based studies, the first whole-genome methylation maps of the prefrontal cortex of humans and chimpanzees at nucleotide-resolution have been recently published, showing a strong relationship between differential methylation and gene expression (ZENG *et al.* 2012).

## 1.2. Regulatory evolution of gene expression.

The regulation of genes can occur at many different levels with multiple possible consequences in the gene expression pattern, such as changes in the structure of the gene, regulatory elements acting in *cis* or *trans,* epigenetic modifications of the chromatin, changes in the sequence of the mRNA, and other post transcriptional modifications. In addition, there could be even post translational mechanisms (which will not be considered for a detailed explanation in this work), such as the addition of functional groups or other proteins or peptides, changes in the chemical nature of the amino acids, protein degradation, etc. This myriad of possible mechanisms has been proposed to provide a great capacity to fine-tune gene expression levels, making regulatory changes one of the main potential sources of genetic variation for evolution (FIGURE 2).

### 1.2.1. The role of regulatory mutations in the evolution.

An important objective of evolutionary genomics is to understand the relationship between mutation, natural selection, and variation in gene regulation (SHIBATA *et al.* 2012). The main goal is to identify the genetic changes and the molecular mechanisms that underlie phenotypic diversity and to understand the evolutionary pressures under which phenotypic diversity evolves (ROMERO *et al.* 2012). However, the relative contributions of

different mechanisms to phenotypic evolution depend upon both what is genetically possible, and what is permitted by natural selection (BOX1) (CARROLL 2005). Even though it has become clear that variation in regulatory mutations often plays a key part in the evolution of morphological phenotypes (STERN and ORGOGOZO 2008), there is still some debate about if phenotypic changes are most likely caused by coding changes or regulatory changes (HOEKSTRA and COYNE 2007, CARROLL 2008).

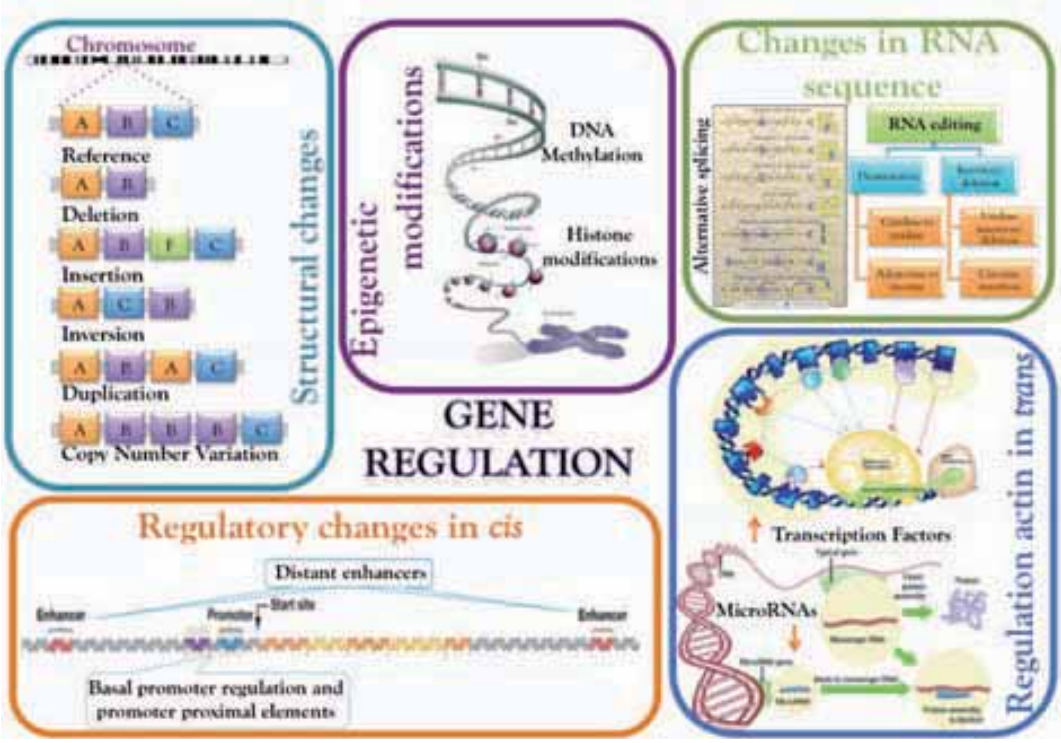



FIGURE 2. Global regulation of gene expression levels. Gene regulation can occur at many different levels: changes in the structure of the gene, such as deletions, insertions, inversions, duplications or copy number variations, regulatory elements acting in *cis* (promoters or enhancers) or *trans* (transcription factors or micro RNAs), epigenetic modifications of the chromatin, like DNA methylation or histone modifications, and changes in the sequence of the RNA, as alternative splicing or RNA edition. This figure has been created based on several images obtained from Google Images web browser (www.google.com), their rights belong to the respective authors. Some images were tracked until scientific publications. Structural changes: Modified from BAKER (2012), Epigenetic modifications: modified from QIU (2006)

It has been considered that one limitation affecting the relative contribution of different genetic mechanisms to phenotypic variation is the ability of the different mutations to alter more than one trait, in other words, the pleiotropy of the mutation

(STERN 2000). Some authors argue that coding mutations affecting protein structure are not susceptible to dynamic change according to the functional necessities. They remain mostly static and are just usually affected by alternative transcription start sites (TSSs), distinct splicing or post-translational modifications. In comparison, phenotypic modifications controlled by changes in regulatory regions affecting the transcription are much more predisposed for adaptation to new and different environments according the respective functional demands (JACOB and MONOD 1961, STERN 2000). In fact, due to their modular organization, a mutation in a specific regulatory region might affect only one part of the overall transcription profile (STERN 2000). For example, the effects of a regulatory mutation could be limited to one stage of the development, particular environmental conditions, or to a single organ or tissue even when the gene is much more widely expressed, resulting in potentially fewer functional trade-offs and less pleiotropic effects than coding mutations (CARROLL 2005, WRAY 2007, LAPPALAINEN and DERMITZAKIS 2010).

BOX 1: Types of natural selection

Natural selection can cause several different types of changes in a population. How the population changes, depend upon the particular selection pressure the population is under and which traits are favored in that circumstance.

Positive selection, (also known as Darwinian selection or directional selection) in this type, traits at one end of a spectrum of traits are selected for, whereas traits at the other end of the spectrum are selected against. Over generations, the selected traits become common, and the other traits become more and more extreme until they are eventually phased out. e.g. The ability of digest lactose as an adult thanks to SNPs in the encoding gene. (BERSAGLIERI *et al.* 2004).

Purifying selection, (also known as negative or stabilizing selection) in this type extreme or unusual traits are eliminated. Individuals with the most common traits are considered best adapted, maintaining their frequency in the population. Over time, nature selects against extreme variations of the trait. e.g. The existence of "living-fosiles" which ecological niche have not to change for millions of years, fossil forms of the species can be almost indistinguishable from their present-day descendants (VANE-WRIGHT 2004).

Balancing selection, in this type, the environment favors extreme or unusual traits and selects against the common traits, favoring diversity. Under balancing selection, the spread of an allele never reaches fixation, and therefore it can initially seem to be undergoing positive selection, but it then undergoes negative selection when its frequency is too high. Thus alleles cannot be classified as advantageous or deleterious. e.g. The resistence to the infection with the malaria parasite *Plasmodium vivax* of heterozygote individuals for the β-globin gene (HAMBLIN and DI RIENZO 2000).

## 1.2.2. Possible mechanisms of gene regulation.

As shown in FIGURE 2, the possible regulatory mechanisms that may affect the expression of a gene can be divided into five different groups according to how they regulate the gene expression. The main concepts of the different mechanisms and some of their most representative examples are described below.

### 1.2.2.1. Genomic structural variations.

As introduced above, the structural changes –also referred to as genomic structural variation– provide a mechanism through which evolutionarily adaptive changes can be introduced (O'BLENESS *et al.* 2012). These changes may occur by mechanisms like deletions or insertions, which result in the loss or gain of genetic material; duplications or copy number variations, which alter the dosage of the same existing genomic material which was already in the genome before (this can account to different copies of a particular gene increasing or decreasing its expression or leading to different functions); or like chromosomal inversions, in which the orientation of a segment of a chromosome is reversed without changing the overall amount of genetic material, which can result in structural problems with meiosis (pericentric inversions), disrupt an open reading frame or alter gene expression through position effects.

Human evolution, for example, has been strongly affected by the insertions of different transposable elements (TEs) (BRITTEN 2010). It has been shown that they may affect genes and their expression (PONICSAN *et al.* 2010), supplying for example transcription factor binding sites (COWLEY and OAKEY 2013), being involved in alternative splicing (LI *et al.* 2001), or even inducing the ectopic expression of genes like in the human amylase gene (*AMY1*). This gene, encoding for a salivary enzyme that catalyzes the first step in digestion of dietary starch and glycogen, constitutes an interesting example of evolutionary relevant gene structural variations. Firstly, it has been shown that during the evolution of this gene family,

the insertion of a processed γ-actin pseudogene in the proximal promoter region of the ancestral amylase gene was followed by two retroposon insertions after the divergence of the New World monkeys from the primate ancestral tree (SAMUELSON *et al.* 1996). Secondly, the amylase gene has approximately three times more copies in humans compared to chimpanzees, and this copy number differences correlate positively with the higher levels of salivary amylase protein (PERRY *et al.* 2007).

#### 1.2.2.2. Regulatory changes in *cis*.

Differences in gene expression may also arise from changes in *cis*-regulation, which is required for the proper temporal and spatial control of gene expression (EPSTEIN 2009, WITTKOPP and KALAY 2012). *Cis*-regulatory sequences, such as promoters or enhancers, are composed of DNA containing binding sites of transcription factors and/or other regulatory molecules that are needed to activate and sustain transcription (WITTKOPP and KALAY 2012). Thanks to the multiple studies of *cis*-regulatory mutations in several organisms, their ability to produce transcription changes capable of altering ecologically relevant traits is now known. But, what kinds of mutations trigger *cis*-regulatory divergence? It has been commonly observed that substitutions of a few nucleotides or even the mutation of a single one can be enough to alter the activity of *cis*-regulatory sequences between species. For example, between *Drosophila* species, multiple single nucleotide substitutions within various transcriptional enhancers only had a relatively small effect on gene expression and on the final phenotype, when taken individually. However, when they were combined, they produced large morphological differences, such as the loss of trichomes (FRANKEL *et al.* 2011) or a divergent pigmentation (JEONG *et al.* 2008, SHIRANGI *et al.* 2009). Similar to this, it has been shown that the human-accelerated conserved non-coding sequence 1 (*HACNS1*) has evolved a novel enhancer activity relative to chimpanzees and rhesus macaques (PRABHAKAR *et al.* 2008). Specifically, the human copy of *HACNS1*, which has 13 nucleotide substitutions within an 81-bp region of the 546-bp conserved element, activates the expression of a reporter gene in transgenic mice in the anterior fore- and hindlimb bud during embryonic development. Orthologous sequences from chimpanzees and rhesus

macaques lack this enhancer activity when assayed in transgenic mice. It has been shown that human specific changes in hindlimb morphology, such as the inflexibility and shortened digits of the human foot, facilitated habitual bipedalism. This suggested that the gain of function in *HACNS1* might have influenced the evolution of these or other human limb features by altering the expression of nearby genes during limb development (PRABHAKAR *et al.* 2008).

In the same way, a single SNP, which disrupts a transcription factor binding site in the *Duffy* gene promoter, makes humans resistant to infection with the malaria parasite *Plasmodium vivax (TOURNAMILLE et al. 1995, HAMBLIN and DI RIENZO 2000);* or two single SNPs located ~14 kb and ~22 kb upstream of the gene encoding lactose (*LCT*) (C/T -13910 and G/A-22018) has been associated with a dietary adaptation to digest lactose as an adult (lactose persistence) (ENATTAH *et al.* 2002). These SNPs are located within introns 9 and 13 of the adjacent *MCM*6 gene. While the C/T -13910 SNP has been associated with the lactose persistence in Europeans, TISHKOFF and colleagues identified three additional SNPs within *MCM6* intron 13 associated with lactose persistence and the increase or *LCT* transcription in East African (TISHKOFF *et al.* 2007).

In addition to nucleotide substitutions, deletions between species are also a very common way to produce changes in *cis*-regulatory activity. Lake threespine sticklebacks (*Gasterosteus aculeatus)* suffered recurrent deletions disrupting certain *cis*-regulatory activity after their isolation from the rest of the oceanic population, which resulted in the reduction or loss in skeletal armor involving the dorsal spines and pelvic girdle (SHAPIRO *et al.* 2004, CHAN *et al.* 2010). Humans have been also affected by specific deletions: 510 putative *cis*-regulatory elements highly conserved between chimpanzees and other mammals are missing from the human genome, which produces an up- or down-regulation of the gene depending on the depleted region (MCLEAN *et al.* 2011). One example of this is the regulation of the tumor suppressor gene *GADD45G,* which was previously commented on in section 1.1.2.4 "Structural changes". Finally, insertions can also alter a *cis*-regulatory activity by introducing new transcription factor binding sites or disrupting spacing between existing ones. Indeed, insertions of transposable elements also have the potential to contribute to *cis*-regulatory

divergence by bringing new *cis*-regulatory elements to existing genes. The pituitary hormone prolactin represents one example of this, in which the insertion of an LTR element 5.8 kb upstream to the TSS before the divergence of Old World monkeys from higher apes resulted in an alternative promoter with an extra exon (GERLO *et al.* 2006).

### 1.2.2.3. Regulatory changes in *trans*.

Transcription factors are proteins acting in *trans* to control the activation or repression of gene expression by binding to particular DNA sequence motifs proximal to a gene's TSS or a more distant enhancer. The large number of transcription factors in the human genome emphasizes their importance as potential regulatory elements with an extremely wide range of possible combinations. They are frequently the terminal components of signaling cascades relaying signals from a variety of sources, which ensure the correct spatio-temporal expression of the genes they control. Transcription factor genes are subject to the same transcriptional regulatory mechanisms as other protein-coding genes.

Some transcription factors might have contributed to human cognitive evolution. This fact is in part based on genomic sequence comparisons between humans and other species, like the work of VARKI and ALTHEIDE (2005 ), and on searches for unusual patterns of sequence change, combined with functional information from human disease studies (SOMEL *et al.* 2013). As previously discussed in section 1.1.2.5 "Small genetic changes", the two amino acid substitutions discovered in the FOXP2 transcription factor, can affect the evolution of gene expression in humans (ENARD *et al.* 2009, KONOPKA *et al.* 2009). These amino acid differences between the human and the chimpanzee versions of FOXP2 were associated with a different regulatory potential of this transcription factor in a human cell line, which could partly explain the differences in mRNA abundance of the genes containing a binding site for FOXP2 between adult human and chimpanzee brains (KONOPKA *et al.* 2009).

In addition to transcription factors, small regulatory RNAs, called microRNAs (miRNAs), can also regulate the expression of genes in *trans.* They are short (20-23 nucleotides) non-coding regulatory RNAs that influence gene expression at a post-transcriptional level (HU *et al.* 2011). The first study of miRNAs expressed in human and chimpanzee brains reported a significant number of differences in the variety of miRNA between both species (BEREZIKOV *et al.* 2006). However, following studies indicated that most miRNA genes are highly conserved among primates in terms of both sequence and expression (LIANG and LI 2009, SOMEL *et al.* 2010). Likewise, it has been shown that among 325 miRNAs expressed in the human, chimpanzee and macaque prefrontal cortex, up to 11% and 31% of the expressed miRNAs diverged significantly in humans versus chimpanzees and humans versus macaques, respectively. Expression differences between humans and chimpanzees were equally distributed between the human and chimpanzee evolutionary lineages (HU *et al.* 2011). One of these miRNAs, miR-184, which has been implicated in driving neural stem cell proliferation, is present in large quantities in the prefrontal cortex and in the cerebellum of humans, but not in chimpanzees or macaques (LIU *et al.* 2010).

#### 1.2.2.4. Epigenetic chromatin modifications.

Epigenetic mechanisms of gene control do not directly rely on the DNA sequence itself, but instead on its higher order modifications and chromatin structure. They can be divided into two classes: DNA methylation and histone modifications. DNA methylation at CpG dinucleotides or covalent modifications of histone proteins can work in synchrony to regulate the gene expression (INBAR-FEIGENBERG *et al.* 2013). DNA methylation is established *de novo* by the DNA methyltranferase enzymes and is maintained through mitosis (MAUNAKEA *et al.* 2010). In mammals, most of the DNA methylation occurs at CpG dinucleotides (IBRAHIM *et al.* 2006, ZILLER *et al.* 2011). They are concentrated in genomic regions called CpG islands, which are frequently methylated when they are located within gene promoters of silent genes (COSTELLO and PLASS 2001), but predominantly unmethylated when they are found surrounded by actively transcribed genes and tumor

suppressor genes (BIRD 2002). Histone modifications are mostly regulated by histone acetyltransferases and deacetylases. DNA is wrapped around histone proteins, which are grouped forming an octamer with two copies of each histone: H2A, H2B, H3, and H4. This structure is called nucleosome and it is considered the basic unit of chromatin which provides structural stability and the capacity to regulate gene expression (INBAR-FEIGENBERG *et al.* 2013). Recently it has been shown that different types of histone modifications are associated with different types of activity states of the chromatin (ERNST *et al.* 2011) Comparisons between the methylation levels of humans and non-human primates in several tissues, are recently providing a large number of gene expression differences, which might be explained –at least partially– by corresponding differences in methylation levels (PAI *et al.* 2011, ZENG *et al.* 2012). Additionally, it has been shown that differential DNA methylation might potentially contribute to the evolution of disease susceptibilities (ZENG *et al.* 2012).

### 1.2.2.5. Modifications of the mRNA.

Modifications in the mRNA sequence can also alter the expression of a gene. This may include alternative splicing, the use of alternative untranslated regions, or post-translational edition events like nucleotide substitutions (such as cytidine to uridine and adenosine to inosine deaminations, which can create new start/stop codons or alterations in splice sites) or nucleotide additions within the RNA molecule (FARAJOLLAHI and MAAS 2010).

Alternative splicing has been classified into diverse subgroups: (1) exon skipping, where the exon is spliced out of the transcript together with its flanking introns; (2) alternative 5' or 3' donor sites, which result from the recognition of two or more splice sites at the ends of an exon; (3) intron retention, in which an intron can remain in the mature mRNA molecule, and (4) mutually exclusive exons, in which one of two exons is retained in mRNAs after splicing, but not both. In addition, RNAs can suffer some changes in the maturation process such as alternative TSSs, which can create alternative 5'-untranslated regions or add additional exons, or multiple polyadenylation sites, which can create several

mRNA that differ in the length of the 3'-end (BLENCOWE 2006, KIM *et al.* 2007). Previous studies using high-throughput sequencing technology have reported that up to 92%~94% of human multiexon genes undergo alternative splicing, suggesting this process as one of the most significant components of the functional complexity of the human genome (PAN *et al.* 2008, WANG *et al.* 2008). In spite of this, the splicing mechanism is not always well performed, creating abnormal spliced mRNAs, which have been found in a high proportion of malignant cells (SKOTHEIM and NEES 2007). It is thought however that the deleterious effects of mis-spliced transcripts are usually safeguarded and eliminated by a cellular post-transcriptional quality control mechanism termed nonsense-mediated mRNA decay (KARAM *et al.* 2013)

Considering possible differences in alternative splicing between humans and chimpanzees, it has been revealed that the majority of orthologous human and chimpanzee genes with splicing level differences do not overlap genes that display transcript level differences, suggesting that alternative splicing has evolved rapidly to increase the number of protein coding genes with altered patterns of regulation between humans and chimpanzees. (CALARCO *et al.* 2007).

## 1.3. Methods for characterization of regulatory regions.

It has been shown that *cis*-regulatory divergence is common between closely related species (WITTKOPP *et al.* 2004, TIROSH *et al.* 2009, MCMANUS *et al.* 2010). Clarifying the molecular mechanisms responsible for regulatory divergence requires identifying the changes in the sequence that alter the *cis*-regulatory elements (CREs) (WITTKOPP and KALAY 2012). Finding regulatory sequences is much more difficult than finding the coding region of genes, since they are scattered across the 98% of the human genome and tend to be short and lack well defined sequence signatures. These regulatory elements can be

divided in three groups depending on their distance to the TSS (FIGURE 3): (1) Core promoters with the closest spatial relation to the TSS; (2) Proximal promoters including different binding sites for transcription factors (TFBS) and response elements (RE); and (3) distal elements, such as enhancers, silencers or insulators that control the activity of the TSS.

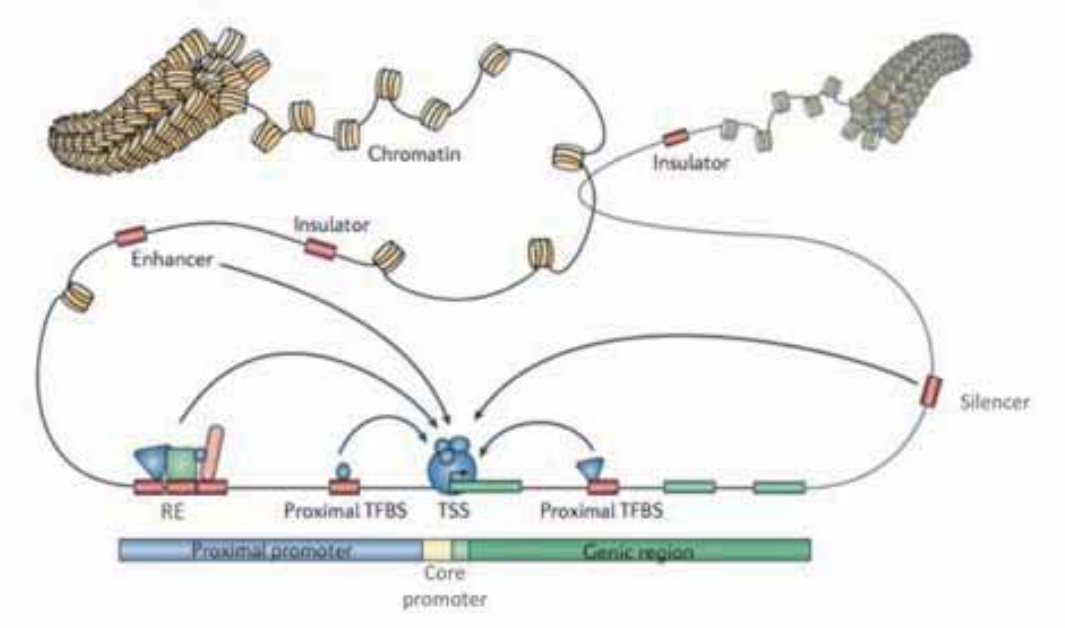



FIGURE 3. Control of gene transcription by different regulatory elements. The promoter region around the transcription start site (TSS) is divided in a short core promoter with the sequence elements that are bound by the RNA-Polymerase II complex and a larger proximal promoter upstream of the TSS, which contains different binding sites for transcription factors (TFBS) and response elements (RE). Distal enhancers and silencers are able to promote or suppress the promoter activity, respectively. The structure of the chromatin can be tightly wrapped or accessible according to insulators, which can block enhancers or prevent the advance of condensed chromatin. Figure modified from LENHARD *et al.* (2012)

### 1.3.1. What is a promoter?

Promoters are crucial for gene regulation. They are the sequences immediately flanking genes where the transcription machinery assembles before initiating the synthesis of mRNA (LENHARD *et al.* 2012). They usually contain a number of binding elements for various

components of the basal transcription machinery, as well as for transcription factors capable of relaying regulatory signals to the promoter (COOPER *et al.* 2006, CONSORTIUM *et al.* 2007). It has been estimated that 20-55% of all human genes have multiple promoters (COOPER *et al.* 2006, KIMURA *et al.* 2006, TASLIM *et al.* 2012). Alternative promoters are important in regulating gene expression and generating protein diversity (LANDRY *et al.* 2003).

Promoters are often described as having two separate parts: the core promoter and the proximal promoter regions. The core promoter is the sequence of about 200 bp near the TSS and it contains the sequence elements that are bound by the RNA-Polymerase II (RNAP-II) complex (VEDEL and SCOTTI 2011). There are several classes of structurally different core promoters in eukaryotes. The best-known class of promoters –and the easiest to identify– possesses a TATA-box. This is an AT-rich element found at -25 to -30 bases from the TSS, which is then bound by a TATA binding protein (TBP), a key component in the assembly of the RNAP-II complex (HERNANDEZ 1993). However, some promoters do not have a TATA-box (referred to as TATA-less). In humans, it seems that only around 20-30% of the promoters contain a TATA-box (SUZUKI *et al.* 2001, JIN *et al.* 2006). A component that is always present in all types of promoters is the initiator (Inr) element that encompasses the TSS. This element has the consensus sequence YYANWYY[1] in mammals (SMALE and BALTIMORE 1989, JIN *et al.* 2006) and the adenosine residue in the sequence marks the beginning of transcription (base +1). Some TATA-box promoters have a sequence called BRE element (B recognition element) which is directly recognized by the transcription factor IIB of the RNAP-II complex and which contributes to the transcription initiation (LAGRANGE *et al.* 1998). In contrast, some TATA-less promoters have a downstream promoter element (DPE), which assists the initiator in controlling the precise initiation of transcription in a similar way as the TATA-box, although with weaker effects (KADONAGA 2002). A motif ten element (MTE), which is conserved from *Drosophila* to humans (LIM *et al.* 2004), and requires the presence of an initiator element but is

[1] According to the International Union of Pure and Applied Chemistry (IUPAC), Roman character Y (pyrimidine) represents the nucleotides C or T, W (weak) represents A or T, and N represent any nucleotide.

independent from TATA-boxes or DPE, completes the core promoter. It promotes transcription by RNA polymerase II when it is located precisely at positions +18 to +27 relative to the TSS (LIM *et al.* 2004).

The other separate segment of a promoter, which is less well defined than the core promoter, is the proximal promoter or extended promoter region. It has been suggested that the core promoter recruits the general transcriptional apparatus and supports basal transcription, while the proximal promoter recruits transcriptional activators, which are necessary for appropriately activated transcription (LEE and YOUNG 2000). Even though the proximal promoter is described as the region immediately upstream (up to a few hundred base pairs) from the core promoter, there is no agreement about when the core promoter ends and the proximal promoter begins. For example, some authors consider the CCAAT-box, a motif located 75-80 base pairs upstream of the TSS, as part of the core promoter (KARAM *et al.* 2013), whereas other publications describe it as part of the proximal promoter (FANG *et al.* 2004). The functional characteristics of the proximal promoter depend on the transcription factor binding sites contained (typically activators very position dependent) and the relative spacing and clustering between proximal and core promoter. Response elements (REs) are short sequences of DNA within the proximal promoter region that are able to bind specific transcription factors and regulate transcription of genes. Specific recognition of the response elements by transcription factors and the ways they assemble on promoters and enhancers is essential for functional and gene-specific transcription initiation (GEORGES *et al.* 2010, PAN *et al.* 2010). Examples of response elements are HSE (heat shock sequence elements), HRE (hormone response elements) and SRE (serum response elements).

## 1.3.2. Other types of regulatory elements.

Cellular responses to environmental signals rely on tight gene regulation. As previously described, gene expression levels can be regulated by several different mechanisms. Although promoters could be considered among the most important gene regulatory elements, there are also other elements involved in the regulation of gene transcription.

### 1.3.2.1. Enhancers and silencers.

Enhancers were, along with promoters, the earliest regulatory elements to be discovered (KHOURY and GRUSS 1983). They are DNA sequences typically no longer than a few hundred base pairs in total, containing multiple binding sites for a variety of transcription factors. Unlike promoters, they have no predictable spatial relationship to the TSS. Enhancers can activate transcription independently of their location, distance or even relative orientation with respect to the promoters of the genes that they regulate (BANERJI *et al.* 1981, BLACKWOOD and KADONAGA 1998). In some instances, they can even activate transcription of genes located in a different chromosome (GEYER *et al.* 1990, LOMVARDAS *et al.* 2006). In some enhancers the interaction with a variety of transcriptional activator proteins (which then interact with the basal transcription machinery to modulate expression) is performed by a looping in chromatin, which physically approaches all the implicated distant elements (CARTER *et al.* 2002, KLEINJAN and VAN HEYNINGEN 2005, WEST and FRASER 2005, SEXTON *et al.* 2009).

Silencer elements are negative regulatory elements acting in *cis*. They are functionally similar to enhancers in the control of the activity of the promoter, but they act by suppressing gene expression rather than promoting it (PETRYKOWSKA *et al.* 2008).

### 1.3.2.2. Insulators.

Other class of regulatory elements in the DNA that may control the expression of a gene are insulators. They have a common ability to protect genes from inappropriate signals produced by their surrounding environment. FELSENFELD and colleagues (WEST *et al.* 2002, GASZNER and FELSENFELD 2006, WALLACE and FELSENFELD 2007, GHIRLANDO *et al.* 2012) have been studying insulators over the last decade. They have defined two different classes of insulators. One class is the enhancer-blocking insulators, which inhibit the action of distal enhancers in is effect to activate the promoter. Enhancer blocking only occurs if the insulator is situated between the enhancer and the promoter, not if it is placed elsewhere. This property can avoid that an enhancer activates the expression of an adjacent gene from which it is blocked, but leaves the unblocked side free to stimulate expression of genes located there. The other class is called barrier insulators. This class prevents the advance of nearby condensed chromatin that might otherwise silence expression (WEST *et al.* 2002, WALLACE and FELSENFELD 2007).

### 1.3.2.3. microRNA binding sites.

As previously mentioned, microRNAs (miRNAs) are short endogenous single-stranded RNA molecules of 21-23 nucleotides which do not produce any protein, unlike other genes. There are at least 800 miRNAs within the human genome, each of which has a different function. They work by binding to partially complementary sites in the mRNA and by interacting with hundreds of potential target genes, thus adding to the complex regulation of the human genome. In particular, it has been suggested that microRNAs are involved in post-transcriptional gene expression silencing and inhibition (KONG *et al.* 2008). There are four main mechanisms by which a microRNA can mediate a post transcriptional gene repression in animal cells: (1) inducing a pronounced target mRNA degradation, which is initiated by deadenylation of the poly(A) tail and decay of target mRNAs (GIRALDEZ *et al.* 2006, WU *et al.* 2006); (2) repressing the translation initiation at either the cap-recognition stage (HUMPHREYS *et al.* 2005) or the 60S subunit joining stage (CHENDRIMADA *et al.*

2007); (3) blocking the elongation of the mRNA (PETERSEN *et al.* 2006); or (4) inducing a proteolytic cleavage of nascent polypeptides (NOTTROTT *et al.* 2006).

### 1.3.3. Computational characterization of promoter regions and other regulatory elements.

Like other regulatory elements, it is essential to identify promoters against the genomic background. Given the considerable length of the human genome, *in silico* methods for predicting promoters computationally have been favored for a long time, instead of characterizing such a long sequence experimentally at tremendous costs in time, effort and financial assets.

Promoters are usually located immediately 5' of their TSSs. As such, early promoter prediction algorithms were TSS predictors heavily relying on known promoter elements, such as the TATA box, to identify promoter regions. With the beginning of the Human Genome Project in the 1990s, the sequencing of complementary DNA (cDNA) and the generation of expressed sequence tags (ESTs) became common, and served as standard markers for expressed genes (ADAMS *et al.* 1991). In general, the 3' ESTs mark the end of transcription reasonably well. However, although ESTs available in the GenBank database (http://www.ncbi.nlm.nih.gov/) could be used to identify putative promoter regions, due to the quick degradation of RNA, the 5' ESTs may end at any point within the transcript and have considerable error margins. In a similar way to ESTs, cap analysis of gene expression (CAGE) is a method based on the sequencing of concatamers of DNA tags from the initial 20 nucleotides of the 5′ end of mRNAs. The advantage in this case is that the sequencing is focused in the 5' ends of capped transcripts, representing the real beginning of the mRNA. Therefore, CAGE allows high-throughput gene expression analysis and the identification of TSS and the analysis of promoter usage (SHIRAKI *et al.* 2003, KODZIUS *et al.* 2006, TAKAHASHI *et al.* 2012). This method has been widely used in the different phases of the

FANTOM project (KAWAJI *et al.* 2006) and the results are available at http://fantom.gsc.riken.jp. The genome-wide analysis of multiple CAGE libraries from human and mouse showed that promoters can be divided into two different groups depending on the profile of their TSSs. In one group, promoters have tightly defined single TSSs that correspond well to the classical definition and have a high probability of containing TATA-boxes. In the other group, promoters are more evolvable and have roughly defined start sites, which were more likely to be situated inside CpG islands (CARNINCI *et al.* 2006).

Other criteria for detecting functional regulatory elements is the computational search of patterns of sequence conservation between species, since these sequences tend to be more conserved than non-functional sequences (HE *et al.* 2009). The sequence conservation was initially used to identify candidate binding sites across diverse species. The technique was then called phylogenetic footprinting (ZHANG and GERSTEIN 2003), and was focused on the detection of sequences that have remained highly conserved during evolution, because they have a major probability to be functional (but the reverse proposal is not true: a sequence can be functional although non-conserved). This method has become more common since the full sequencing of multiple genomes from vertebrates (LINDBLAD-TOH *et al.* 2011) or closely related non-vertebrate species has been achieved, such as *Drosophila* or yeast (DERMITZAKIS *et al.* 2005, LOOTS and OVCHARENKO 2007). The comparison of aligned genome sequences of multiple species versus the sequence of one selected species can create a conservation profile. However, even when the conserved non-coding sequences are identified, it is still a challenge to find the ones that have real regulatory function, and separate them from other conserved elements, such as unknown exons or RNA genes (KHAMBATA-FORD *et al.* 2003).

Typically, transcription factor binding sites are very short sequences of 6-12 bp, in rare cases up to 18 bp, which usually can contain a certain amount of ambiguous bases and therefore makes their reliable identification very difficult. Considering them in terms of a position weight matrix (BUCHER 1990), describing the probability of each base being any one of the four possible bases, a promoter will have a great number of potential binding

sites simply by chance, although they will be functional only in a small number of cases. A variety of databases are available that contain the weight matrices of known transcription factors (FAZIUS *et al.* 2011, KULAKOVSKIY *et al.* 2013). Additionally, an increasing number of software applications have been released to predict transcription factor binding sites computationally. Some of these programs, like regulatory Vista (LOOTS *et al.* 2002) or the combination of ECR (Evolutionary Conserved Regions) browser (OVCHARENKO *et al.* 2004) with Multi-TF, merge database searches of transcription binding motifs with comparative sequence analysis, reducing the number of incorrectly predicted transcription factor binding sites and increasing the chances of identifying functional ones across species. These transcription factor binding sites are often used to search putative promoter sequences, but they must be considered with caution in the absence of experimental data confirming their functionality. Computational regulator searches must therefore be complemented with experimental verification of the predictions.

### 1.3.4. Experimental characterization of regulatory elements.

The development of new technologies for studying the functional characteristics of non-coding DNA on a genome-wide scale and the decreasing cost of doing large-scale experiments has promoted the interest in scanning the genome for promoter elements in an unfocused and unbiased way, without necessarily relying on previous *in silico* predictions. Traditionally, the functional characterization of putative promoters was performed following laborious and intensive experimental methods that required some kind of knowledge of the candidate region in advance, such as the presence of a confirmed TSS. These classical methods, nowadays still in use with more advanced vectors and reporter enzymes, are based on the cloning of the putative promoters into a reporter plasmid and the transfection into an *in vitro* model system (either cultured cells or model organisms), followed by nested deletions to determine the limit of the minimum sequence necessary to drive expression (MYERS *et al.* 1986, TAKAOKA *et al.* 1998, WALLICH *et al.* 1998). When the

focus of the characterization is aimed at the actual promoter function and not at its exact position, the reporter assays can be performed to analyze the transcriptional activity of the promoter. In this case, the region of interest is cloned into a plasmid with a reporter gene that encodes a visually identifiable fluorescent or luminescent protein and allows measuring the transcriptional activity. The most common reporter genes are the gene that encodes jellyfish green fluorescent protein, which causes expressing cells to glow green under blue light (TSIEN 1998), or the oxidative enzyme luciferase from different species of either firefly *Photinus pyralis* or the sea pansy *Renilla reniformis*, which triggers a chemical catalytic reaction that produces oxyluciferin in an electronically excited state, when excited with light, and then emits visible light by itself (GOULD and SUBRAMANI 1988, MARQUES and ESTEVES DA SILVA 2009). Another available marker is the red fluorescent protein from the gene *dsRed* (BAIRD *et al.* 2000, CAMPBELL *et al.* 2002). The reporter gene is then placed under the control of the target promoter and the reporter gene product's activity is quantified.

Another typical method of promoter characterization is the DNA footprinting to find the DNA sequences that bind to particular transcription factors. The method is based in the amplification of the region of interest by polymerase chain reaction (PCR) followed by the labeling of the resulting DNA. The addition of the protein of interest to some of the labeled DNA and the random cutting of the DNA with a cleavage agent in a sequence independent manner, will allow the comparison of the labeled DNA with and without bound protein. The portion of unbound DNA template run in a gel in a ladder-like distribution, meanwhile the DNA template with the protein will result in a ladder distribution with a break in it; the "footprint", where the protein has protected the DNA from the cleavage agent (HAMPSHIRE *et al.* 2007).

To characterize regulatory regions, another useful approach is to investigate genetic variants affecting gene expression. One well-established method is the mapping of expression quantitative trait loci (eQTLs). It works mostly in *cis* close to the target gene, but also identifies eQTLs in *trans* (STRANGER *et al.* 2005, STRANGER *et al.* 2007, STRANGER *et al.* 2007, PICKRELL *et al.* 2010, FEHRMANN *et al.* 2011). This method is available for a wide range of cell types and organisms. However, nowadays, other techniques based in massive

sequencing are maybe more accurate for the study of regulatory sequences, such as identifying nucleosome-depleted DNase I hypersensitive sites by DNase-seq, or analyzing protein interactions with DNA by chromatin immunoprecipitation coupled with sequencing (ChIP-seq).

Chromatin accessible DNase I hypersensitive sites, which were one of the first signatures used in the search of regulatory sequences, have three properties that make them exceptionally helpful for evolutionary analyses of gene expression (BOYLE *et al.* 2008, SONG and CRAWFORD 2010). Firstly, DNase I sensitivity provides a precise and quantitative marker of regions of open chromatin, and is well correlated with a variety of other markers of active regulatory regions, including promoters, enhancers, silencers, insulators and locus control regions (DEGNER *et al.* 2012, SHIBATA *et al.* 2012). Secondly, only active regulatory elements are distinguished by a heightened sensitivity to DNase I cleavage, which means that DNase-seq can be used to identify evolutionary changes in a regulatory element in a specific tissue and at a specific developmental stage. Finally, DNase I hypersensitive sites only represent about 2% of the genome, making it possible to focus the analysis on regions that are involved in transcriptional regulation and ignore regions that are not (NATARAJAN *et al.* 2012, SHIBATA *et al.* 2012).

The study of the chromatin profile provides other efficient way of detecting cis-regulatory elements. Experimentally, chromatin immunoprecipitation (ChIP) methods have revolutionized the study of protein–DNA interactions, opening the possibility for generating genome-wide maps within the cell of interest. Briefly, the technique relies on the treatment of cells with a chemical agent that covalently cross-links any proteins bound to DNA. An antibody is then used to immunoprecipitate and to isolate the protein of interest, consequently also precipitating the DNA fragments bound to that protein. Afterwards, the cross-linking can be reversed by heat and acid hydrolysis, liberating the DNA fragments for analysis. The isolated DNAs can then be identified by hybridization to microarrays, also known as DNA chips (ChIP-chip) or by sequencing (ChIP-Seq) (DOWELL 2010). This technique has been widely use to locate the binding sites of different transcription factors in several cell types, as for example in the ENCODE project (CONSORTIUM *et al.* 2007). In the

last decade, genomic studies with ChIP-chip and ChIP-Seq of specific histone modifications have been of particular interest, providing precise insight into the chromatin state and correlating it with regulator binding, transcriptional initiation and elongation, enhancer activity and repression (BARSKI *et al.* 2007, BIRNEY *et al.* 2007, HEINTZMAN *et al.* 2007, ERNST and KELLIS 2010, ERNST *et al.* 2011).

## 1.4. The thrombospondin family.

Thrombospondins (THBSs, also called TPSs) are multidomain calcium-binding extracellular glycoproteins found throughout the body of vertebrates and lower metazoans (ADAMS and LAWLER 2011, RISHER and EROGLU 2012). Thrombospondins have been well conserved during animal evolution, being present in insects (ADAMS *et al.* 2003), crustacea (ADAMS 2004) and chordates. In vertebrate genomes, five forms of thrombospondin proteins named THBS1-THBS5 (thrombospondin-5 usually known as cartilage oligomeric matrix protein, COMP) are encoded by different genes named also as *THBS1-THBS4* and *THBS5/COMP,* in italics to distinguish proteins from genes. They fall into two groups, termed A and B, according to their oligomerisation status and molecular architecture (ADAMS and LAWLER 1993). Group A thrombospondins (THBS1 and THBS2) form trimers of subunits composed of an N-terminal domain (THBS-N), a coiled-coil oligomerisation domain, a von Willebrand Factor type C (VWC) domain, three thrombospondin repeats (TSRs), three epidermal growth factor (EGF)-like repeats, a calcium-binding wire, and a lectin-like repeat. Group B thrombospondins (THBS3-THBS5) form pentamers, lack VWC modules and TSRs, but have additional EGF-like repeats (ADAMS and LAWLER 2004, CARLSON *et al.* 2008, ADAMS and LAWLER 2011). A representation of the different structures of THBS family members can be found in FIGURE 4.

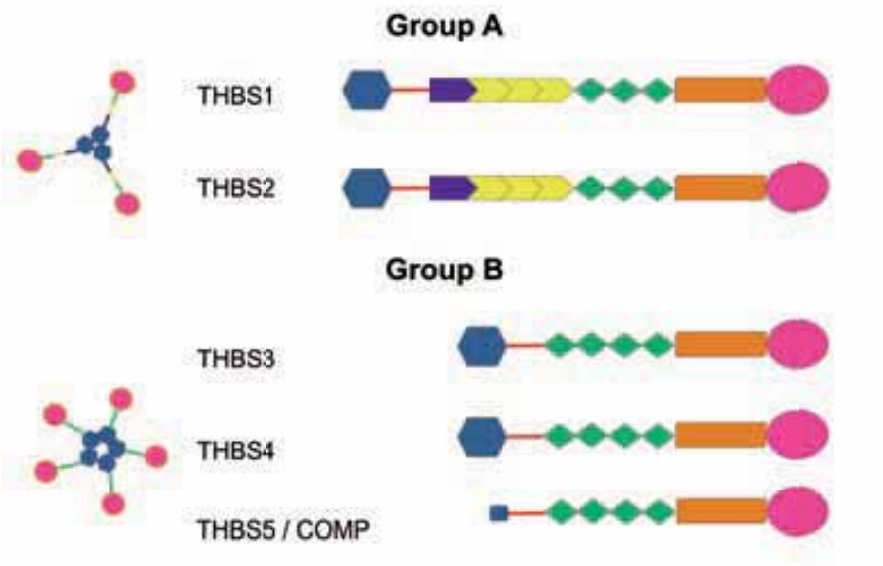



FIGURE 4. Structure of thrombospondin family members. Group A (THBS1 and THBS2) form trimers. Group B (THBS3 to THBS5) form pentamers. The different modules are colored as follows: the N-terminal domain (THBS-N) in blue, the oligomerization coiled-coil domain in red, the von Willebrand Factor type C domain (VWC) in purple, the three thrombospondin repeats (TSRs) of the group A in yellow, the epidermal growth factor (EGF)-like repeats in green, the calcium-binding wire in orange, and the lectin-like module in pink. Figure redrawn from CARLSON *et al*. (2008).

### 1.4.1. Functions of thrombospondins and their implication in the central nervous system.

The prototypic member of the family, described for first time by BAENZIGER *et al.* in 1971 (BAENZIGER *et al.* 1971), is THBS1. It was identified as a glycoprotein released from the $\alpha$-granules platelets in response to stimulation with thrombin, which gave it its name. Thrombospondins modulate transitory or long-term interactions with other extracellular matrix components in different processes of angiogenesis, inflammation, osteogenesis, cell proliferation and apoptosis (BENTLEY and ADAMS 2010, RISHER and EROGLU 2012). Thrombospondins connect and interact with many cellular receptors and other extracellular matrix molecules through their various domains. For example, they bind integrins and calcium through EGF domains, CD36 and transforming growth factors through TSRs, and integrins as well as heparin through the N-terminal domain (BORNSTEIN 2009, MURPHY-

ULLRICH and IOZZO 2012). However, the actions of thrombospondins depend on the expression and availability of their binding partners in the given tissue environment (MUSTONEN *et al.* 2013).

In the nervous system, thrombospondins have been reported to promote neurite growth (ARBER and CARONI 1995). Particularly, within the cerebral cortex, thrombospondins are primarily expressed by astrocytes during the early postnatal development (CAHOY *et al.* 2008, EROGLU 2009). In 2005, CHRISTOPHERSON and colleagues (CHRISTOPHERSON *et al.* 2005) suggested that thrombospondins were able to regulate synapse formation in the developing central nervous system. In previous studies, they discovered that when purified rat retinal ganglion cells were cultured in defined serum-free conditions (appropriate for their survival and growth), they formed few synapses. However, when retinal ganglion cells were cultured with astrocytes, even without direct contact, the condition was sufficient to enhance synapse numbers. These results indicated that one or more soluble signals in the astrocyte-conditioned media existed to mediate these effects. They screened a list of astrocyte-secreted proteins for their ability to induce synapse formation, finding that purified THBS1 and THBS2 (trimeric thrombospondins of group A) significantly increased the number of synapses formed between retinal ganglion cells when added to the cultures. Furthermore, immunodepletion of THBS2 from the astrocyte-conditioned media reduced the media's synaptogenic effect to control levels, concluding that they were necessary and sufficient signals coming from astrocytes to stimulate excitatory synaptogenesis between retinal ganglion cells.

Few years later, EROGLU *et al.* (2009) determined that not only THBS1 and THBS2 were capable of inducing synapse formation, but also the pentameric thrombospondins of group B also had synaptogenic functions (FIGURE 5). They suggested that the epidermal growth factor (EGF)-like domains, common to all thrombospondins present in vertebrates, mediated its synapse-inducing activity via a GABA (gamma amino butyric acid) receptor $\alpha 2\delta$-1, which binds directly to thrombospondins and is required for central nervous system synapse formation. These results complement well with those of ARBER and CARONI (1995), who localized THBS4 in synapse-rich territories of rat brains.

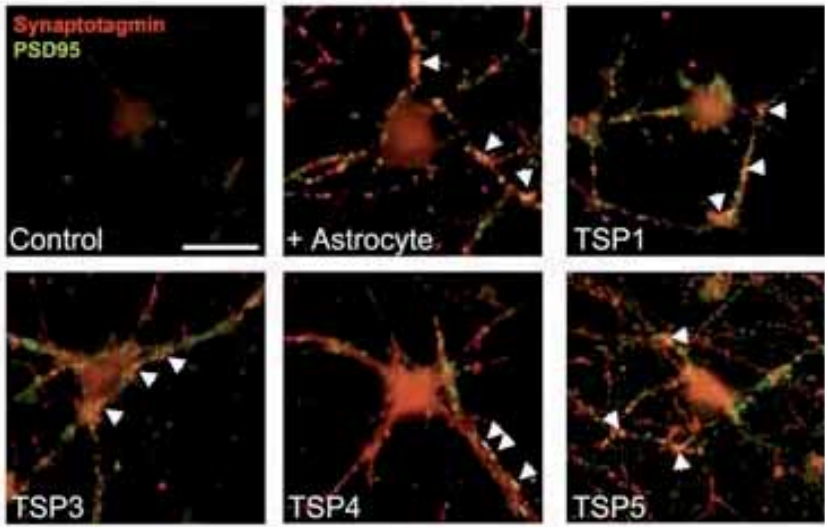



FIGURE 5. Immunostaining of rat retinal ganglion cells. Retinal ganglion cells colocalization of presynaptic synaptotagmin (red) and postsynaptic PSD-95 (green). White arrows point to colocalized synaptic puncta. Figure extracted from EROGLU *et al.* 2009.

Considering the critical roles of thrombospondins in developmental synaptogenesis, it is not surprising to find that many of these matricellular proteins have been shown to be upregulated after central nervous system injury, such as traumatic brain injury (TBI), stroke, epilepsy, and Alzheimer's disease. Moreover, abnormalities in its expression may underlie cortical dysfunction in neurological disorders characterized by aberrant synaptic networks (RISHER and EROGLU 2012). THBS1, THBS2 and THBS4 have been localized in amyloid plaques in the frontal and hippocampal cortex of individuals with Alzheimer's disease and non-demented control cases (BUÉE *et al.* 1992, CÁCERES *et al.* 2007). This does not necessarily implicate them in the pathogenesis of Alzheimer's disease. However, it is possible that one or more molecules that were modified in human evolution enhance Aβ toxicity in humans, and THBS4 and THBS2 could be among those molecules (CÁCERES *et al.* 2007).

## 1.4.2. The thrombospondin-4 gene.

The thrombospondin-4 *(THBS4)* gene was first identified in the African clawed frog (*Xenopus laevis*) genome. Lawler and colleagues (LAWLER *et al.* 1993) showed that *THBS4* is expressed with high levels in adult heart and skeletal muscles, and named it as *THBS4* based on the differences in the tissue distribution with *THBS3*. They also suggested that the THBS4 protein could have been produced by a gene duplication event that occurred 925 million years ago and separated the branches of THBS1 and THBS2 of the THBS3 and THBS4 based on phylogenetic trees constructed from the multisequence alignment of thrombospondin sequences from human, mouse, chicken, and frog. Additionally, a posterior duplication event 644 million years ago could have separated the branches of THBS3 and THBS4 (LAWLER *et al.* 1993). In humans, *THBS4* is localized in chromosome 5, position q13, and encodes a 961 amino acids protein.

### 1.4.2.1. Gene expression analysis in *THBS4*.

Microarray data comparing the adult cortex of humans, chimpanzees and macaques showed that the *THBS4* gene has a manifold higher expression in humans compared to the other species (CÁCERES *et al.* 2003). To determine which tissues and cell types were involved in thrombospondin expression changes during human evolution, CÁCERES *et al.* in 2007 carried out an ample analysis of gene-expression patterns of thrombospondins in humans, chimpanzees, and macaque monkeys. First, they analyzed microarray data available from 6 different cortical and sub-cortical brain regions, showing that the *THBS4* mRNA presented approximately 6-fold higher levels in humans compared to chimpanzees in most cortical areas and the caudate nucleus, all of which belong to the telencephalon, but not in the cerebellum (FIGURE 6A). Similarly, although with a smaller inter-species difference, around 2-fold higher mRNA levels were found for *THBS2* in the same areas. To quantify gene-expression differences more accurately, CÁCERES and collaborators (CÁCERES *et al.* 2007) measured the transcript levels of thrombospondins in the frontal and temporal cortex by real-time RT-PCR from 11 adult humans, 5 chimpanzees, and 10 macaques (5 rhesus

and 5 pigtails). These results showed similar expression patterns as in the microarrays. In the frontal and the temporal cortex of humans there were significantly more *THBS4* and *THBS2* mRNA molecules than in chimpanzees and macaques, but there were not differences between these non-human primate samples (FIGURE 6B). This suggested an evolutionary up-regulation of *THBS4* and *THBS2* in forebrain regions in adult humans.

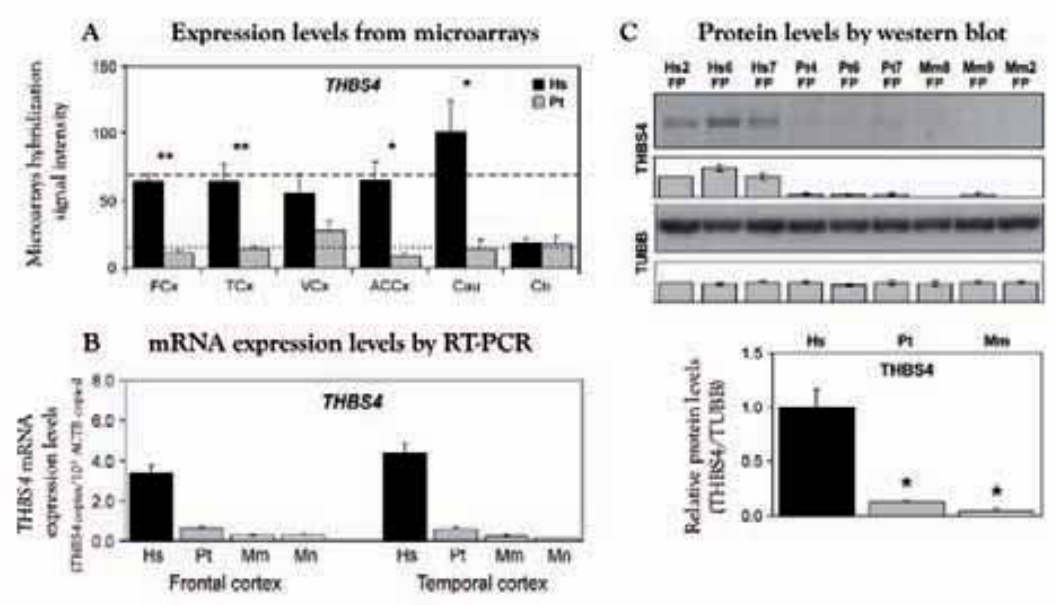



FIGURE 6. Analysis of *THBS4* gene expression between species. A. *THBS4* expression levels from microarrays. Brain regions: FCx, frontal cortex; TCx, temporal cortex; VCx, primary visual cortex; ACCx, anterior cingulate cortex; Cau, caudate nucleus; Cb, cerebellar vermis. B. Quantification of mRNA expression levels of *THBS4* in frontal and temporal cortex of different primate species by real-time RT-PCR. C. Quantification of THBS4 protein levels in primate frontal cortex by Western blot analysis. Species included were humans (Hs), chimpanzees (Pt), rhesus macaques (Mm), and pigtail macaques (Mn). *$P$ = 0.05; **$P$ = 0.01. Figure modified from CÁCERES *et al.* 2007

In addition, CÁCERES and colleagues determined whether these higher mRNA levels resulted in increased protein levels: western blot analysis were carried out with samples from the frontopolar regions of 3 humans, 3 chimpanzees, and 3 rhesus macaques showing that THBS4 and THBS2 protein levels in humans were significantly higher than in non-human primates (FIGURE 6C), in concordance with gene expression analysis. Finally, they analyzed microarray data from heart, kidney, liver, and testis samples to compare expression levels of *THBS4* and *THBS2* between humans and chimpanzees, showing no significant differences in expression levels, except in testis. In addition, both *THBS4* and *THBS2* expression in

heart appeared to be higher in chimpanzees than in humans, contrarily to the results in the other two tissues.

### 1.4.2.2. Cellular localization of *THBS4*.

Originally, *THBS4* expression was mainly detected in large far-projecting neurons and on the cell bodies of Purkinje cells in the adult mouse nervous system (ARBER and CARONI 1995). In humans, *THBS4* has been present in multiple cell types, including endothelial cells and vascular smooth muscle cells from brain blood vessels and coronary arteries (STENINA *et al.* 2003, CÁCERES *et al.* 2007), astrocytes and oligodendrocytes glia cells, or pyramidal neurons in the frontal cortex of humans and nonhuman primates with no clear differences in its distribution. Thus, apparently the *THBS4* upregulation in humans is not due to an increase in the *THBS4* expressing cells between species, but rather to a higher expression within each cell. The major histological difference found between species was the higher density of THBS4 protein labeling in the neuropil of humans, consistent with a role for thrombospondins in synapse formation (CÁCERES *et al.* 2007).

### 1.4.2.3. Known effects of *THBS4* SNPs.

The search for different conformations or changes in the structure and sequence of a gene may help to understand how the gene evolved and its functional consequences. The study of the genetic diversity of *THBS4* in humans has been mostly focused on non-synonymous polymorphisms present in the sequence and the attempt to relate them to known functions of *THBS4* and the possible phenotypic consequences of these variants.

Single-nucleotide polymorphisms (SNPs) affecting the coding regions of *THBS4* have emerged as novel risk factors correlated with familial premature myocardial infarction (MI) (TOPOL *et al.* 2001). The knowledge on thrombospondin-4 in myocardial infarction is limited, but *THBS4* gene expression has been shown to increase highly and chronically after myocardial infarction (MUSTONEN *et al.* 2008, LYNCH *et al.* 2012). Nevertheless, what

happens after a myocardial infarction? It has been shown that the inflammatory cells migrating for the remodeling of the infarct areas produce matrix metalloproteinases (MMPs) capable of degrading the extracellular matrix both in the infarcted zone and in a lesser extent in the non-infarcted zone (JUGDUTT 2003). Reperfusion after myocardial infarction in THBS1-deficient mice resulted in a prolonged inflammatory response in the infarct border zone and extensive cardiac remodeling (FRANGOGIANNIS *et al.* 2005), suggesting that THBS1 serves as a barrier in the infarct border zone limiting the expansion of the inflammatory and fibrotic response into the non-infarcted myocardium (MUSTONEN *et al.* 2013).

The *THBS4* variant responsible for adding this gene to the list of myocardial infarction risk factors is a non-synonymous polymorphism changing an alanine amino acid to a proline at position 387 of the amino acid sequence (Ala387Pro), and particularly changing a guanine nucleotide to a cytosine at position 1186 (1186G/C, rs1866389) located in exon 9 of the gene (TOPOL *et al.* 2001). In contrast to the Ala387 isoform, *THBS4* with Pro387 has been reported to interfere with adhesion properties and proliferation of vascular endothelial cells, and in a process mediated by the THBS4 receptor Mac-1, the Pro387 isoform appeared to be a stronger activator of neutrophil responses than the Ala387 isoform (STENINA *et al.* 2003, PLUSKOTA *et al.* 2005, STENINA *et al.* 2005, CORSETTI *et al.* 2011). This association between the SNPs in the *THBS4* gene and myocardial infarction has not been consistently corroborated. Depending on the study, the CC genotype of the SNP in *THBS4* has been implicated with a higher or lower risk of myocardial infarction than the GG genotype (CUI *et al.* 2004, WESSEL *et al.* 2004, CUI *et al.* 2006). In other studies, the GC genotype of the SNP in *THBS4* has been connected with myocardial infarction (TOPOL *et al.* 2001) or an association has not been found (KOCH *et al.* 2008)

## 1.5. Objectives

With the development of techniques that allow large-scale comparisons of gene expression levels, it became possible to identify hundreds of genes differentially expressed in different brain and other tissues of humans, chimpanzees and other non-human primate species (CÁCERES *et al.* 2003, KHAITOVICH *et al.* 2005, XU *et al.* 2010, HU *et al.* 2011, LIU *et al.* 2011, PAI *et al.* 2011). The majority of these studies suggested that gene expression changes in the human cortex implicated predominantly increased expression, and that many of the genes up-regulated in humans could be related to higher levels of neuronal activity (CÁCERES *et al.* 2003, PREUSS *et al.* 2004, KHAITOVICH *et al.* 2006, SOMEL *et al.* 2009, BABBITT *et al.* 2010) or specific brain functions (ZHANG *et al.* 2011). However, they also generated a big list of candidate genes that could be responsible of some of the human brain uniqueness. Therefore, the detailed characterization of each of these genes should be considered to gain insight in the specific genetic changes involved and to determine the role that natural selection may have had in their fixation.

This work is focused on the *THBS4* gene. As suggested in section 1.4 and subsections within, we have put special attention into this calcium-binding extracellular glycoprotein because microarray data and mRNA and protein expression validation experiments showed that the *THBS4* gene has several times higher expression in the adult cortex of humans than in chimpanzees and macaques (CÁCERES *et al.* 2003, CÁCERES *et al.* 2007). In addition, its role in synapse formation makes it a very interesting candidate for human brain function. The main objective of this work was to use the *THBS4* gene as a model to increase the knowledge about the mechanisms implicated in gene expression regulation at the genomic level in general, and specifically in gene expression changes in human brain evolution. The following specific goals were to be achieved:

1. Characterize the possible *THBS4* mRNAs to determine their possible transcription start site/s.

Preliminary analysis of the human genome sequence revealed a putative TSS for *THBS4* upstream the reference TSS. Amplification and sequencing of the possible mRNA will be carried out for its validation.

2. Study the expression patterns of both isoforms in different human cell types and tissues.

Differential expression between the two *THBS4* isoforms could suggest the specialization of each of them to carry out specific functions in different tissues and would be important to determine which of the isoforms is predominantly expressed in the brain. We will perform real-time RT-PCR to quantify diverse non-brain tissues, brain tissues (adult and fetal tissues) and few human cell lines.

3. Measure the expression differences between species of both isoforms.

How these two isoforms are expressed in humans and non-human primate species, such as chimpanzees and rhesus macaques, could give insight to explain the increased *THBS4* gene expression observed in the human brain. To quantify the expression of both mRNAs, we will perform real-time RT-PCR from the frontal cortex of multiple humans, chimpanzees and macaques.

4. Search the possible causes regulating *THBS4* differential expression.

Several causes can be regulating the increased expression of *THBS4* in the human brain. We will investigate possible sources of the expression differences such the genomic context in which *THBS4* is located and the structural changes that could have occurred in that region between species, changes in each promoter region affecting their transcriptional activity in humans and chimpanzees, specific sequences for transcription factor binding sites, differences in the promoter methylation affecting differentially both species, or changes in potential enhancers that could be increasing the expression of *THBS4.*

5. Evaluate the variation of expression levels of ***THBS4*** within human populations.

The study of the *THBS4* variation within humans could help to understand how the gene is regulated. We will quantify the *THBS4* expression of multiple adult brain samples and correlate the levels with the different haplotypes present for *THBS4* within human populations.

6. Get insights about ***THBS4*** promoters evolution.

Data about genome conservation available online could provide information about how conserved are both *THBS4* isoforms between differences species and which of them is more antique. We will search for possible evidences that help to understand how or when *THBS4* promoters had evolved.

# 2

# MATERIALS AND METHODS


– JORGE CHAM, *Piled Higher and Deeper* (2013) –

# 2

# MATERIALS AND METHODS

## 2.1. Samples.

The molecular characterization of thrombospondin-4 (*THBS4)* expression has required many samples from human and non-human primates. The use of different tissues and cell lines provided the flexibility to extract and obtain all the nucleic acids required, both DNA or RNA. In addition, as cell lines can be grown *in vitro*, they were used for transfection with different plasmid DNA and for chromatin immunoprecipitation (which requires big amounts of sample).

The primary function of the total RNA samples were the retro-transcription into complementary DNAs (cDNAs), which, in turn, were used for gene expression detection and quantification in different tissues and species. Plasmid DNAs were generated from the cloning of genomic DNA from humans and chimpanzees in pGL3 vectors for transcription activity assays. Genomic DNA was also modified and cloned in pGEM-T vector to study the methylation levels in a specific region.

### 2.1.1. Commercial RNAs.

To determine how *THBS4* is expressed in different tissues in the human body we took advantage of total RNAs available from different companies (TABLE 3), which offered an extensive variety of total RNAs isolated from many tissue types.

TABLE 3. **Commercial total RNAs from humans used in this study.** Summary table including: company, catalog number, origin, concentration, lot, sex, age and information about sample's treatment with DNase I.

| Company | Catalog Number | Origen | Conc. (µg/µl) | Lot | Sex | Age | DNA free |
|---|---|---|---|---|---|---|---|
| Stratagene* | 540009 | Colon | 1.00 | 1140402 | Female | 53 years | No |
| Stratagene* | 540011 | Heart | 0.90 | 1140403 | Female | 63 years | No |
| Stratagene* | 540013 | Kidney | 0.90 | 0760025 | Female | 67 years | No |
| Agilent Technologies | 540017 | Liver | 0.90 | 1140405 | Female | 34 years | No |
| Stratagene* | 540025 | Placenta | 1.00 | 0850456 | Female | 28 years | No |
| Stratagene* | 540035 | Spleen | 0.90 | 1140582 | Female | 38 years | No |
| Stratagene* | 540049 | Testes | 1.10 | 0760573 | Male | 72 years | No |
| Stratagene* | 540071 | Ovary | 1.00 | 1240179 | Female | 30 years | No |
| Agilent Technologies | 540141 | Thymus | 0.90 | 1240354 | Male | 45 years | No |
| Agilent Technologies | 540020 | Lung | 0.90 | 0350707 | unknown | unknown | No |
| Agilent Technologies | 540029 | Skeletal Muscle | 0.9 | 50540029 | Female | 71 years | No |
| Ambion** | AM6768 | Caudate Nucleus | 1.00 | 10030213 | Male | 86 years | Yes |
| Ambion** | AM6870 | Hippocampus | 1.00 | 4050016 | Male | 23 years | Yes |
| Ambion** | AM6762 | Thalamus | 1.00 | 10030263 | Male | 86 years | Yes |
| Ambion** | AM6778 | Globus Pallidus | 1.00 | 4050072 | Male | 23 years | Yes |
| Amsbio | R1244051-50 | Fetal Brain. Frontal Lobe | 2.15 | B111145 | Male | 20 weeks | Yes |
| Amsbio | R1244049-50 | Fetal Brain. Diencephalon | 3.39 | A706020 | Male | 22 weeks | Yes |
| Amsbio | R1244040-50 | Fetal Brain. Cerebellum | 1.27 | B111158 | Male | 36 weeks | Yes |

* Currently Agilent Biotechnologies

** Currently Life Technologies.

## 2.1.2. Cell lines.

Cells derived from tumors frequently proliferate indefinitely in culture and are referred to as immortal cell lines. These have been particularly useful for many types of experiments, because they provide a continuous source of cells and are a good model to simulate *in vivo* conditions and functions. In this study, human cell lines (TABLE 4) have been used to isolate nucleic acids for subsequent experiments, for plasmid DNA transfections, and for chromatin immunoprecipitation.

TABLE 4. Human cell lines used in this study.

| Cell type | Name | Tissue (disease) | Sex | Age | Ethnicity | Growth mode | Culture media* |
|---|---|---|---|---|---|---|---|
| Neuroblast | SK-N-AS | Brain (bone marrow metastasis) | F | 6 years | Caucasian | Adherent | DMEM +10% FBS |
| Neuroblast | SH-SY5Y | Brain (bone marrow metastasis) | F | 4 years | Unknown | Adherent | DMEM +10% FBS |
| Glia | T98G | Brain (glioblastoma multiforme) | M | 61 years | Caucasian | Adherent | DMEM +10% FBS |
| Glia | U87 | Brain (glioblastoma. astrocitoma) | M | 44 years | Caucasian | Adherent | DMEM +10% FBS |
| Kidney | HEK293 | Embryonic kidney | F | Fetus | Unknown | Adherent | DMEM +10% FBS |
| Epithelial | Caco-2 | Colon (colorectal adenocarcinoma) | M | 72 years | Caucasian | Adherent | MEM +10% FBS |
| Epithelial | HeLa | Cervix (adenocarcinoma) | F | 31 years | African | Adherent | MEM +10% FBS |
| B Lymphocyte | NA 10856 | Blood | M | 46 years | Caucasian | Suspension | MEM +10% FBS |
| B Lymphocyte | NA 10860 | Blood | M | 50 years | Caucasian | Suspension | MEM +10% FBS |
| B Lymphocyte | NA 12057 | Blood | F | 75 years | Caucasian | Suspension | MEM +10% FBS |
| B Lymphocyte | NA 12872 | Blood | M | Unk | Caucasian | Suspension | MEM +10% FBS |

F: female, M: male, Unk: unknown, DMEM: Dulbecco's modified eagle medium, MEM: minimum essential medium, FBS: fetal bovine serum

### 2.1.3. Tissues.

Tissue samples provide the possibility to extract all kinds of nucleic acids needed for gene expression analysis. However, tissues samples are not easy to obtain, and when they are available, ideal conditions where subjects would be matched by age, sex, cause of death, post-mortem delay, etc. are even more difficult to accomplish. Interspecies studies like this one, aiming at the comparison of humans *versus* other primates, are faced with the additional challenge of obtaining appropriate non-human primate material.

In this study we took advantage of the sample collection available at the laboratory of Professor Todd M. Preuss at the Yerkes National Primate Research Center of Emory University (Atlanta, GA, USA), the laboratory of Professor Xavier Estivill at the Centre for Genomic Regulation (Barcelona, Spain) and in our own laboratory at the Institut de Biotecnologia i de Biomedicina of the Universitat Autonòma de Barcelona. Information about origin, species, age, sex, cause of death, time post mortem, type of tissue and usage of the samples is available in TABLE 5. The Preuss laboratory human and non-human primate tissue samples were originally provided, in turn, from different research centers and universities. Human tissues came from the University of Maryland Brain and Tissue Bank for Developmental Disorders (UMBTB, Hs1-Hs11) and the Alzheimer's Disease Research Center of Emory University (ADRC, Hs12-Hs14). Chimpanzee and macaque samples were obtained from the New Iberia Research Center (NIRC, Pt4 and Mm1-Mm3) and the Yerkes National Primate Research Center (YNPRC, Pt5-Pt22 and Mm10-Mm22). Human samples (Hs30-Hs37) from Estivill laboratory came from the University Hospital of Bellvitge (HUB). The chimpanzee (Pt23) and the gorilla (Gg1-Gg2) samples available in the Cáceres laboratory were obtained from the Animal Tissue Bank of Catalunya (BTAC) of the Universitat Autonòma de Barcelona. All chimpanzee, macaque and gorilla samples used in this thesis project came from animals that died of natural causes or were euthanized for medical reasons. Human brain samples were obtained from individuals who died of causes unrelated to neurological disorders. Case samples were named based on previous publications by CÁCERES *et al.* (2003 and 2007).

TABLE 5. Brain tissue samples from different species used in this study.

| Case | Institution id. | Species | Source | Sex | Age | Cause of death / Disease process | PMI (hours) | Tissue preservation | Tissue analyzed | Use |
|---|---|---|---|---|---|---|---|---|---|---|
| Hs1 | 813 | *H. sapiens* | UMBTB | F | 30 years | Accident, multiple injuries | 14 | Frozen | FP | A.1. A.3. B. C. |
| Hs2 | 1029 | *H. sapiens* | UMBTB | M | 31 years | Accident, multiple injuries | 12 | Frozen | FP | A.1. A.3. B. C. |
| Hs6 | 1863 | *H. sapiens* | UMBTB | F | 40 years | Accident, multiple injuries | 7 | Frozen | FP | A.1. A.3. C. |
| Hs7 | 1903 | *H. sapiens* | UMBTB | M | 42 years | Cardiovascular disease | 7 | Frozen | FP | A.1. A.3. B. C. |
| Hs8 | 1134 | *H. sapiens* | UMBTB | M | 48 years | Cardiovascular disease | 15 | Frozen | FP | A.1. A.3. B. C. |
| Hs9 | 1570 | *H. sapiens* | UMBTB | M | 42 years | Cardiovascular disease | 14 | Frozen | FP | A.1. A.3. C. |
| Hs10 | 1673 | *H. sapiens* | UMBTB | M | 47 years | Cardiovascular disease | 11 | Frozen | FP | A.1. A.3. B. C. |
| Hs11 | 1832 | *H. sapiens* | UMBTB | M | 75 years | Cardiovascular disease | 15 | Frozen | FP | A.1. A.3. C. |
| Hs12 | OS02-35 | *H. sapiens* | ADRC | F | 74 years | Terminal cancer | 6 | Frozen | FP | A.1. A.3. C. |
| Hs13 | OS99-08 | *H. sapiens* | ADRC | F | 87 years | Lung cancer | 3 | Frozen | aFCx | A.1. A.3. C. |
| Hs14 | OS03-394 | *H. sapiens* | ADRC | F | 30 years | Unknown | 5 | Frozen | FP | A.1. A.3. C. |
| Hs30 | A02-15 | *H. sapiens* | HUB | F | 49 years | Unknown | 7 | Frozen | aFCx | A.3. C |
| Hs31 | A02-98 | *H. sapiens* | HUB | M | 79 years | Unknown | 7 | Frozen | aFCx | A.3. C |
| Hs32 | A03-25 | *H. sapiens* | HUB | F | 65 years | Unknown | 4 | Frozen | aFCx | A.3. C |
| Hs33 | A03-34 | *H. sapiens* | HUB | M | 53 years | Unknown | 3 | Frozen | aFCx | A.2. A.3. C |
| Hs34 | A03-62 | *H. sapiens* | HUB | M | 62 years | Unknown | 3 | Frozen | aFCx | A.3. C |
| Hs35 | A04-162 | *H. sapiens* | HUB | M | 67 years | Unknown | 5 | Frozen | aFCx | A.3. C |
| Hs36 | A05-5 | *H. sapiens* | HUB | M | 58 years | Unknown | 4 | Frozen | aFCx | A.3. C |
| Hs37 | A05-38 | *H. sapiens* | HUB | M | 78 years | Unknown | 2 | Frozen | Cb | A.2 |
| Pt4 | 86A005 | *P. troglodytes* | NIRC | F | 16 years | Lymphoma, euthanized | 3.5 | Frozen | FP | A.1. B |
| Pt5 | Yn03-09 | *P. troglodytes* | YNPRC | F* | 21 years | Liver failure | 1 | Frozen | FP | A.1. B |
| Pt6 | Yn04-04 | *P. troglodytes* | YNPRC | M | 27 years | Myocardial fibrosis | 4 | Frozen | FP | A.1. B |
| Pt7 | Yn04-30 | *P. troglodytes* | YNPRC | M | 24 years | Congest heart failure | 3 | Frozen | FP | A.1. B |
| Pt8 | Yn05-12 | *P. troglodytes* | YNPRC | M | 28 years | Myocardial fibrosis | 8 | Frozen | aFCx | A.1 |

(Continuation TABLE 5)

| Case | Institution id. | Species | Source | Sex | Age | Cause of death / Disease process | PMI (hours) | Tissue preservation | Tissue analyzed | Use |
|---|---|---|---|---|---|---|---|---|---|---|
| Pt17 | Yn06-108 | *P. troglodytes* | YNPRC | F | 44 years | Atherosclerosis | 0.5 | Frozen | FP | A.1 |
| Pt18 | Yn06-147 | *P. troglodytes* | YNPRC | M | 43 years | Myocardial fibrosis | 2.5 | Frozen | FP | A.1 |
| Pt19 | Yn07-25 | *P. troglodytes* | YNPRC | F | 47 years | Unknown | 1 | Frozen | FP | A.1 |
| Pt20 | Yn07-147 | *P. troglodytes* | YNPRC | M | 18 years | Myocardial fibrosis | 3 | Frozen | FP | A.1. B |
| Pt21 | Yn07-387 | *P. troglodytes* | YNPRC | M | 40 years | Myocardial fibrosis | 2.5 | Frozen | FP | A.1 |
| Pt22 | Yn09-438 | *P. troglodytes* | YNPRC | F | 40 years | Unknown, euthanized | < 1 | Frozen | FP | A.1 |
| Pt23 | N457/03 | *P. troglodytes* | BTAC | M | 6 years | Respiratory failure | 5 | Frozen | FP | D |
| Mm1 | V344 | *M. mulata* | NIRC | F | 4 years | Moribund, euthanized | 1.5 | Frozen | FP | A.1 |
| Mm2 | V401 | *M. mulata* | NIRC | M | 5 years | Moribund, euthanized | 2 | Frozen | FP | A.1 |
| Mm3 | 96O087 | *M. mulata* | NIRC | F | 9 years | Moribund, euthanized | 2 | Frozen | FP | A.1 |
| Mm10 | Yn04-03 | *M. mulata* | YNPRC | M | 21 years | Enteritis, euthanized | 2 | Frozen | FP | A.1 |
| Mm11 | Yn04-05 | *M. mulata* | YNPRC | F | 19 years | Cancer | 1 | Frozen | FP | A.1 |
| Mm20 | Yn04-21 | *M. mulata* | YNPRC | M | 12 years | Unknown, euthanized | 3 | Frozen | FP | A.1 |
| Mm21 | Yn08-308 | *M. mulata* | YNPRC | F | 4 years | Eye study, euthanized | 2.5 | Frozen | FP | A.1 |
| Mm22 | Yn08-309 | *M. mulata* | YNPRC | F | 4 years | Eye study, euthanized | 2.5 | Frozen | FP | A.1 |
| Gg1 | Z01/03 | *G. gorilla* | BTAC | M | 25 years | Appendicitis | 6 | Frozen | aFCx | D |
| Gg2 | Z02/03 | *G. gorilla* | BTAC | M | 40 years | Adenocarcinoma, euthanized | 15 | Frozen | aFCx | D |

UMBTB: University of Maryland Brain and Tissue Bank for Developmental Disorders; ADRC: Alzheimer's Disease Research Center; NIRC: New Iberia Research Center; YNPRC: Yerkes National Primate Research Center; HUB: University Hospital of Bellvitge; BTAC: Animal Tissue Bank of Catalunya.
Use: (A) Real-time RT-PCR: (A.1) *THBS4* primate expression, (A.2) *THBS4* expression in humans, (A.3) *THBS4* allele-specific expression. (B) Bisulfite-based methylation analysis. (C) Pyrosequencing. (D) Analysis of gorilla inversions. FP. frontal pole; aFCx. anterior frontal cortex; Cb. cerebellum; PMI. post-mortem interval

### 2.1.3.1. Tissue collection.

One of the difficulties for having tissue samples in the laboratory is the fact that they usually have to be acquired at necropsy. In collaboration with Dr. Francesc Alameda from the Anatomo-Patological service at Hospital del Mar (Barcelona, Spain), we had the opportunity to collect human tissue samples from nine different subjects (TABLE 6). For each tissue of interest, 2-4 cm of sample were cut after necropsy and placed in a zip-lock plastic bag labeled with the case number and the tissue. Considering that the ultimate use of the tissue was genetic work only, samples bags were immediately frozen in liquid nitrogen (wrapped in aluminum foil) until they were stored inside a -80º C freezer to avoid RNA degradation.

TABLE 6. Tissues samples collected from human necropsy.

| Institution id. | Specie | Source | Sex | Age (years) | Cause of death / Disease process | PMI (hours) | Tissue preservation | Tissues collected | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | FP | TP | VCx | CbV | Thy | Lng | Liv | Pan | Spl | Kid | Hrt | Test | Ova |
| Hs10-01 | *H.sapiens* | HDM | M | 67 | Diabetic. Renal failure. | <12 | Frozen | | | | | X | X | X | X | X | X | X | X | |
| Hs10-02 | *H.sapiens* | HDM | F | 87 | Renal and heart failures | 4 | Frozen | X | X | X | X | X | X | X | X | X | X | X | | |
| Hs10-03 | *H.sapiens* | HDM | M | 66 | Multi-organ failure | <24 | Frozen | X | X | X | X | X | X | X | X | X | X | X | X | |
| Hs10-04 | *H.sapiens* | HDM | M | 81 | Cardiovascular disease | <12 | Frozen | X | X | X | X | X | X | X | X | X | X | X | X | |
| Hs10-05 | *H.sapiens* | HDM | M | 76 | Peritonitis | <12 | Frozen | | | | | X | X | X | X | X | X | | X | |
| Hs10-06 | *H.sapiens* | HDM | F | 60 | Ovary cancer | <24 | Frozen | | | | | X | X | X | | X | X | X | | |
| Hs10-07 | *H.sapiens* | HDM | F | 85 | Pulmonary embolism | <12 | Frozen | X | X | X | X | X | X | X | X | X | X | X | | X |
| HS10-08 | *H.sapiens* | HDM | M | 69 | Cardiovascular disease | 16 | Frozen | X | X | X | X | X | X | X | X | X | X | X | X | |
| Hs10-09 | *H.sapiens* | HDM | M | 67 | Pulmonary embolism | 8 | Frozen | X | X | X | X | X | | X | X | X | X | X | X | |

HDM: Hospital del Mar; PMI: post-mortem interval; FP: frontal pole; TP: temporal pole; VCx: visual cortex; CbV: cerebellar vermis; Thy: thymus; Lng: lung; Liv: liver; Pan: pancreas; Spl: spleen; Kid: kidney; Hrt: heart; Test: testis; Ova: ovary.

## 2.2. Nucleic acid isolation.

Total RNA was isolated from cell lines and tissues by using the miRNeasy Mini Kit (Qiagen), which enables the purification of total RNA including small RNA. 100-200 mg of each tissue sample were collected from frozen tissue on a cold surface to conserve RNA integrity. Tissue homogenization was performed with POLYTRON® PT 2100 (Kinematica) directly in the QIAzol Lysis Reagent to avoid RNA degradation. Depending on the toughness and size of the sample, 2-5 disruption cycles of 15 sec at maximum speed were done (waiting 2-3 seconds in between) to avoid samples from reaching excessively high temperatures. Cell lines were grown in a monolayer in 100 mm Petri dishes and lysed in place when 85-90% confluence was reached. For each type of sample, 1400 μl of QIAzol Lysis Reagent were added. This amount doubles the one recommended by the manufacturer, but a higher volume of lysis reagent helps during the homogenization of tissues, and does not affect the final concentration of cell-line RNA. Homogenized lysates were divided in 2 different tubes before continuing with the RNA extraction. During isolation, total RNA was treated with the RNase-Free DNase Set (Qiagen) to ensure that all traces of DNA that could interfere in the experiment were removed. At the end, 2 vials with 30 μl of RNA were obtained from each initial sample. Concentration and quality of the different RNAs were quantified with a Nanodrop spectrophotometer (Thermo Scientific) and/or a Qubit Fluorometer (Life Sciences) and checked by gel electrophoresis through 1% agarose gel stained by ethidium bromide.

Genomic DNA was extracted following an optimized protocol derived from SAMBROOK *et al. (*1989) and RAPLEY (2000). Briefly, 80 mg of frozen tissue was collected on a cold surface (even that genomic DNA is not as susceptible to degradation as RNA, tissues should be manipulated with caution to conserve RNA integrity for subsequent experiments). Tissues were disrupted at room temperature with a pestle in 1 ml of 0.1% SDS lysis buffer. To eliminate any RNA residue, samples were incubated for 1 h at 37 ºC with a final concentration of 50 ng/μl of RNase A (Life Technologies). Complete homogenization was ensured by leaving the sample over night with a final concentration of

100 μg/ml of proteinase K at 56ºC with gentle agitation. To purify the genomic DNA, several steps are often necessary to inactivate and/or remove enzymes or other proteins that derive from cell extracts. This removal of proteins was carried out by extracting the aqueous solution of nucleic acids with phenol, phenol-chloroform and phenol-chloroform isoamyl alcohol. After purification, genomic DNA was precipitated with ethanol and diluted in 400 μl of ultrapure water. Extracted DNA was quantified with a Nanodrop spectrophotometer (Thermo Scientific) and/or a Qubit Fluorometer (Life Sciences), and checked by gel electrophoresis through a 1% agarose gel stained by ethidium bromide.

Plasmid DNA transformed in a bacterial strain (see point 2.5 of MATERIALS AND METHODS) was grown in 50 ml standard liquid Luria Bertani (LB) medium for 16h. Plasmid DNA extraction was performed using the QIAGEN Plasmid Midi Kit combined with EndoFree Plasmid Buffer Set (Qiagen). This is based on an alkaline lysis system, followed by the selective binding by gravity flow to an anion-exchange resin, allowing purification of plasmid DNA from RNA, proteins, dyes, bacterial lysates and low molecular weight impurities. Finally, purified plasmid DNA was redissolved in a suitable volume of endotoxin-free TE (Tris-EDTA) buffer. Moreover, when small amounts of plasmid DNA were needed, and the final use did not require an endotoxin free sample, QIAprep Miniprep Kit was used to perform the extraction. Plasmid DNA purified with QIAprep Miniprep Kit allows a faster purification of up to 20 μg of high-copy plasmid DNA from 2.5 ml overnight cultures in LB medium. Phenol extraction and ethanol precipitation are not required with this kit, being able to elute the plasmid DNA with water directly from the QIAprep spin column.

## 2.3. Cell-line culture and media.

Cells were grown in a mycoplasma free laboratory under strict aseptic conditions. All cell culture work was carried out in a class II vertical laminar-flow biological safety cabinet that provides personnel, environment and product protection. The cell lines used, their

sources, and their basal media requirements are listed in TABLE 4 (see section 2.1.2 Cell lines in MATERIALS AND METHODS).

Most common media for growing the cell lines were DMEM (Dulbecco's Modified Eagle Medium), which is a widely used basal medium for supporting the growth of many different mammalian cells, and MEM (Minimum Essential Medium), which can be also used with a variety of suspension and adherent mammalian cells. A 10% of fetal bovine serum (FBS) was added to the cell culture to provide growth factors, vitamins, amino acids, etc. to the cells. Serum was stored frozen to preserve the stability of components, such as growth factors. Prior to use in the cell growth medium, the fetal bovine serum was heated for 30 min at 56 °C in a water bath to destroy heat-labile complement proteins. To prevent cell cultures from bacterial or fungal contamination, an antibiotic and antimycotic mix with penicillin, streptomycin, and fungizone (PSF, Gibco) was added to the media (1%).

Cell lines were generally maintained in 25 $cm^2$ and 75 $cm^2$ flasks (Corning or NUNC), containing 10 ml or 15 ml of medium respectively. Medium was changed every two to three days. Prior to subculture, cells were always checked for any contamination. Adherent cells were washed with Phosphate Buffered Saline (PBS) and then removed from their flasks by enzymatic detachment with trypsin. Trypsin reaction was stopped with DMEM (or MEM) media, centrifugated and resuspended in new media. Cells grown in suspension were centrifugated to allow washing with PBS, and after a second centrifugation to remove PBS, cells were resuspended in new media. When a specific number of cells per milliliter were required, cells were counted with a Neubauer haemocytometer.

## 2.4. RT-PCR and PCR.

Reverse transcription polymerase chain reaction (RT-PCR) was used to analyze expressed genes by reverse transcribing the RNA of interest into its complementary DNA

(cDNA) through the use of a reverse transcriptase. Subsequently, the cDNA was amplified using traditional PCR or real-time PCR.

cDNA was synthesized from 1 µg of the DNase I-treated RNA with SuperScript® III First-Strand Synthesis System (Invitrogen). If a commercial RNA data sheet did not have detailed information on whether the sample was DNA free, 2.5 µg of total RNA was treated with DNase I (DNA-free™ Kit, Ambion) for 30 min at 37 °C previous to the cDNA synthesis. This step degraded double- and single-stranded DNA and chromatin that could be contaminating the RNA.

During the initial experiments, only oligo(dT) primer, which is used to hybridize to the 3'-poly(A) tail of the mRNA molecules, was added to the reaction. However, all the RNA molecules without poly(A)-tails were not transcribed and there was a bias towards the 3' ends of the mRNAs, with no good representation of the 5' ends. To avoid this, a mix with oligo(dT) primer and random hexamers was used in all the subsequent final experiments. Retrotranscriptase negative reactions were obtained from 0.2 µg of each sample to exclude the presence of contaminant DNA, which would have been amplified by PCR, giving false positive results.

To accomplish the cDNA (RT-PCR) or DNA (PCR) amplification, different polymerases were used according to the amplification length, GC content, and need for proofreading activity, which included basic Taq polymerase (Biotherm or Roche), AmpliTaq Gold® 360 DNA polymerase (Life technologies), Expand High Fidelity DNA polymerase (Roche) or Phusion® High-Fidelity DNA Polymerase (New England BioLabs). Most commonly used reactions were performed in a total volume of 25 µl including 10 pmol of each primer (sequences available in TABLES 7 and 9 to 14 in MATERIALS AND METHODS and TABLE 15, 21 and 23 in RESULTS), 200 µM nucleotides (dNTPs), 1.5 mM MgCl2, 1.5 units of *Taq* DNA polymerase, and 50-100 ng of cDNA or genomic DNA. PCR cycling steps were typically performed by a initial denaturation step of 7 min at 94ºC, 35 rounds of 30 sec at 94ºC for denaturalization, 30 sec of annealing at 55-65ºC depending on the primer pair used, and 30-120 sec for elongation at 72ºC depending on template length.

Primers were designed using the Primer 3 software (KORESSAAR and REMM 2007, UNTERGASSER *et al.* 2012) and checked against the human SNP database to avoid nucleotide changes between individuals. Those to be used for amplification in other species were compared against the corresponding chimpanzee or rhesus genome assemblies to avoid sequence changes between species. To visualize the PCR results, ~10 µl of the PCR product were electrophoresed in a 1-2% agarose gel and stained in ethidium bromide, which was used as a fluorescent tag for detection under ultraviolet light.

## 2.5. Cloning and transformation.

DNA cloning is a method in which a specific fragment of cDNA or genomic DNA is inserted into a plasmid vector and incorporated into cultured bacteria host cells. Then a large number of identical DNA molecules can be obtained from its replication during growth of the host cells. Generally, a plasmid vector contains at least three elements: a multiple cloning region where the foreign DNA fragment can be inserted; an antibiotic-resistance gene, which blocks or inhibits antibiotics to allow selective growth of the host cell; and a replication origin to allow the plasmid to replicate in the host cell.

### 2.5.1. Cloning into pGL3 Vectors.

pGL3 Luciferase Reporter Vectors (Promega) were used for the quantitative analysis of the transcriptional activity of promoter isoforms and enhancers potentially regulating *THBS4* gene expression. These pGL3 vectors contain a modified coding region for firefly (*Photinus pyralis*) luciferase that has been optimized for monitoring transcriptional activity in transfected eukaryotic cells. To evaluate the transcriptional activity of both *THBS4* promoters using luciferase assays, segments of 1.2 and 6 kb upstream of the reference

promoter and 3 kb upstream of the alternative promoter were selected and cloned in pGL3 plasmids (see section 2.8 Luciferase assay in MATERIALS AND METHODS).

Plasmid constructions of the reference isoform promoter were obtained previously by Mario Cáceres. Briefly, due to problems with the PCR amplification, ~6 kb fragments corresponding to the reference promoter and first exon of *THBS4* were isolated from human BAC CTD-2211I018 and chimpanzee BAC RP43-41P12 by restriction digestion with *MunI* and gel-purification. These *MunI* fragments were cloned in a pCR 2.1 Topo vector (Invitrogen), previously linearized with the restriction enzyme *EcoRI*, which generates compatible ends. From these plasmids, using a *KpnI* restriction site from the vector, *KpnI-NaeI* fragments of 6.0 kb from humans and 5.7 kb from chimpanzees were cloned in an empty pGL3 Basic vector and pGL3 Enhancer vector linearized with *KpnI* and *SmaI*, resulting respectively in plasmids pMCL-007 and pMCL-011 (human fragment) and pMCL-008 and pMCL-012 (chimpanzee fragment). The cloned fragments included the upstream region of the *THBS4* reference isoform to the beginning of the first exon. In addition, a smaller 1.2 kb *SacI-NaeI* fragment from each species, corresponding to the region closest to the reference *THBS4* transcription start site (TSS), was cloned in a pGL3 Basic vector linearized with *SacI* and *SmaI*, generating human plasmid pMCL-009 and chimpanzee plasmid pMCL-010.

To generate the plasmid constructions of the alternative *THBS4* promoter, 3 kb upstream the predicted TSS (including the TSS itself), were amplified from 2 humans with different haplotype sequences in this region (HapMap samples NA10856 and NA12057) and 1 chimpanzee (BAC clone RP4-341P12). Primers were designed including a 5' tail of 12 nucleotides with the restriction enzyme site for *XhoI* and extra bases to permit the enzyme-DNA binding (TABLE 7). PCRs were carried out in 50 µl reactions with the Expand High Fidelity PCR System (Roche) to avoid introduction of sequence errors by the polymerase during PCR amplification.

TABLE 7. Sequence and combination of primers used in quantitative analysis of the transcriptional activity of promoter isoforms. Colored base pairs correspond with an extra 5' tail added to the primer with the restriction enzyme site for *XhoI* (blue) and extra bases to permit the enzyme-DNA binding (orange). Hs: human; Pt: chimpanzee

| Primer name | Sequence (5'-> 3') | Amplicon (bp) | Gene/Region |
|---|---|---|---|
| Alt.THBS4. Luc-3kb-Fw | ACACACCTCGAGACTCCTTGGCATCTCTGTATATC | 3136 (Hs) 3142 (Pt) | Upstream alternative *THBS4* |
| Alt.THBS4. Luc-3kb-Rv | ACACACCTCGAGGCCAAGCTGCTCCTAAGAGTT | | |

PCR products were purified with the QIAquick PCR Purification Kit (Qiagen) to eliminate excess primers and dNTPs from the PCR. 5 µg of pGL3-Basic or pGL3-Enhancer Vector (Promega) and the purified PCR products were digested with 80 units/each of *XhoI* restriction enzyme (New England BioLabs) overnight in a 100 µl reaction at 37ºC. Enzyme inactivation was carried out for 20 min at 65ºC. Directly subsequent to this step, pGL3 vectors were dephosphorylated with Calf Intestinal Alkaline-Phosphatase (CIP, New England BioLabs) to avoid their recircularization during ligation. PCR products and vectors were run in a 1x agarose gel and purified with QIAquick Gel Extraction Kit (Qiagen) to eliminate salts, undigested vector and small restriction fragments that could affect the ligation process. Ligation was performed overnight with between 50-100 ng of DNA (depending on the construction) with 2:1 to 3:1 insert/vector ratio in a 10 µl reaction with 200 units of T4 DNA Ligase (New England Biolabs) at 16ºC. Next, 5 µl of the ligation reaction were transformed in 50 µl of JM109 Competent Cells (Promega) by heat shock following the manufacturer's recommendations, and scattered on LB plates with 100 mg/ml of ampicillin. Since pGL3 Luciferase Reporter Vectors do not allow blue/white screening, all colonies resulting from transformation were checked for successful ligation by PCR. Positive colonies were grown overnight in LB media with 100 mg/ml of ampicillin at 37ºC and glycerinated in triplicates (TABLE 8).

TABLE 8. Reporter plasmid constructs used in this study.

| Name | Species | Insert size | Insert region | Plasmid Vector | Use |
|---|---|---|---|---|---|
| pMCL-006 | *H. sapiens* | 6.0 kb | Upstream *refTHBS4* | pGL3-Basic | Transcriptional activity of reference *THBS4* promoter |
| pMCL-008 | *P. troglodytes* | 5.7 kb | Upstream *refTHBS4* | pGL3-Basic | |
| pMCL-009 | *H. sapiens* | 1.2 kb | Upstream *refTHBS4* | pGL3-Basic | |
| pMCL-010 | *P. troglodytes* | 1.2 kb | Upstream *refTHBS4* | pGL3-Basic | |
| pMCL-011 | *H. sapiens* | 6.0 kb | Upstream *refTHBS4* | pGL3-Enhancer | |
| pMCL-012 | *P. troglodytes* | 5.7 kb | Upstream *refTHBS4* | pGL3-Enhancer | |
| pMCL-022 | *H. sapiens* | 3.0 kb | Upstream *altTHBS4* | pGL3-Basic | Transcriptional activity of alternative *THBS4* promoter |
| pMCL-024 | *H. sapiens* | 3.0 kb | Upstream *altTHBS4* | pGL3-Basic | |
| pMCL-025 | *P. troglodytes* | 3.0 kb | Upstream *altTHBS4* | pGL3-Basic | |
| pMCL-032 | *H. sapiens* | 3.0 kb | Upstream *altTHBS4* | pGL3-Enhancer | |
| pMCL-033 | *H. sapiens* | 3.0 kb | Upstream *altTHBS4* | pGL3-Enhancer | |
| pMCL-034 | *P. troglodytes* | 3.0 kb | Upstream *altTHBS4* | pGL3-Enhancer | |
| pMCL-035 | *H. sapiens* | 3.0 kb | Enhancer_Alu | pMCL-022 | Transcriptional activity of alternative *THBS4* promoter in presence of specific enhancer candidate. |
| pMCL-036 | *P. troglodytes* | 3.0 kb | Enhancer_Alu | pMCL-025 | |
| pMCL-037 | *H. sapiens* | 3.0 kb | Enhancer_4 | pMCL-022 | |
| pMCL-045 | *P. troglodytes* | 3.0 kb | Enhancer_4 | pMCL-025 | |
| pMCL-039 | *H. sapiens* | 3.0 kb | Enhancer_15 | pMCL-022 | |
| pMCL-040 | *P. troglodytes* | 3.0 kb | Enhancer_15 | pMCL-025 | |
| pMCL-041 | *H. sapiens* | 3.0 kb | Enhancer_3 | pMCL-022 | |
| pMCL-042 | *P. troglodytes* | 3.0 kb | Enhancer_3 | pMCL-025 | |
| pMCL-043 | *H. sapiens* | 3.0 kb | Enhancer_7 | pMCL-022 | |
| pMCL-050 | *P. troglodytes* | 3.0 kb | Enhancer_7 | pMCL-025 | |

In addition, plasmid DNA constructions pMCL-022 (human) and pMCL-025 (chimpanzee), formed by the pGL3-Basic vector and the *THBS4* alternative promoter of each species, were used as plasmid vectors for analyzing the transcriptional activity of this promoter in the presence of candidates enhancers that have been computationally selected (see section 2.11. Bioinformatic prediction of enhancers in MATERIALS AND METHODS). Primers were designed to amplify approximately 3 kb around the candidate enhancer in one human (NA12872) and one chimpanzee (BAC RP4341P12) sample, including a 12 nucleotides 5' tail with the restriction enzyme site for *SalI* and extra bases to permit the enzyme-DNA binding (TABLE 9). PCRs were carried out in a 50 µl reaction with Phusion®

High-Fidelity DNA Polymerase (New England BioLabs). PCR products were also run in a 1% agarose gel and purified with the QIAquick Gel Extraction Kit (Qiagen). Purified PCR products and 5 µg of pMCL-022 and pMCL-025 were digested with 40 units of the restriction enzyme *SalI* Hi-fi (New England BioLabs) overnight in a 100 µl reaction at 37 ºC. Enzyme inactivation, vector dephosphorylation, digested vector and insert purification, ligation, transformation, and colonies screening were performed as explained above for the plasmid constructions of the alternative *THBS4* promoter.

TABLE 9. Sequence and combination of primers used to amplify putative enhancer regions
Colored base pairs correspond with an extra 5' tail added to the primer with the restriction enzyme site for *SalI* (blue) and extra bases to permit the enzyme-DNA binding (orange). Fw: forward; Rv: reverse; Hs: human; Pt: chimpanzee

| Primer name | Sequence (5'-> 3') | Amplicon (bp) | Gene/Region |
|---|---|---|---|
| EnhR1.Fw | ACACACGTCGACCTAACACAGCCCCAGAGAGC | 3208 (Hs & Pt) | Enhancer Region 1 |
| EnhR1.Rv | ACACACGTCGACATAAAACGAGCCCATGTTCG | | |
| EnhR3.Fw | ACACACGTCGACCAAGCTCCTACCTGGTCTGC | 2884 (Hs) 3486 (Pt) | Enhancer Region 3 |
| EnhR3.Rv | ACACACGTCGACGGGTTTCTGTGGGGTGTAAA | | |
| EnhR4.Fw | ACACACGTCGACCCCGATTTTCATCATCTGCT | 2695 (Hs) 2698 (Pt) | Enhancer Region 4 |
| EnhR4.Rv | ACACACGTCGACTGGGTTCAAGCAATTCTCCT | | |
| EnhR7.Fw | ACACACGTCGACAATGGGCTCCACTCACAAAG | 3109 (Hs) 3106 (Pt) | Enhancer Region 7 |
| EnhR7.Rv | ACACACGTCGACCCTGTTGTTCAAGGGCAAAA | | |
| EnhR15.Fw | ACACACGTCGACACCCCTCAAAATGACAGTGG | 2942 (Hs) 2951 (Pt) | Enhancer Region 15 |
| EnhR15.Rv | ACACACGTCGACGGCAGGAGAACAGGGTGATA | | |
| EnhR17.Fw | ACACACGTCGACACACAAGTGATGGTGGCTGA | 3121 (Hs) 3131 (Pt) | Enhancer Region 17 |
| EnhR17.Rv | ACACACGTCGACAAGCAACTCCAAAACCCTCA | | |
| EnhALU.Fw | ACACACGTCGACAGTGTTCCCCAGTGTTCCTG | 2905 (Hs) 2589 (Pt) | Enhancer Region ALUY |
| EnhALU.Rv | ACACACGTCGACGCTTGTGTTATTGGGCTGGT | | |

## 2.5.2. Cloning into pGEM-T Vectors.

pGEM®-T Easy vector (Promega) is a linearized plasmid with a single 3'-terminal thymidine at both ends. The T-overhangs at the insertion site increase the efficiency of ligation of PCR products by preventing recircularization of the vector and providing a compatible overhang for PCR products generated by certain thermostable polymerases that

tend to add a single 3'-adenine to each end. This Vector was used for cloning methylation analysis PCR products (see section 2.9. Bisulfite-based methylation analysis in MATERIALS AND METHODS). pGEM®-T Easy Vectors allow blue/white screening of the colonies thanks to the presence of the *LacZ* gene, which induces the formation of β-galactosidase enzyme and the precipitation of X-gal (added in plates), producing the characteristic blue colonies. However, successful ligation disrupts the *LacZ* and thus no functional β-galactosidase can be formed, resulting in white colonies.

## 2.6. Sanger sequencing.

In several moments of this thesis project it has been necessary to sequence different DNA fragments and plasmids to verify that the different experiments were going in the right direction or as step of the experiment itself. Roughly, sequencing was used to verify the alternative mRNA isoform of *THBS4,* or to corroborate, at least partially that the different PCR products cloned for transcriptional activity quantification did not include any mutation during the PCR amplification that could interfere with the results.

In general, prior to Sanger sequencing, single band PCR products (without primer dimers) were purified with the QIAquick PCR Purification Kit (Qiagen) or ExoSAP-it (Affimetrix) to remove the excess of primers and dNTPs from the PCR. PCR amplifications from complicated regions, where it was not possible to get a single band product, were purified from the agarose gel with the QIAquick Gel Extraction Kit (Qiagen). Standard Sanger sequencing was carried out at three different services depending on the agreements of the host laboratories where this thesis project has been performed: The Servei de Genòmica of the Universitat Pompeu Fabra (Spain), the Sequencing Service of Macrogen company (South Korea) and the Genomic Service of Beckman Coulter Inc. (United States).

## 2.7. Real-Time RT-PCR.

Real-time RT-PCR quantification experiments were carried out in a LightCycler 480 Real-Time PCR System (Roche) with LightCycler 480 SYBR Green I Master (Roche), and in an ABI Prism 7900HD Sequence Detection System (Applied Biosystems) using the iTaq SYBR Green Supermix with Rox (BioRad). Primers for quantitative real-time RT-PCR were designed in specific regions conserved across humans, chimpanzees and macaques, to ensure that amplification efficiency is the same in these species without affecting the product quantification measures (TABLE 10). Quantitative RT-PCR reactions were performed in a total volume of 10 µl in 384-well plates (for Roche System) or 20 µl in 96-well optical plates (for Applied Biosystems System) with 5 or 10 µl of the PCR mix, respectively, and 2 µl of a 1/20 dilution of the cDNA (or 4 µl for 96-well plates). PCR cycling conditions were 95ºC for 8 min and 45 cycles of a denaturation step at 95ºC for 10 sec followed by annealing and extension at 60ºC for 30 seg and 72ºC for 15 seg for the Roche System. For the Applied Biosytesms System, conditions were 95ºC for 10 min and 40 cycles of a denaturation step at 95ºC for 15 sec followed by annealing and extension at 60ºC for 1 min. A dissociation curve step was added at the end of each run in both systems to ensure that only one specific product was amplified in each reaction.

For each sample, the mRNA of interest and ß-Actin were both amplified in duplicate or triplicate (depending of the number of samples analyzed and the number of wells/plate) in at least 2 independent replicates of the real-time RT-PCR experiments. Moreover, serial dilutions from a known amount of molecules of each *THBS4* mRNA isoform and the housekeeping gene ß-Actin were used as standard curve to control the efficiency of each PCR and to quantify the number of molecules in each sample. Standard curve templates were amplified by PCR, purified and quantified by NanoDrop 2000 spectrophotometer (Thermo Scientific) and/or a Qubit Fluorometer (Life Sciences). Using a housekeeping gene such as *ß-Actin*, which is expressed constitutively in most of the tissues, served as an internal control for differences in cDNA concentration among samples and was used to normalize the results. Results were analyzed by using the LightCycler® 480 Software version 1.5

(Roche) or the Sequence Detection System Software versions 1.7 or 2.3 (Applied Biosystems) depending of the machine used. Average expression levels of the replicates within each experiment were compared by a Student's t-test.

TABLE 10. Sequence and combination of primers used for real-time RT-PCR quantification.

| Primer name | Sequence (5'-> 3') | Amplicon (bp) | Gene/Region |
|---|---|---|---|
| THBS4-3 | GTTGCAGAACCTGGCATTCAG | 54 | *3'-end THBS4* |
| THBS4-4 | CCCTGGACCTGTCTTAGACTTCA | | |
| THBS4-28 | CCACCCCCCAGGTCTTTG | 122 | Reference *THBS4* |
| THBS4-30 | CAGCTTGAAGGTGGAAATCACA | | |
| THBS4-31 | CCTTCGCCTTCCACCATGA | 131 | Alternative *THBS4* |
| THBS4-30 | CAGCTTGAAGGTGGAAATCACA | | |
| ACTB-3 | CTGGAACGGTGAAGGTGACA | 195 | *3'end ACTB* |
| ACTB-4 | GGGAGAGGACTGGGCCATT | | |
| ACTB-5 | ACGAGGCCCAGAGCAAGA | 62 | *5'end ACTB* |
| ACTB-6 | GACGATGCCGTGCTCGAT | | |
| GAPDH-1 | CTGACTTCAACAGCGACACC | 120 | *GADPH* |
| GADPH-2 | CCCTGTTGCTGTAGCCAAAT | | |
| 18srRNA-1 | GCAATTATTCCCCATGAACGA | 52 | *18s* rRNA |
| 18srRNA-2 | AAGCTTATGACCCGCACTTACTG | | |

## 2.8. Luciferase assay.

Plasmid DNA including the reference and alternative *THBS4* promoter regions were extracted under the same endotoxin-free conditions and quantified carefully with NanoDrop 2000 (Thermo Scientific) and Quant-iT™ dsDNA High-Sensitivity (HS) in a Qubit Fluorometer. Concentration of all of plasmids was normalized to 300 ng/µl to ensure that the same amount of plasmid DNA was transfected into the cells. In addition, 40 ng/µl of the *renilla* pRL-TK plasmid vector was prepared to be used as a control. *Firefly* luciferase and *renilla* luciferase use different substrates in their luminescent reaction. In addition, the luminescence of *firefly* luciferase can be quenched without affecting the activity of *renilla* luminescence. Thus, the amount of each luciferase product can be individually defined and used to control for differences in transfection efficiency or growth conditions between cells.

Optimization of conditions like the number of wells per plate, the number of cells per well, the amount of transfected DNA, the medium conditions before transfection, the time after transfection, and even the reporter assay method used was required for a proper performing of the experiments. As final conditions, cell lines were grown in 12-well plates with DMEM Media - L-Glutamine (Invitrogen) supplemented with 10% FBS (Invitrogen) and 1% PSF (Gibco). 125.000 cells/well were seeded for T98G and HEK293T cell lines and 150.000 cells/well were seeded for SH-SY5Y, SK-N-AS and Caco-2 cell lines. A total of 400 ng of the each plasmid construction and 4 ng of the *renilla* luciferase reporter plasmid were transfected 24 hours after seeding with 1.2 µl of FuGENE®6 Tranfection Reagent (Roche and Promega) in a ratio 3:1 to DNA (3 µl Fugene 6:1 µg DNA). 48 hours after transfection, the cells were washed with PBS and lysed with 1x passive lysis buffer of the Luciferase reporter assay system (Promega) for 15 minutes at room temperature on a platform shaker. Meanwhile, Lumi vials tubes (Berthold Technologies) were prepared with 100 µl of Luciferase Assay Reagent II, containing the luciferin substrate of the *firefly* luciferase. A test assay showed that HEK293T cell lines saturate the maximum level of luminescence of the luminometer. Therefore, after the lysis reaction, a 1:10 dilution of the lysate was performed in new 1x passive lysis buffer. To induce the first luminescent reaction, 20 µl of sample lysate was added to one tube. Then the sample was carefully mixed without vortexing and *firefly* luminescence was quantified with a Lumat LB 9507 Single Tube Luminometer (Berthold Technologies). Directly thereafter, 100 µl of the Stop & Glo reagent was added to the tube, mixed vigorously and quantified for *renilla* luminescence. The Stop & Glo Reagent of the kit quenches *firefly* luminescence and simultaneously provides the substrate for *renilla* luciferase, enabling quantification of *renilla* luminescence without disturbance from *firefly* luminescence.

For the luciferase assay experiments that measured the transcriptional activity of each *THBS4* promoter in human and chimpanzee samples, the *firelly/renilla* ratio of each sample were normalized by the *firelly/renilla* ratio of the pGL3-Basic vector. However, when this technique was used for measuring the transcriptional activity in presence of a candidate enhancer region (MATERIALS AND METHODS section 2.5.1 and 2.11), samples were

normalized to the *firelly/renilla* ratio of the human pMCL-022 plasmid, which includes 3 kb upstream of the alternative promoter cloned in the *XhoI* restriction enzyme site of a pGL3 basic vector. Average ratios of the replicates within each experiment were compared by a Student's t-test.

## 2.9. DNA methylation analysis.

Genomic DNA treatment with sodium bisulfite converts cytosine residues into uracils, but leaves 5-methylcytosines unaffected, which then can be used to characterize DNA methylation in a region of interest. One of the main limitations of this treatment is the integrity of the DNA. Conditions necessary for complete conversion, such as long incubation times, elevated temperature, and high sodium bisulfite concentration, can lead to the degradation of the incubated DNA. EZ DNA Methylation gold kit (Zymo Research) was used to perform the DNA conversion. We opted for this commercial kit because it combines DNA denaturation and bisulfite conversion processes into one step, achieving a much faster bisulfite conversion. Two different conversion conditions were assayed: condition 1, 10 min at 98ºC, 2.5 hours at 63ºC and up to 20 hours at 4ºC; condition 2, 10 min at 98ºC, 10 min at 53ºC, 8 cycles of 6 min at 53ºC and 30 min at 37ºC and a final step up to 20 hours at 4ºC. The starting amount of DNA was also optimized (400 – 1000 ng), since large amounts of DNA often lead to incomplete conversions and small amounts can be problematic if there is extensive degradation. Finally, since degradation occurs due to depurinations causing random strand breaks (EHRICH *et al.* 2007), the longer the desired PCR amplicon, the more limited the number of intact template molecules would likely be. Initially, two forward and two reverse primers were designed with an expected product size variable between 555 and 728 bp depending of the different combinations used. However, shorter product sizes had to be considered due to the impossibility to optimize the PCR amplification.

After the optimization tests, 400 ng of genomic DNA of frontal cortex samples of 5 humans and 5 chimpanzees were treated with sodium bisulfite following the manufacturer's instructions. Final conversion conditions were 10 min at 98ºC, 2.5 hours at 63ºC and up to 20 hours at 4ºC. Primers for amplification of sodium bisulfite-treated DNA were designed replacing all cytosines present in the forward primer for thymines and all the guanines in the reverse primer for adenines (TABLE 11 and FIGURE 7). This reduction of the primers GC-content was compensated by increasing the primer length to 24–32 bp. CpG dinucleotides are generally not recommended in the primer sequences to avoid a potential bias towards unmethylated CpGs. However, primers should contain non-CpG cytosines to ensure the specific amplification of the converted DNA and to supply stability to the primers, due to the stronger bonding of G and C bases and the raise of the melting temperature. DNA after sodium bisulfite treatment may result in incomplete conversion of cytosines. Therefore, PCR with non-modified primers was used as negative control to ensure that there was not amplification from the original DNA and only converted DNA was amplified (TABLE 11).

TABLE 11. Sequence and combination of primers used for transformed (T) and non-transformed DNA amplification.

| Primer name | Sequence (5'-> 3') | Amplicon (bp) | Gene/Region |
|---|---|---|---|
| THBS4-48 | GCAATCAAACATTTTAGACCAAGTG | 425 | CpG island alternative *THBS4* |
| THBS4-51 | AGGCTGACTTGCCCCAGTTT | | |
| THBS4-48-T | GTTTGTAATTAAATATTTTAGATTAAGTG | 430 | CpG island alternative *THBS4* |
| THBS4-51-T | CAAACTAACTTACCCCAATTT | | |
| THBS4-50 | GAGCCCAGGGCAATGCAG | 383 | CpG island alternative *THBS4* |
| THBS4-53 | GCCAAGCTGCTCCTAAGAGTT | | |
| THBS4-50-T | GGAGGGAGTTTAGGGTAATGTAG | 392 | CpG island alternative *THBS4* |
| THBS4-53-T | TTAAACCAAACTACTCCTAAAAATT | | |
| pU/M13fw | GTTTTCCCAGTCACGAC | Variable | CpG island alternative *THBS4* |
| pUC/M13rv | CAGGAAACAGCTATGAC | | |

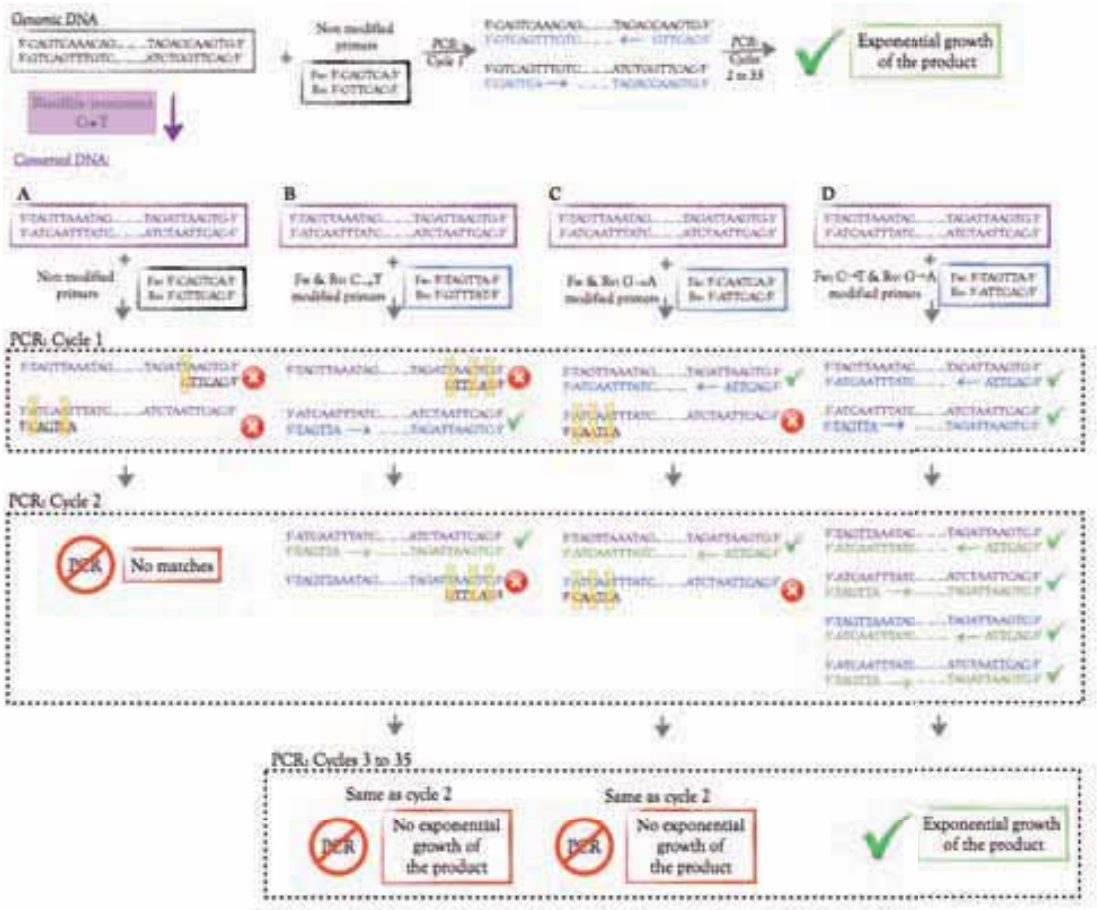



FIGURE 7. **Primer design for sodium-bisulfite treatment experiments.** For a proper exponential growth of the PCR product, it should be considered that primers designed for genomic DNA (black) would not match the converted genomic DNA (purple), because all non-methylated cytosines (C) had been changed into thymines (T) (Case A). Moreover, when the sequences of both forward (Fw) and reverse (Rv) primers, were modified following the same rule, changes of C to T (Case B) or G to A (Case C), one strand would be lost in each cycle, thus avoiding the exponential growth of the product. However, if cytosines present in the forward primer were changed for thymines and in the reverse primer guanines for adenines, converted DNA would pair with both primers and a successful PCR could take place (Case D).

A 550 bp CpG island localized immediately upstream of the *THBS4* alternative promoter was amplified in two different overlapping fragments of 425 bp and 383 bp with AmpliTaq Gold® 360 DNA Polymerase (Invitrogen), which adds a single 3'-adenine overhang to each end of the PCR product due to its lack of 3' to 5' proofreading activity. PCR products for both fragments and for each sample were cloned in a pGEM®-T Easy Vector System (Promega) and 5 µl of the ligation were transformed in JM109 Competent

Cells (>$10^8$ cfu/µg, Promega), as described in the section 2.5 "cloning and transformation" of MATERIALS AND METHODS. The possibility of carrying out blue/white screening of the colonies thanks to the presence of the *LacZ* gene, allowed a much easier discrimination of the positive colonies. However, some of the colonies presented a bluish coloration, not being able to visually determine or not if the insertion had taken place. Positive colonies were checked by PCR with the universal pUC/M13 vector primer pair to confirm that they had the fragments of interest. PCR products of 8-10 different colonies for each fragment and each sample were initially sent to be sequenced from both strands through pUC/M13 universal primers in two 96 well plates at Beckman Coulter Inc. (United States). After checking the resulting sequences, there were some samples in which sequencing did not work properly. We then took advantage of the plasmid DNA clones that were grown in a LB+ampicillin plate for re-sequencing at Macrogen company (South Korea) samples that did not work properly the first time plus some extra samples that had not been considered initially. Sequences were aligned with CLC DNA workbench 6 Software (CLCBio). Methylation levels of the different clones in a specific CpG position were compared by a Student's t-test.

## 2.10. ChIP-Seq.

SH-SY5Y, SK-N-AS and T98G human cell lines were used for ChIP-Seq experiments (chromatin immunoprecipitation coupled to high-throughput sequencing) to try to identify active enhancers. Before harvesting, cell lines were treated with a crosslinking solution with 1% formaldehyde for 20 min at room temperature. Incubation time and method were consistent between cell lines, since too little crosslinking would not sufficiently preserve the chromatin structure and too much crosslinking would hamper the ChIP procedure. Crosslinked chromatin was sheared with a Bioruptor® Standard (Diagenode) for 17 min in 30 sec ON/OFF cycles. Next, 30 µg of chromatin was immunoprecipitated using Dynabeads® Protein A (Invitrogen) coupled with the p300 antibody (C-20) (sc-585, Santa Cruz Biotechnology). Rabbit Control IgG - ChIP Grade antibody (ab46540, Abcam) was

used as a negative control to measure IgG nonspecific binding. A chromatin input control (non-immunoprecipitated) was also used as a reference control. Finally, 6 different regions that were previously identified on the human genome assembly as binding p300 in different cell lines (OVCHARENKO *et al.* 2004, CONSORTIUM *et al.* 2007, HEINTZMAN *et al.* 2007, HEINTZMAN *et al.* 2009, VISEL *et al.* 2009) were used as positive control candidates and quantified by real-time PCR (TABLE 12) before sequencing the samples.

TABLE 12. Sequence and combination of primers used to test immunoprecipitated DNA by real-time RT-PCR quantification.

| Primer name | Sequence (5'-> 3') | Amplicon (bp) | Gene/Region |
|---|---|---|---|
| P300-R1Fw | TGGGCCTTGTAGTTTTCTTTCC | 68 | chr1: 61630551-61631050 |
| P300-R1Rv | TGTTTCACTCTGCACACTTTTCTTT | | |
| P300-R2Fw | CTCGGGACTGAATGTTACCCAT | 82 | chr5:124359201-124359700 |
| P300-R2Rv | AAACCATCCAACAAGCATTAAAAGA | | |
| P300-R3Fw | CAGAAACCTGGGTGAAAGGA | 140 | chr14:52994751-52995250 |
| P300-R3Rv | TCCTCCAACCCTCACATCAG | | |
| P300-R4Fw | AGGCAAGAGACAAGAGAATGTAACC | 69 | chr11:116448451-116448950 |
| P300-R4Rv | CACTTGGAAACTGTGACTCAGCAT | | |
| P300-R5Fw | AAATTCTGCCTCTCTTTCAAAACC | 66 | chr2:234390601-234391100 |
| P300-R5Rv | CCTGGAAATGCCCCAGATAA | | |
| P300-R6Fw | GCCACCAGCTTCACATAAAC | 90 | chr13:29814451-29814950 |
| P300-R6Rv | ATTGACCTGACCTCCAAGAA | | |
| ncP300-R1Fw | TCCTCCATCTGACCCTGAAC | 126 | chr19:56031354-56032021 |
| ncP300-R1Rv | GGAGGGTTTGGGGAAATAGA | | |
| ncP300-R2Fw | CTACCTCTGCACCCACCAAT | 107 | chr1:114516353-114516852 |
| ncP300-R2Rv | AGAGAAGCTTTGCCCATTCA | | |

DNA from two independent p300 ChIP experiments for each cell line were subjected to deep sequencing using the Solexa Genome Analyzer (Illumina) at the CRG Genomics Core Facility. Additionally, due to the final amount of IgG ChIP was below the minimum required for sequencing they were not used as control. Instead, the inputs of the two experiments were mixed in a unique sample (each cell line separately) and subjected also to deep sequencing. The use of input as control for ChIP experiments provided exact information about the amount of chromatin available before carrying out the immunoprecipitation. The binding sites for p300 were detected with *Pyicos* software, a tool

for the analysis of high-throughput sequencing mapped reads (ALTHAMMER *et al.* 2011). Initially, all the estimated regions with coverage equal or higher to 10 were considered for the binding sites detection. After that, regions that were also detected in their respective negative control were discarded. Single-ended sequences were mapped to the hg19 human genome assembly. The database *RefGenes* available in the same UCSC webpage was used for studying the overlapping of the different regions with genes.

## 2.11. Bioinformatic prediction of enhancers.

The information available about the human genome (hg18) in UCSC Genome Browser (http://genome.ucsc.edu) was used to search for enhancer sites that could regulate the alternative *THBS4* mRNA expression. Particularly 50 kb up- and downstream of the alternative *THBS4* (chr5: 79.322.874 - 79.414.863 ± 50 kb) were analyzed in 5 kb windows. The first criterion for selecting an enhancer site was based on three ENCODE regulation tracks that can be found within the ENCODE integrated regulation track. If at least two of them were fulfilled, the site was considered a possible enhancer. The three regulation tracks were:

- ENCODE enhancer- and promoter-associated histone mark (H3K4Me1) on 8 cell lines: Epigenetic modifications to the histone proteins present in chromatin influence gene expression by changing how accessible the chromatin is to transcription. A specific modification of a particular histone protein is called a histone mark. This track shows the levels of enrichment of the H3K4Me3 histone mark across the genome as determined by a ChIP-seq assay. The track displays cell lines with different colors in the same vertical space. Scale of the vertical view range can be configured in the track settings. The region was accepted if the peak had a value of 50 or more over a maximum value of 100 in at least one cell line.
- ENCODE digital DNaseI hypersensitivity clusters: Regulatory regions tend to be DNase I sensitive due to changes in nucleosome organization in active chromatin. This track shows DNase hypersensitive areas assayed in a large collection of cell types.

A gray box indicates the extent of the hypersensitive region, and the darkness is proportional to the maximum signal strength observed in any cell line. The number to the left of the box shows how many cell lines are hypersensitive in the region. If there was at least one cluster in the region, it was accepted.

- ENCODE transcription factor ChIP-seq: This track shows regions where transcription factors bind to DNA as assayed by ChIP-seq in different cell lines used in the ENCODE project. A gray box encompasses the peaks of transcription factor occupancy. The darkness of the box is proportional to the maximum signal strength observed in any cell line. If there was at least one cluster in the region, it was accepted.

To do a precise selection and reduce the number of candidates to be analyzed experimentally, a second criterion based on the UCSC track ChromHMM was applied. This track displays a chromatin state segmentation based on a ChIP-seq pipeline to generate a genome-wide chromatin data set for each of the nine cell lines used (ERNST *et al.* 2011). To summarize their results, ERNST and collaborators selected 15 chromatin states that showed distinct biological enrichments and were consistently recovered. They distinguished six broad classes: promoter, enhancer, insulator, transcribed, repressed and inactive states. Among these classes, active, weak and poised promoters differ in expression level, strong and weak candidate enhancers differ in expression of proximal genes, and strongly and weakly transcribed regions also differ in their positional enrichment along transcripts. Similarly, repressed regions differ from heterochromatic and repetitive states, which are also enriched for H3K9me3. Only regions labeled as strong or weak candidate enhancers were selected for further analysis. Finally, evolutionary conservation in all the regions was checked using the ECR browser.

## 2.12. Allele-specific expression quantification.

Two different methods were used for the analysis of the expression of the two main human *THBS4* haplotypes, taking the SNP at exon 3 (rs438042) as a marker: allele-specific real-time RT-PCR (AS-PCR) and pyrosequencing. AS-PCR quantified the expression levels of each allele for a specific SNP by real time RT-PCR in presence of a standard curve with a known number of molecules. The pyrosequencing technology, also denominated as "sequencing by synthesis", is based on the synthesis of a single strand of DNA by copying base by base a complementary strand. This reaction is carried out in presence of a chemiluminiscent enzyme, which generates an amount of light proportional to the number of nucleotides added in every step. It allows the genotyping of SNPs or sequencing of unknown regions. In addition, if one allele is more expressed than the other in a heterozygous sample, the incorporation of the complementary nucleotide will produce a higher luminescent peak.

### 2.12.1. AS-PCR

To achieve selective amplification of each allele of the A/T polymorphism understudy, specific primer pairs were designed. The same reverse primer was used to amplify both alleles, in combination with two practically identical forward primers. Only a modification in the last base at the 3'-end results in one forward primer matching one of the alleles, but not the other - and *vice versa* (TABLE 13). Even though both primer pairs are practically identical, they might not have the same efficiency. Therefore, to control primer efficiency, a standard curve of five different mixes was created from two cDNA stocks homozygous for both SNP variants. Each mix had a known percentage of each allele, from 100% of one allele and 0 % of the other to the other way around, having in between the intermediate proportions of 75%-25% and 50%-50%. Since an initial common mix with primers, SYBR green and water is uniformly distributed among the different wells in the preparatory steps

of real time RT-PCR, primers variations equally affect control and study samples, thus allowing detection of potential problems with the efficiency of the primers. In the same way, in order to verify that the amount of cDNA was not interfering with the results, total *THBS4* levels were also amplified and used as internal control. Real-time PCRs were performed in 96-wells plates with iTaq SYBR Green Supermix with Rox (BioRad). The procedures for the real-time RT-PCR and primer sequences for the *THBS4* 3'-end (TABLE 10) are described at MATERIALS AND METHODS. 2.7 Real-Time PCR.

TABLE 13. Sequence and combination of primers for allele-specific real time RT-PCR.

| Primer name | Sequence (5'-> 3') | Amplicon (bp) | Gene/Region |
|---|---|---|---|
| THBS4-Ex3 FwA | GGACTTTCCAGAGGAAGCCA | 169 | rs438042 quantification |
| THBS4-Ex3-Rv | TGTGTCATTTGACCCAAGAACT | | |
| THBS4-Ex3-FwT | GGACTTTCCAGAGGAAGCCT | 169 | rs438042 quantification |
| THBS4-Ex3-Rv | TGTGTCATTTGACCCAAGAACT | | |

## 2.12.2. Pyrosequencing

The first step for pyrosequencing experiments was the amplification of the region of the A/T polymorphism by PCR with a biotinylated forward primer and a non-biotinylated reverse primer (TABLE 14).

TABLE 14. Sequence and combination of primers for pyrosequencing experiments.

| Primer name | Sequence (5'-> 3') | Amplicon (bp) | Gene/Region |
|---|---|---|---|
| THBS4-piro-Fw | CTCCCAGAAACCTGAGACCA | 79 | rs438042 amplification |
| THBS4-piro-Rv | CACCAGCTTCAGCTCTTCCA | | |
| THBS4-piro-S | TCTTCCAAGAAGTCCTG | – | rs438042 genotyping |

Pyrosequencing was performed using the PyroMark Gold Q96 Reagents (Qiagen). A total of 40 µl of PCR product was prepared according to the manufacturer's protocol. Samples of biotinylated PCR products were immobilized by Streptavidin Sepharose High Performance beads (GE Healthcare). The beads were then aspirated with PyroMark Q96

Vacuum Prep Workstation (Qiagen) and incubated with 70% ethanol, denaturing buffer, and washing buffer. Subsequently, the beads were released into pyrosequencing reaction plates, containing annealing buffer and the sequencing primer, located a 1 bp away from the SNP to be quantified (TABLE 13). Primer annealing was performed by heating the samples at 80 ºC for 2 min and by cooling at room temperature for 5 min prior to pyrosequencing. Pyrosequencing reaction was carried out in PSQ™ 96MA System (Qiagen). Pyrosequencing data were quantified and background corrected using the PSQ™ 96MA version software (Qiagen).

## 2.13. Common bioinformatic analysis.

To carry out many of the aforementioned techniques, it has been necessary the previous bioinformatic search and analysis of several genomic sequences. The first step for approaching to the study of a gene is obtaining its sequence from the database *nucleotide* at the website of the National Center for Biotechnology Information (NCBI, http://www.ncbi.nlm.nih.gov). After obtaining the mRNA sequence from the gene, it can be copy in the Blat search at the University of California Santa Cruz (UCSC) genome browser website (http://genome.ucsc.edu/), which will find the matches between the provided sequence and the chosen genome (for example human). This allowed us to, determine the beginning and end of the gene's exons, and when working within primates sequence, it gave reliable information of the homolog splicing sites in the closely-related species. When the mRNA o DNA sequence of diverse species has been obtained, MUSCLE (MUltiple Sequence Comparison by Log- Expectation) or ClustalW2 was used to perform a multi-alignment of different sequences (both available at http://www.ebi.ac.uk/). Visualization and edition of the alignments was performed with software programs like BioEdit (http://www.mbio.ncsu.edu/bioedit/bioedit.html) or CLC main workbench (http://www.clcbio.com). When it was required the comparison and alignment of two sequences that might not align perfectly, BLAST (Basic Local Alignment Search Tool, available at http://blast.ncbi.nlm.nih.gov/) was used to find regions of local similarity

between sequences without needing an exact or nearly exact match to find a hit (as it is required for Blat).

# 3

# RESULTS

*After climbing a great hill, one only finds that there are many more hills to climb.*

- NELSON MANDELA -

# 3

# RESULTS

By the time this thesis project started (May 2008), several microarray studies to identify genes with expression differences between the cerebral cortices of humans and chimpanzees had already been published (ENARD *et al.* 2002, CÁCERES *et al.* 2003, KHAITOVICH *et al.* 2004, KHAITOVICH *et al.* 2005). In addition, using rhesus macaques as an outgroup it had been possible to identify gene expression changes ascribed specifically to the human lineage (CÁCERES *et al.* 2003). Particularly, CÁCERES and colleagues had begun the characterization of two synaptogenic thrombospondins in the human brain, thrombospondin-2 (*THBS2*) and thrombospondin-4 (*THBS4*), identifying around 2-fold and 6-fold greater mRNA expression, in human cerebral cortices compared to chimpanzees and macaques respectively (CÁCERES *et al.* 2007). The main question then was to identify the origin of these expression differences, especially those of *THBS4*.

## 3.1. Computational characterization of *THBS4* regulatory changes.

To characterize the regulatory changes that could be involved in *THBS4* upregulation in humans we looked into the different possible molecular causes of this expression change. We searched possible variations within the *THBS4* genomic context between humans and chimpanzees and localized the possible transcripts expressing *THBS4*.

### 3.1.1. The *THBS4* genomic context.

To begin with the characterization of *THBS4* and to identify the possible regulatory changes responsible for its up-regulation in the human brain, we first studied the genomic context of *THBS4* in humans and chimpanzees. Taking advantage of the sequenced genomes available at the UCSC Genome Browser website (http://genome.ucsc.edu/), we identified the human *THBS4* gene in chromosomal region 5q14.1. It is flanked on each side by genes transcribed in the opposite direction: the metaxin-3 *(MTX3)* gene, the transcription start site (TSS) of which is located about 44 kb upstream of the defined TSS for the *THBS4* gene, and the serine incorporator 5 (*SERINC5)* gene, whose 3' end is found around 28 kb downstream of the *THBS4* 3' end. The hg18 version, available since March 2006, was used as a reference of the human genome for this characterization. Once the next version of the human genome (hg19) was available in February 2009, the genomic context of *THBS4* was rechecked to corroborate that the region had not been modified.

After localizing the gene, we looked for possible structural changes between the genomes of humans and chimpanzees that could be responsible for the up-regulation of *THBS4*. First, using the information available at UCSC Genome Browser and the data of PERRY *et al.* (2008), who performed a comparative genomic hybridization to identify CNVs among the genomes of humans and chimpanzees, we made sure that there was a single copy of *THBS4* in the human and in the chimpanzee genome sequences. In addition, we searched for other possible structural changes (such as insertions, deletions or inversions of DNA) within the gene sequence or adjacent regions where possible regulatory elements could be likely expected. Considering an area spanning ±50 kb around the *THBS4* gene in human, chimpanzee, gorilla, orangutan and macaque sequences, we performed two alignments using the MultiPipMaker server (http://bio.cse.psu.edu/). This software allows to align sequences between a contiguous reference sequence, in our case humans (FIGURE 8A) or chimpanzees (FIGURE 8B), and one or more secondary sequences, in our case the four remaining species. As a result, a stacked set of percent identity plots (MultiPip)

comparing the reference sequence with subsequent sequences is obtained (SCHWARTZ *et al.* 2003).

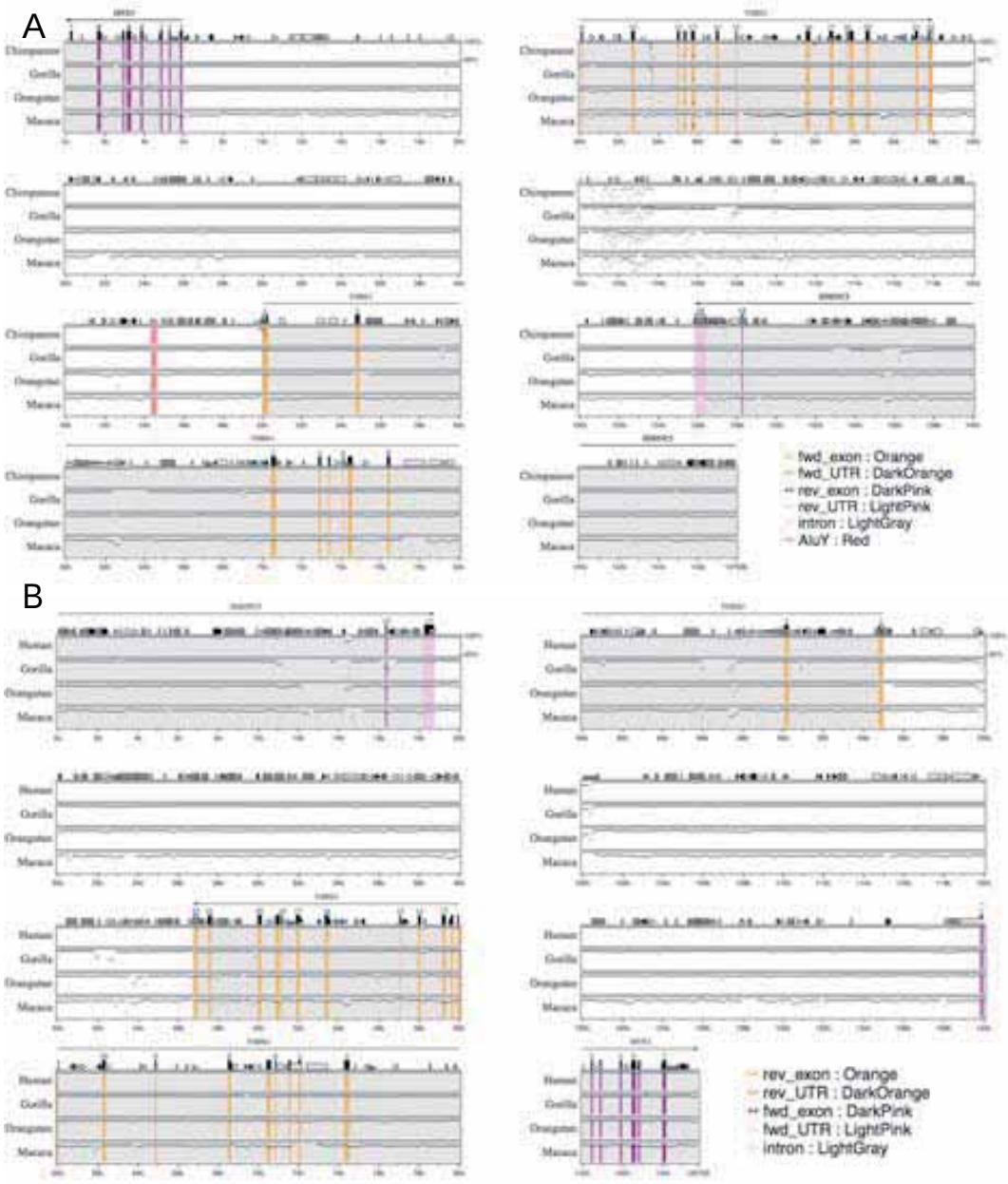



FIGURE 8. MultiPipMaker alignments for ±50 kb of the ***THBS4*** gene in human, chimpanzee, gorilla, orangutan and macaque sequences. Human (A) or chimpanzee (B) sequences were used as alignment reference. Gene name and transcriptional orientation are specified for each gene. Exons/introns are colored as indicated in the legend and are indicated as tall black boxes at the top of the figure. Repetitive elements are represented by different symbols.

Additional information as annotations indicating the name, position, transcriptional orientation of the genes and icons that represent repetitive elements can be added supporting the reference sequence. A human-specific AluY transposable element insertion upstream the *THBS4* gene stood out in the alignments of humans against the other species. We did not find further relevant structural variation in the sequences of both species. Small variation between human and chimpanzee sequences in this portion of the genome could be explained by minor insertions or deletions ("indels") that appear in both species since their division from a common ancestor around 5-7 million years ago (Mya) (VARKI and ALTHEIDE 2005).

However, this analysis pointed out that the reading directions of the gene are opposite between the human and chimpanzee. The most likely explanation is that *THBS4* is located inside a known chimpanzee-specific inversion (SZAMALEK *et al.* 2005). The breakpoints of this inversion of about 77 Mb in length are located in the human region 5p15 (chr5: 18,589,070–18,589,078) and 5q15 (chr5: 95,947,182–95,947,192), whereas the reference sequence for *THBS4* locates the gene at chr5q14.1 (chr5: 79,366,926–79,414,863), which is around 16 Mb away from the closest inversion breakpoint. This long distance to the breakpoint provides a large highly conserved region in both species, indicating that possible regulatory elements for gene expression might have the same overall organization in humans and in chimpanzees.

### 3.1.2. Identification of the *THBS4* promoter.

Assuming that there is a similar *THBS4* genomic context between both species, the next step was to confirm the localization of the transcription start site of the gene to determine the promoter region involved in the transcription of the gene. We analyzed the expressed sequence tags (ESTs), which provide information about the expressed region of the gene, available at http://genome.ucsc.edu/. We found 15 ESTs supporting the transcription start site defined for the *THBS4* RefSeq mRNA (NM_003248). To complement the search with

additional supporting information for this region rich in ESTs, we checked the data available from the Cap Analysis of Gene Expression (CAGE, http://fantom.gsc.riken.jp/), which detects distinct transcription start sequences and marks their location (CAGE tags) (SHIRAKI *et al.* 2003, SHIMOKAWA *et al.* 2007). These CAGE tags are short sequences focused in the 5' ends of capped transcripts representing the beginning of the mRNA. We found 35 CAGE tags (9 of which originated from brain samples) in a region of 162 bp, supporting the *THBS4* RefSeq transcription start site. The beginning of these CAGE tags was variable within the 162 bp, 3 tags map in the position -33 in reference to the described TSS, 15 map at position -1, 11 at position +36 and 4 at position + 108. The current version of the *THBS4* reference sequence mRNA (NM_003248.4, updated on January 12th, 2013) has 3233 bp, which contains 22 exons and encodes for a protein with 961 aa (FIGURE 9).

Continuing with the analysis of the *THBS4* region, we found that about 44 kb upstream from the TSS defined for the RefSeq mRNA, there are 12 additional ESTs. The 5'-ends of these short sequences are located no more than 50 bp away from the exon 1 of *MTX3* gene, and the two last exons of the ESTs correspond to the exons 2 and 3 of *THBS4* mRNA. This suggests that there could be a non-defined transcript that may belong to the *THBS4* gene. Moreover, to provide additional support for the found ESTs, we identified 15 CAGE tags (12 of which were originated from the brain) in a region of 78 bp around the 5'- end of these ESTs. These findings suggested that the respective sequence could belong to an alternative TSS for *THBS4* that could be related to brain expression.

To validate if there is an alternative promoter for *THBS4* as the ESTs and CAGE tags suggest, we designed different pairs of primers targeting from the exon 1 of the ESTs to diverse exons within the *THBS4* RefSeq sequence (TABLE 15). We performed RT-PCR to the exon 4 (616/668 bp), exon 8 (1170/1222 bp), exon 16 (2127/2179 bp) and exon 22 (2912/2967 bp) using RNA samples from the human neuroblastoma cell line SH-SY5Y (FIGURE 9). Initially, we expected to have some variation in the size of the fragments since the 3'-end of exon 2 presents two conformations with a variation of 52 bp between the ESTs. However, for all the amplifications, bands of the expected size considering the shorter form of the exon 2 were obtained, with a detectable lighter band in some of the samples

corresponding to the longer alternative exon 2. To be sure that the amplified fragments belonged to an alternative isoform of *THBS4*, the longer fragments (2127 and 2912 bp) were sequenced. For this process, we took advantage of the available primers that were used in the RT-PCR amplification plus additional internal primers that facilitate the covering of the entire sequence length in both the positive and the negative strands. The resulting sequences confirmed the existence of an alternative mRNA isoform for *THBS4* of 3115 bp (chr5: 79,322,875-79,414,866). This isoform has 23 exons, two of which are specific for the alternative isoform, named "alternative exon 1" (AE1) and "alternative exon 2" (AE2). The other 21 exons are shared with the reference isoform of the gene, corresponding with exons 2 to 22 (FIGURE 9).

TABLE 15. Sequence and combination of primers used to validate ***THBS4*** alternative mRNA and other ***THBS4*** transcripts.

| Primer name | Sequence (5'-> 3') | Amplicon (bp) | Gene/Region |
|---|---|---|---|
| THBS4-20 | GGGTAATTTGGGGGCTCTTGA | 661/668 | *THBS4* AE1 →E4 |
| THBS4-19 | CCTCTCACCACCAGCTTCAGC | | |
| THBS4-20 | GGGTAATTTGGGGGCTCTTGA | 1170/1222 | *THBS4* AE1 →E8 |
| THBS4-21 | AGACCTGCTTGTTTGACTTGG | | |
| THBS4-20 | GGGTAATTTGGGGGCTCTTGA | 2127/2179 | *THBS4* AE1 →E16 |
| THBS4-24 | TGTTGCTATCCTCCTGGGCTG | | |
| THBS4-20 | GGGTAATTTGGGGGCTCTTGA | 2912/2967 | *THBS4* AE1 →E22 |
| THBS4-25 | GCGGTCGAAATTCTGGGTTTG | | |
| THBS4-22 | TAGGTCAGAATTGCTCAGGGA | 786 | Antisense ESTs (*THBS4*) |
| THBS4-23 | GCTGTCATCTCTAACTGTCCTCAT | | |
| THBS4-AE3-40 | AACACCCAGCAAATAACCATACTAAT | 105 | *THBS4* AE3→E3 |
| THBS4-41 | TGAAAACCACCAAATGCACCT | | |

Comparing both isoforms, between the first methionine in the exon 1 of the *THBS4* reference gene and the amino acid in position 26, there is the signal peptide, which was predicted and localized by the search of the amino acid reference sequence in software like *SignalP 4.1 Server* (PETERSEN *et al.* 2011). This sequence (for this gene: *MLAPRGAAVLLLHLVLQRWLAAGAQA*) is supposed to help with the insertion of the protein into the membrane of the endoplasmic reticulum, where it is usually cleaved off by the signal peptidase and presumably degraded rapidly (MARTOGLIO 2003). The newly

described mRNA lacks exon 1 of the reference isoform, and consequently its initiation of translation and its signal peptide, suggesting it should produce a slightly different protein. The exact localization of the first amino acid and the structure of the protein encoded by this newly discovered mRNA is still unknown. However, according to the graphical analysis tool *ORF Finder* (http://www.ncbi.nlm.nih.gov/), which finds all open reading frames in an amino acid sequence, the alternative mRNA may start the translation in the exon 2 (common to both isoforms) and produce a protein of 870 aa. This amino acid sequence was also analyzed with the *SignalP 4.1 Server* but no signal peptide was detected.

### 3.1.3. Other *THBS4* transcripts.

The analysis of the ESTs sequences around the *THBS4* gene region detected some other different and low frequency transcripts in humans, which were examined for a possible association to the gene expression differences. We identified at least two other different transcripts: an antisense transcript and a transcript with an extra exon.

At the 3'-end of *THBS4,* there are several ESTs in the negative strand of the DNA and one of them generates an extra transcript overlapping the *THBS4* gene, but in the opposite direction. These ESTs can be summed up in the UCSC gene BC047373 (not a Refseq mRNA). To analyze this region more carefully, we aligned the BC047373 sequence with several other antisense-ESTs and the equivalent sequences in the available primate species that already had their genomes sequenced. To evaluate the antisense expression by PCR, primers were designed for regions that were common in most of aligned ESTs (TABLE 15). After several tries of PCR optimization, no antisense transcript could be successfully amplified in a sufficient level to be visually detected in a semi-quantitative electrophoresis. The characterization of this possible antisense RNA was then set aside, since it was unlikely that this low expressed transcript was responsible for the *THBS4* over-expression in the human brain.

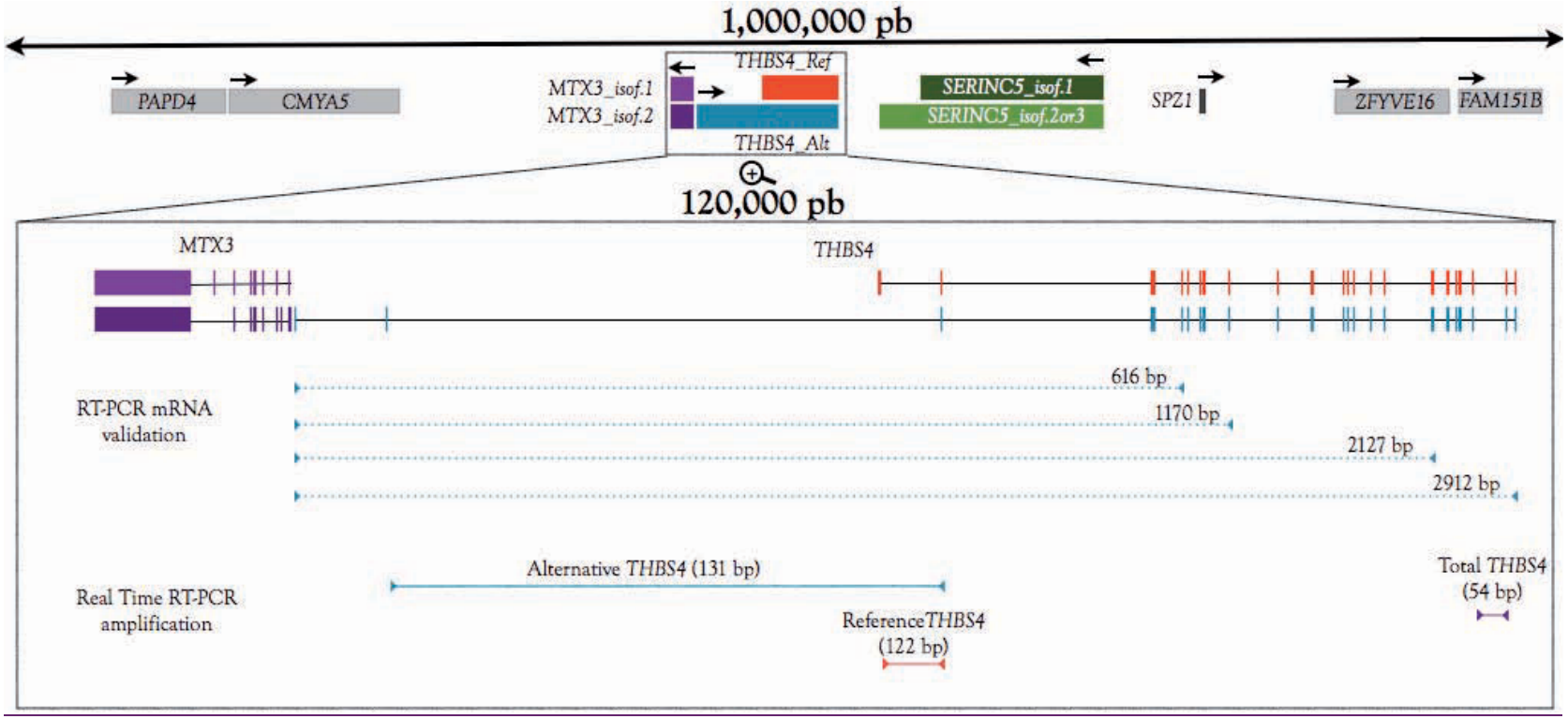



FIGURE 9. ***THBS4*** genomic context and experimental analysis of the two THBS4 isoforms. Representation of 1 Mb and 120 kb of the Chr5q14.1 region, enclosing both mRNA isoforms of the *THBS4* gene and surrounding genes at the top. Dotted lines below the gene diagrams correspond with the RT-PCR amplifications for mRNA validation, from the putative alternative exon 1 to the exon 4 (616 bp), exon 8 (1170 bp), exon 16 (2127 bp) and exon 22 (2912 bp) of the reference mRNA isoforms. Primer localization for real time RT-PCR quantification of *THBS4* total mRNA expression (54 bp) and both alternative (183 bp) and reference (122 bp) mRNA isoforms are represented at the bottom.

The second detected variant transcript within the *THBS4* region corresponds to the spliced EST DA759537 (FIGURE 10). It presents what could be an extra exon sequence between the described reference exon 2 and exon 3 of the gene. It has not been confirmed if this EST with an alternative exon 3 (AE3) prolongs its mRNA until the known 3'end of *THBS4* and encodes for a protein. If so, it might do it from one of the two methionines within the alternative exon 3, since the exon 1 from the reference isoform is not present, and the alternative exon codification produces a shift in the reading frame of the amino acid sequence from the ATG located at exon 2. To try to quantify the specific alternative exon sequences by real-time RT-PCR, we designed primers in the alternative exon 3 and in exon 3 (TABLE 15). However, the low levels, or even lack, of the extra exon in the selected individuals, precluded getting a consistent detection of the alternatively spliced sequences by real-time RT-PCR with SYBR green, making any possible quantification non-reliable.

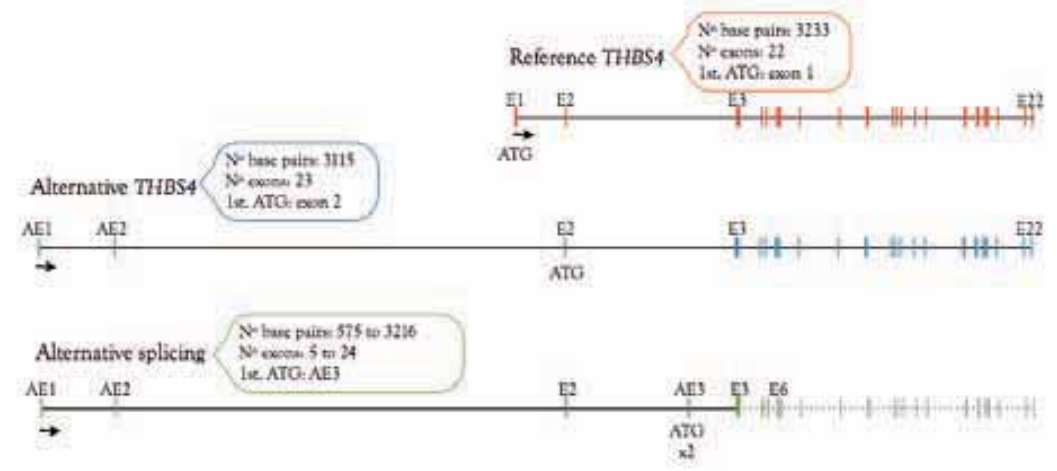



FIGURE 10. Schematic representation of ***THBS4*** isoforms. Representation of both confirmed mRNAs for *THBS4* and the alternative spliced EST DA759537. Localization of the first available methionine codon without breaking the reading frame of the possible proteins is represented by ATG. Non-confirmed sequence of an alternatively spliced isoform is represented in gray. Number of base pairs and exons for the spliced isoform are indicated considering the possibility (or not) that the mRNA is prolonged until the known 3'-end of *THBS4*. PCR amplifications until the E6 have been performed without confirming by sequencing.

## 3.2. Expression analysis of the *THBS4* promoters.

After confirming the presence of an alternative isoform of *THBS4* mRNA, which has not previously been described, we studied where it is expressed, whether it accounts for a significant fraction of the total expression of the gene (especially in brain regions), and if this new isoform is human-specific or generally expressed in non-human primates and whether it shows expression differences between species.

### 3.2.1. *THBS4* isoforms expression in human tissues.

Variation in gene expression patterns across diverse tissues could be related to both functional and anatomical differences in tissue structure and organization, even within the same organ, such as in the brain. To compare expression levels of each *THBS4* transcript and to investigate the potential function of the alternative mRNA isoform in different human tissues, quantitative real-time RT-PCRs of the common 3'-end (FIGURE 8A) and of the isoform-specific 5'-end regions of the two mRNAs were carried out (FIGURE 8B).

Human RNA from 11 diverse non-brain tissues, 9 brain tissues (6 adult and 3 fetal tissues) and 5 cell lines were examined by real-time RT-PCR (see TABLES 3 and 4 in the MATERIALS AND METHODS section for sample description and TABLE 10 for primer sequences). Two independent expression experiments were performed to assure reliability. In order to analyze the usage of each promoter more accurately and to normalize differences in RNA content between samples, we used the housekeeping gene ß-actin (*ACTB*) as a control. Although it is difficult to find a good normalization control across diverse tissues, *ACTB* could give an idea of the relative expression levels of *THBS4* in different tissues. One single run of the glyceraldehyde-3-phosphate dehydrogenase (*GADPH*) gene and the 18S ribosomal RNA real-time RT-PCRs were performed in parallel to test for the differences

between different housekeeping genes and ensure the reliability of *ACTB*. Similar results with the other housekeeping genes compared with *ACTB* was found for the quantification of all the tissues, with the exception of skeletal muscle that was remove from the final experiments. Final values of expression were shown as the number of molecules of *THBS4* gene *per* 1000 molecules of *ACTB* gene expression using a standard curve with a known number of molecules of each transcript.


FIGURE 11. ***THBS4*** gene expression by real-time RT-PCR in diverse human tissues and cell lines. Quantification of mRNA expression levels of total *THBS4* measured at the 3'-end of the gene common between isoforms (A), and alternative and reference isoform of *THBS4* measured at the 5'-end of each mRNA (B). RNA samples used come from 5 human cell lines, 10 human non-brain tissues, and 9 human brain tissues (6 adult tissues and 3 fetal tissues). Graphs represent the average number of *THBS4* molecules of each isoform for $10^3$ molecules of *ACTB* in the *y* axis, with the standard error indicated by error bars.

Expression of total *THBS4* mRNA was previously reported in heart, brain and testicular tissue from analysis of microarray data (CÁCERES *et al.* 2006). Our results after amplifying the 3'-end of *THBS4* indicated that in addition to the three tissues reported previously,

other tissues such as ovary, thymus and colon also express the *THBS4* gene (FIGURE 11A), with skeletal muscle tissue also showing *THBS4* expression, even that it was not considered for the graph. Interestingly, when investigating the 5'-end specific *THBS4* isoforms, the alternative mRNA was expressed in brain tissues and female and male sexual organs beyond the expression levels of the reference mRNA. This includes also most of the cell lines analyzed, which have mainly a nervous tissue origin. In fact, among all the tissues and cell lines analyzed, only colon and heart samples showed higher expression of the reference isoform compared to the alternative isoform (FIGURE 11B). It is important to mention that the expression levels for the 3'-end of *THBS4* were considerably higher than the observed for the quantification of the two isoforms at the 5'-end. It is known that RNA degradation has a 5' to 3' directionality, which could explain these differences.

### 3.2.2. Brain expression of *THBS4* isoforms in primates.

After evaluating the levels of both transcripts of *THBS4* in humans, and determining that the alternative mRNA was the major responsible isoform for the expression of the gene in brain tissues, the next question to answer was whether both isoforms were also expressed in other non-human primates such as chimpanzee and rhesus macaque. Initial tests with semi-quantitative PCR with 3 chimpanzee cDNA and 1 macaque cDNA samples available at that moment, revealed that the alternative *THBS4* isoform was not human specific. Thanks to a first short-term exchange scholar fellowship for a 4-month stay at Emory University (Atlanta, USA), we could perform a proper quantification of *THBS4* expression patterns between species. We carried out real-time RT-PCR of the two isoforms in frontal cortex regions using RNA samples from 11 humans, 11 chimpanzees and 8 macaques in collaboration with the laboratories of Professors James W. Thomas and Todd M. Preuss from Emory University and Yerkes National Primate Research Center (Atlanta, USA). Following the procedure used for the analysis of the different human tissues, the RNA differences were normalized using the housekeeping gene *ACTB* as a control. *THBS4* expression levels were then quantified relative to *ACTB* expression. The results for the quantification of the total *THBS4* mRNA expression showed around 7 and 12 times higher

levels in humans compared to chimpanzees ($P$<0.001) and macaques ($P$<0.001) respectively, similar to the results described by CÁCERES *et al.* (2007). When the expression of the different isoforms was analyzed, the alternative isoform was expressed, respectively, around 5 and 17 times higher in humans relative to chimpanzees ($P$<0.001) and macaques ($P$<0.001), and the reference isoform was expressed, respectively, around 7 and 6 times higher in humans relative to chimpanzees ($P$<0.001) and macaques ($P$<0.001) (FIGURE 12). Although human expression for the alternative isoform is only ~5-fold higher than in the chimpanzee, this difference is comparable with the ~6-fold observed for total *THBS4* expression (CÁCERES *et al.* 2007). Considering that the analysis of both isoforms in humans revealed that brain tissues expresses *THBS4* primarily as the alternative isoform, it is reasonable to suspect the regulation of the alternative isoform as the potential cause responsible for the most part of the differences between species and the higher *THBS4* expression in the human brain. However, it is interesting to note that the expression of both isoforms across species in brain cortex appeared to show the same pattern.

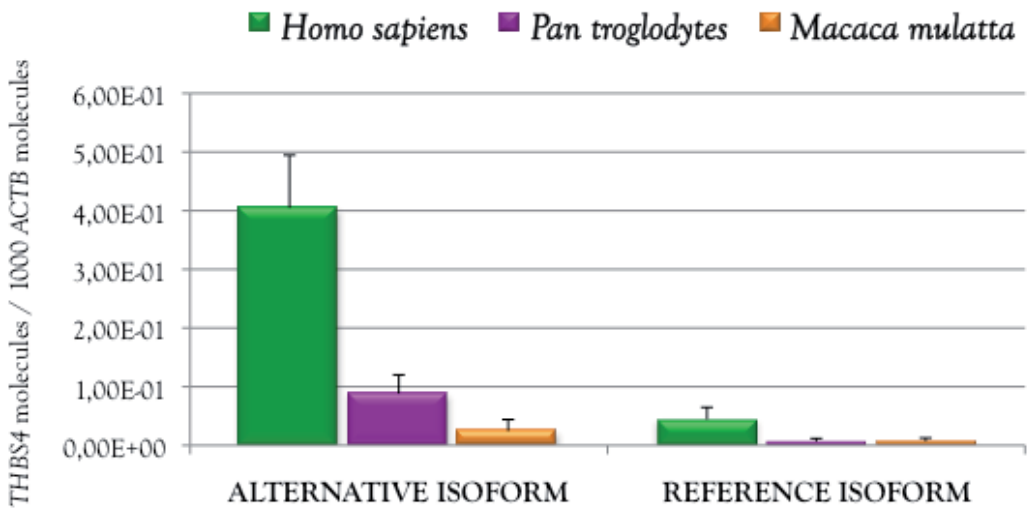



FIGURE 12. ***THBS4*** gene expression among primate frontal cortex by real-time RT-PCR. Quantification of the alternative and reference *THBS4* mRNA expression levels in frontal cortex of 11 humans (*Homo sapiens*), 11 chimpanzees (*Pan troglodytes*) and 8 rhesus macaques *(Macaca mulata)* by real-time RT-PCR. Graphs denote the average number of *THBS4* molecules for $10^3$ molecules of *ACTB* in the *y* axis, with the standard error indicated by error bars.

In the evaluation of *THBS4* expression in the eight samples of rhesus macaques, after performing the real-time RT-PCRs between species, we detected the existence of a single nucleotide change in the macaque sequence where the forward primer specific of the alternative isoform should bind. This sequence change could result in lower amplification efficiency of the *THBS4* alternative isoform in rhesus and thus could affect the expression quantification. Therefore, although the results obtained were in line with the expectations, we considered them not to be reliable for publication (RUBIO-ACERO *et al.,* in prep.).

## 3.3. Possible causes of *THBS4* expression differences.

Searching for the *THBS4* expression differences in the human brain, we have found that most of the *THBS4* present in this tissue come from the alternative mRNA isoform, rather than the reference one. Additionally, we have seen how the alternative isoform is not specific from the human lineage, being also the predominat source of *THBS4* in the frontal cortex of chimpanzees and macaques. Considering the higher *THBS4* expression in both human mRNA isoforms in comparison to the non-primate species, we searched for the possible cause of this expression differences.

### 3.3.1. Interspecific differences in *THBS4* promoters.

In order to collect information about the sequence differences in *THBS4* promoters, we first counted the single nucleotide variations and small indels that have occurred within the gene regulatory region (2 kb upstream and 2 kb downstream of both TSSs) between the genome sequence of humans and chimpanzees, using the rhesus genome as an outgroup, in order to determine the lineage specificity (TABLE 16). To search if the differences between humans and chimpanzees at each side of the TSS showed the same pattern, we performed a

Fisher's Exact Test for both *THBS4* isoforms. As expected, no significant differences were found, suggesting that the changes between species or regions were evenly distributed in both alternative (*P*=0.351) and reference (*P*=0.135) promoter regions analyzed. However, there was a tendency to accumulate more nucleotide changes in the upstream region of the human reference promoter and in the upstream region of the chimpanzee alternative promoter.

TABLE 16: Quantification of human or chimpanzee-specific nucleotide changes at the regulatory sequences of ***THBS4***. Result indicates the distribution of the differences, including single nucleotide variations and small indels between humans and chimpanzees in 2 kb upstream and 2 kb downstream of the alternative and the reference TSSs. In parenthesis, number of single nucleotide variations and number of indels.

| | Alternative promoter | | | Reference promoter | | |
|---|---|---|---|---|---|---|
| | Human lineage | Chimpanzee lineage | Total | Human lineage | Chimpanzee lineage | Total |
| Changes 2 kb upstream | 7 (4,3) | 23 (20,3) | 30 (24,6) | 15 (13,2) | 5 (5/0) | 20 (18,2) |
| Changes 2 kb downstream | 8 (8,0) | 13 (13,0) | 21 (21,0) | 13 (13,0) | 12 (10,2) | 25 (23,2) |
| Total | 15 (12,3) | 36 (33,3) | 51 (45,6) | 28 (26,2) | 17 (15,2) | 45 (41,4) |

## 3.3.2. Quantification of transcriptional activity of *THBS4* promoters.

To determine if the differential expression between humans and non-human primates could be due to differences in transcription, we quantified the transcriptional activity of both promoter regions in humans and chimpanzees by luciferase reporter assays. For the reference isoform of *THBS4* we considered two different fragments of ~1.2 and ~6 kb, including the region upstream of the reference TSS and the first nucleotides of the transcript. The reason for choosing two different lengths was the presence of a human-specific AluY repetitive element around 5.6 kb upstream of the reference TSS that could be

regulating the activity of the promoter. For the alternative isoform, a fragment of ~3 kb upstream of the TSS, including the first nucleotides of the transcript was selected (FIGURE 13). Additionally, it has been found that there are two main different haplotypes in humans along different parts of *THBS4* and upstream region. . To investigate if the transcriptional activity of the alternative promoter was differentially regulated in the different haplotypes, two different humans representing the two haplotype sequences were selected for the reporter assays.

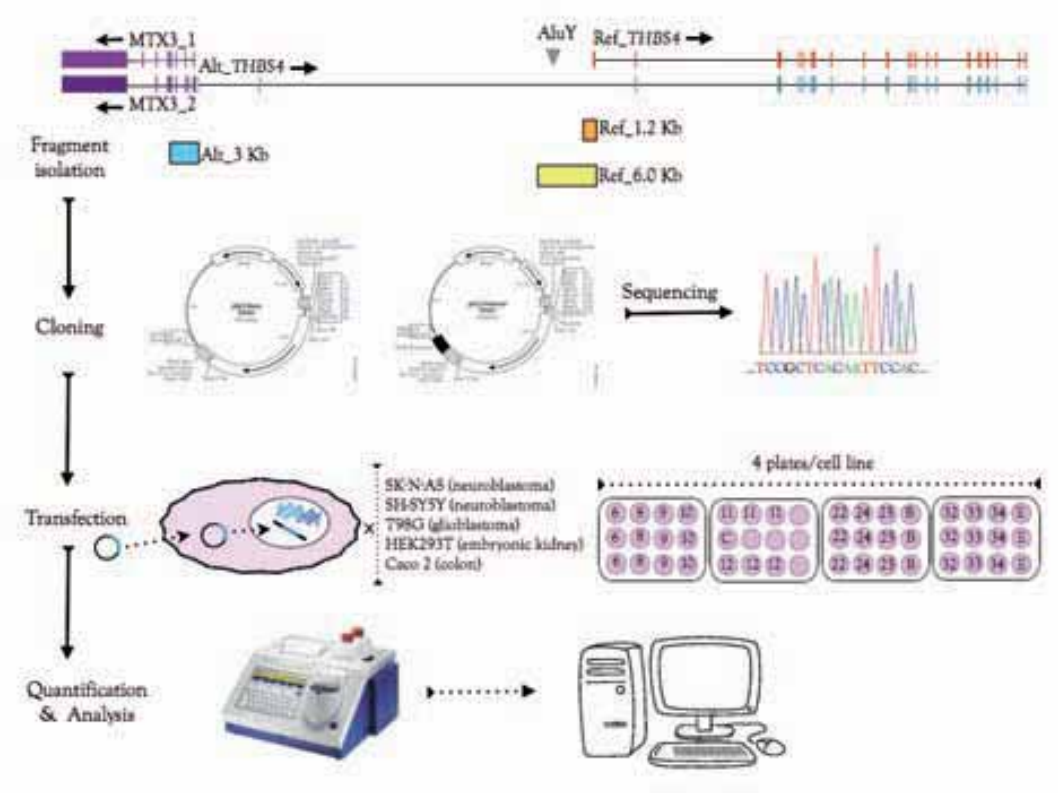



FIGURE 13. Experimental design for the transcriptional activity quantification of *THBS4* promoters. The rectangles in the upper part of the diagram represent the different regions used in the study and their localization: blue for the alternative isoform (in two humans and one chimpanzee samples), and orange and yellow for the reference (in one human and one chimpanzee sample). Fragments of interest were cloned in pGL3-Basic and pGL3-Enhancer vectors. Selected clones were sequenced to ensure that no changes in the sequences have been produced during the amplification and cloning process. All plasmids, with exception of a single pGL3-Control, were transfected in triplicate for each of the five cell lines used. Luciferase quantification and analysis of the results was performed 48 h after transfection. Different steps used in this work are schematized from top to bottom and from left to right. See details of each of these processes in the MATERIALS AND METHODS sections 2.5.1 and 2.8.

With the exception of the plasmids of the reference promoter, which had been constructed previously, the 3 different alternative promoter fragments (2 human and 1 chimpanzee) were cloned in pGL3-Basic and pGL3-Enhancer plasmids (See 2.5 "Cloning and transformation" in MATERIALS AND METHODS). Since pGL3 Luciferase Reporter Vectors do not allow blue/white screening, between 80-100 colonies of the six alternative promoter constructs (4 human and 2 chimpanzee) were checked in order to obtain 2-3 clones for each fragment. Before transfection, 300 ng of the different clones were cut with *BamHI*, *HindIII* and *XhoI*, to verify that the plasmids had the expected size and that no aberrations were produced in the cloning process. In addition, the insert ends of the selected plasmid from each type were sequenced to ensure that no single nucleotide mutations had been introduced either.

Four different types of human cell lines expressing *THBS4* (2 from neuroblastoma, 1 from glioblastoma and 1 from embrionaric kidney) were used to better represent the potential regulatory regimes found in brain cells. Cell lines were grown in 12 well plates and transfected with the 12 plasmids in study, plus the pGL3-Basic and pGL3-Enhancer plasmids used to normalize the values obtained, and pGL3-Control plasmid to ensure that the transfection worked correctly. Four plates with cells were required for each cell line and experiment. Non-transfected wells with grown cells were used as a visual control of the cells to ensure that the transfection did not damage the cells viability during the 48 h until luciferase quantification.

For each cell-line, three different independent experiments were carried out. In each experiment, all plasmids, with the exception of a single pGL3-Control, were transfected in triplicate. Correlation between most of the experiments was very high, with a Pearson's $r \geq 0.95$ for SK-N-AS and HEK293T cell lines and $r \geq 0.99$ for SHSY-5Y cell lines. Regarding the differences in the transcriptional levels between promoters, we found that the human embryonic kidney (HEK293T) cell line and the neuroblastoma cell lines (SK-N-AS and SH-SY5Y) used for the experiment showed a significantly higher luciferase activity for the alternative promoter, compared to the reference promoter ($P_{SK\text{-}N\text{-}AS}<0.001$, $P_{SHSY\text{-}5Y}<0.001$, $P_{HEK293T}<0.001$) (FIGURE 12). These results agree well with the observed expression patterns

for both isoforms in the cell lines. However, transfection into glioblastoma cell lines (T98G), in which correlation between experiments resulted in a Pearson's $r$<0.70, showed really low transcriptional levels for both isoforms, making results variable between the samples and thus indicated low reliability.

Specifically, for the plasmids cloned with 3 kb of the alternative isoform promoter we found no significant differences between the two different humans representing the two haplotype sequences in *THBS4* when cloned in pGL3-Basic ($P_{SK\text{-}N\text{-}AS}$=0.641; $P_{SHSY\text{-}5Y}$=0.591; $P_{HEK293T}$=0.441) or pGL3-Enhancer ($P_{SK\text{-}N\text{-}AS}$=0.504; $P_{SHSY\text{-}5Y}$=0.629; $P_{HEK293T}$=0.100). Considering the promoter activity between species, we found no significant differences between humans and chimpanzees in presence of the universal enhancer ($P_{SK\text{-}N\text{-}AS}$=0.547, $P_{SHSY\text{-}5Y}$=0.264, $P_{HEK293T}$=0.564) or not ($P_{SK\text{-}N\text{-}AS}$=0.583, $P_{SHSY\text{-}5Y}$=0.452, $P_{HEK293T}$=0.626). However, there was a slight tendency of the human plasmids to have more activity than the chimpanzee plasmids, emphasized specially in the plasmids with pGL3-Enhancer (FIGURE 14). Therefore, under these experimental conditions, the 3 kb sequence selected from the human and chimpanzee alternative promoter did not appear to be responsible by itself for the higher expression of *THBS4* detected in the human brain.

On the contrary, with regard to the plasmids with the reference isoform promoter, the chimpanzee fragments had an inclination to have more transcriptional activity compared to the humans ones with and without enhancer (FIGURE 14), even though for the transfections performed in SK-N-AS and SHSY-5Y plasmids this was just a tendency with no significant differences in *THBS4* transcriptional activity between species (pGL3-Basic 1.2 kb fragment: $P_{SK\text{-}N\text{-}AS}$=0.214, $P_{SHSY\text{-}5Y}$=0.355, pGL3-Basic 6.0 kb fragment: $P_{SK\text{-}N\text{-}AS}$=0.370, $P_{SHSY\text{-}5Y}$=0.578, or pGL3-Enhancer 6.0 kb fragment: $P_{SK\text{-}N\text{-}AS}$=0.425, $P_{SHSY\text{-}5Y}$=0.431). For the transfections in HEK293T cell lines we found around 1.5-times significantly higher activity in the chimpanzee fragments cloned in pGL3-Basic (1.2 kb fragment: $P$=0.019; 6.0 kb fragment: $P$=0.027) and pGL3-Enhancer (6.0 kb fragment $P$=0.005) vectors than in human ones (FIGURE 14). This suggests that transcriptional activity of the *THBS4* reference promoter in non-brain tissues might display a distinctive pattern than brain tissues.

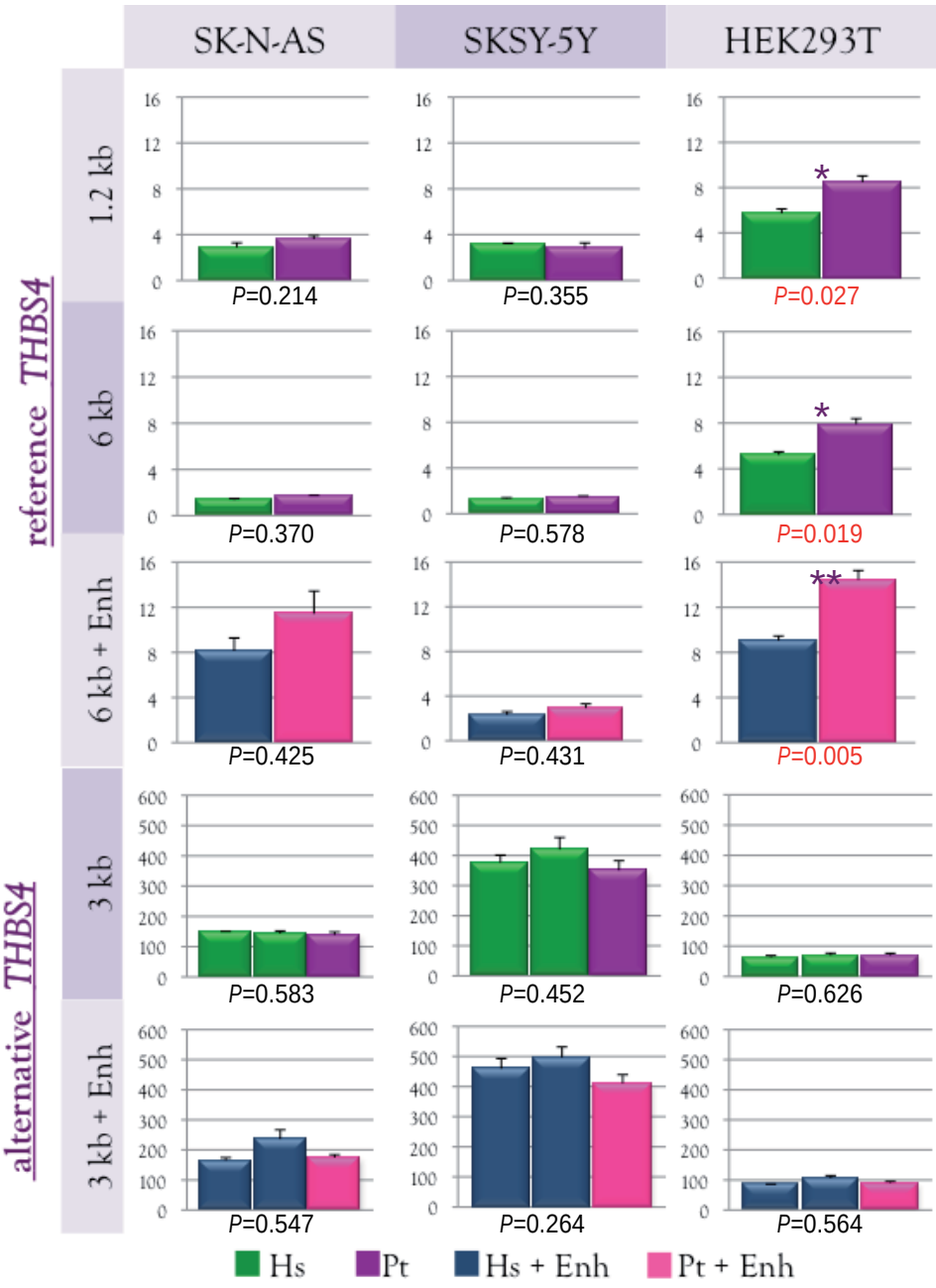



FIGURE 14. Transcriptional activity quantification of *THBS4* promoters in humans and chimpanzees. Luciferase activity in different cell lines transfected with pGL3-Basic and pGL3-Enhancer plasmid with the human (Hs) and chimpanzee (Pt) cloned segments of ~1.2 (no-enhancer) and ~6 kb upstream of the reference promoter and ~3 kb upstream of the alternative promoter (including two different human sequences). See FIGURE 10 for region localization. Graphs represent the average ratio of *firefly* luminescence divided by the *renilla* luminescence in the *y* axis, with the standard error indicated by error bars. For the difference between humans and chimpanzees: *, $P<0.05$; **, $P<0.01$.

Interestingly, in the microarray expression analysis of *THBS4* in different tissues of CÁCERES *et al.* (2003), a similar higher *THBS4* expression in heart in chimpanzees than in humans was found (CÁCERES *et al.* 2007), which is consistent with the higher transcriptional activity of the chimpanzee reference promoter.

After obtaining the first of the luciferase activity datasets in these four cell lines, we were concerned about the low activity of the reference promoter. In fact, only in HEK293T cells the two fragments tested for this promoter showed clear transcriptional activity above background with and without enhancer, whereas there was a clear increase in the transcription of the 6 kb fragment in SK-N-AS with enhancer. To solve this problem, we looked for cell lines from a tissue that primarily expresses the reference isoform. We performed some test experiments in Caco2 cell lines, a colon carcinoma tissue line. Contrary to the expected, luminiscence emitted after the transfection in Caco2 cells was significantly higher for the alternative promoter compared to the reference one (FIGURE 15). Only the 6 kb reference promoter fragment cloned in pGL3-Enhancer vector showed a clear high transcriptional activity in humans and chimpanzees, being this last species the one who shows the higher levels ($P$=0.044). This result correlates well with the significant differences showed in HEK293T cell lines, supporting that the reference isoform in non-brain tissues could be more active in chimpanzee in comparison to humans.

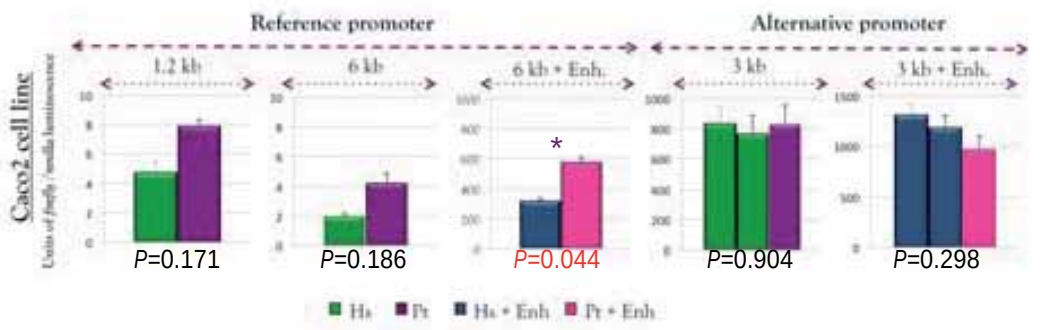



FIGURE 15. Transcriptional activity quantification of ***THBS4*** promoters transfected in Caco2 cell lines. Luciferase activity in Caco2 cell lines transfected with pGL3-Basic and pGL3-Enhancer plasmid with the human (Hs) and chimpanzee (Pt) cloned segments of ~1.2 kb (no-enhancer) and ~6 kb upstream of the reference promoter and ~3 kb upstream of the alternative promoter (including two different human sequences). See FIGURE 10 for region localization. Graphs represent the average ratio of *firefly* luminescence divided by the *renilla* luminescence in the *y* axis, with the standard error indicated by error bars. *$P$<0.05 for the difference between humans and chimpanzees for the 6 kb reference promoter fragment with enhancer

### 3.3.3. Searching for transcription factor binding sites near the *THBS4* gene.

The localization of putative transcription factor binding sites (TFBS) within the promoter region of humans and chimpanzees could be of special interest to understand how the alternative promoter is regulated. In particular, it could give us two important sources of information: TFBS affected by sequence changes between species, and the identification of possible regulators acting in *trans* on the alternative promoter. However, bioinformatic prediction of TFBS is a complicated problem, since the binding matrices of transcription factors tend to be very degenerated and many potential binding sites are usually found, most of which are false positives. To identify potential TFBS, we focused on a region around 5 kb upstream and 2 kb downstream from the alternative *THBS4* promoter region. This 7 kb sequence was then mapped with blat search genome software (http://genome.ucsc.edu) into other mammal species such as chimpanzee, rhesus, marmoset, mouse, cow and dog to obtain their corresponding sequences. We then used the Mulan online program for the alignment of the sequences. This software is integrated with the MultiTF program that identifies evolutionarily conserved TFBS shared by all analyzed species, helping to decode the sequence structure of regulatory elements that are functionally conserved among different species and lowering the false positive rate of TFBS predictions (OVCHARENKO *et al.* 2005). Mulan and MultiTF are publicly available at http://www.dcode.org. The human and chimpanzee genomic sequences were independently used as references of the alignment with the other species. This allows the identification of TFBS conserved in one species but not in the other. All the transcription factor families available by the MultiTF software and a matrix similarity parameter threshold of at least 0.85 were considered for the analysis of both groups of sequences. The TFBS matrix similarity parameter defines the level of identity required between a consensus sequence and the genomic sequence (WINGENDER *et al.* 1996). This threshold usually results in extremely high levels of false-positive TFBS predictions for transcription factor matrices with insufficient experimental evidence for the consensus sequence or for those that have relatively short binding sites, but ensures a uniform level of sequence similarity

between the consensus sequence and detected TFBSs (OVCHARENKO *et al.* 2005). Overall, 142 TFBS were predicted and practically all of them were the same in both species (FIGURE 16). The comparison of the resulting TFBS showed that the alternative promoter has an extra predicted binding site for the EN1 transcription factor around 1.7 kb upstream the TSS when aligning humans with other species and two extra binding sites for the LRF_Q2 and CHCH transcription factors around 4.8 kb upstream the TSS when aligning chimpanzees with the other mammals. Therefore, the high level of conservation of the TFBS between human and chimpanzees agrees well with the lack of differences in promoter activity.

TRANSFAC® is a private knowledge base containing published data on eukaryotic transcription factors, their experimentally-proven binding sites, and regulated genes (http://www.biobase-international.com/). Even though the results are not updated, a snap shot from the 2005 information can be found in the TRANSFAC public database (http://www.gene-regulation.com/). The EN1 transcription factor is encoded by a gene called engrailed homeobox 1 (NM_001426) located in humans in the chromosome 2. EN1 plays an important role in the development of the central nervous system (BELL *et al.* 2012). Epigenetic suppression of EN1 is common to most astrocytomas (WU *et al.* 2010). No specific information was found for the transcription factor LFR_Q2. Two classes of transcription factors answer to the acronym LFR precluding our search for further information: leukemia/lymphoma-related factor, which belongs to a family of transcriptional repressors (AGGARWAL *et al.* 2010), and luman-recruiting factor; which is a endoplasmic reticulum-bound cellular transcription factor (AUDAS *et al.* 2008). Regarding to the transcription factor CHCH, it might refer to a highly conserved zinc finger transcription factor also called Churchill, which is involved in neural induction during the embryogenesis (LEE *et al.* 2007). However, no specific information was found on TRANSFAC® public database. None of these transcription factors have been previously associated to *THBS4* expression.

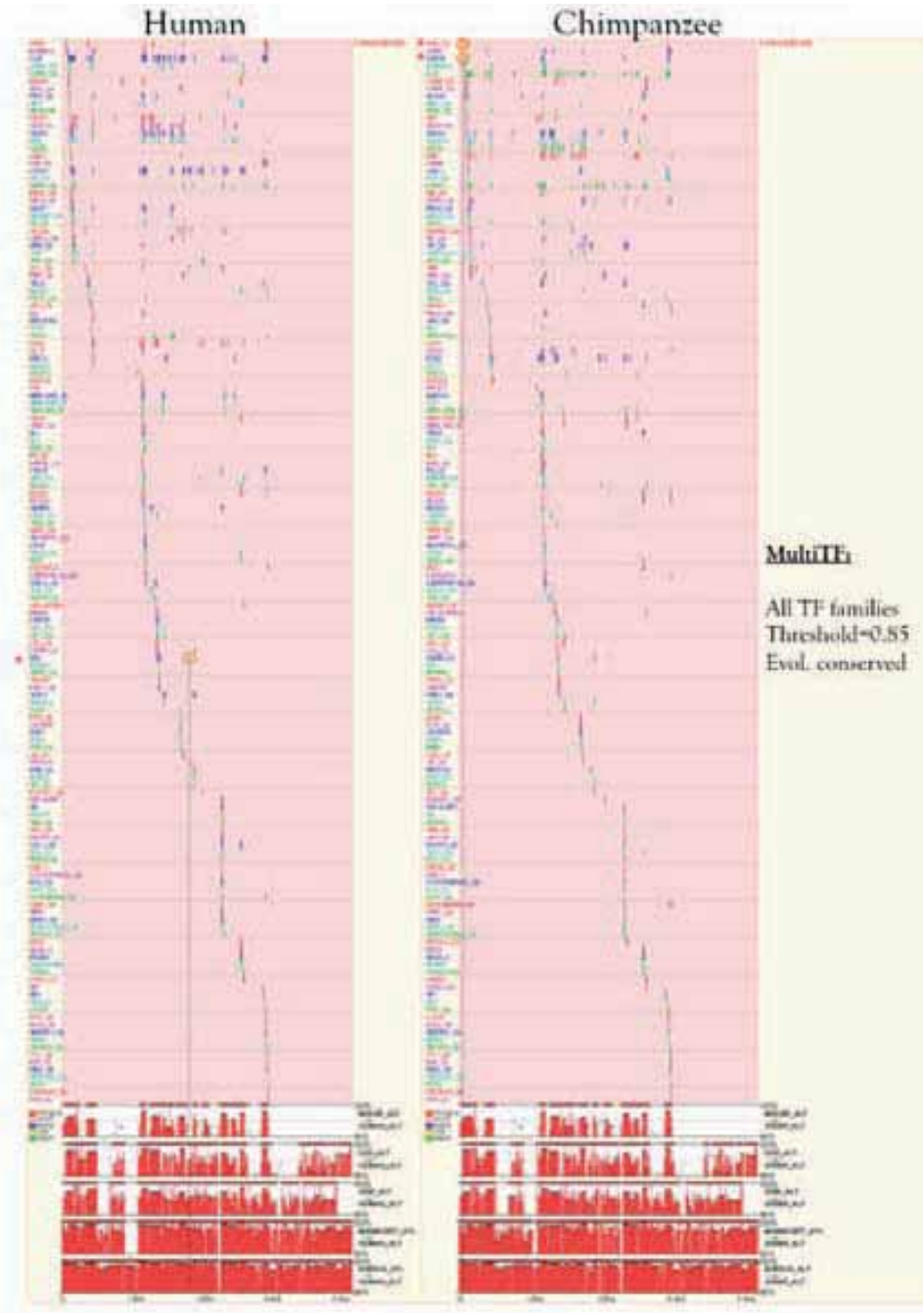



FIGURE 16. Conserved transcription factor binding sites (TFBS) located 5 kb upstream and 2 kb downstream of the alternative *trombospondin-4* isoform TSS. The alignment and search for TFBS conserved among different species allowed the detection of an extra TFBS for EN1 when aligning humans with other species and two extra TFBS, LRF_Q2 and CHCH, when aligning chimpanzees with the other mammals. TFBS specific for humans or chimpanzees are encircled and marked with an asterisk for easier localization. Parameters used in MultiTF analysis included all the TF families available and a threshold of matrix similarity parameter of at least 0.85 to analyze both groups of sequences.

### 3.3.4. Analysis of promoter CpG methylation in primates.

The next step of this study, which was centered exclusively on the characterization of the alternative promoter, involved the possible effect of DNA methylation in 5'-CpG-3' dinucleotides (CpGs) on *THBS4* expression. DNA methylation provides instructions to the gene expression machinery that contribute to determine where and when the gene should be expressed by regulating the promoter. It has been shown that hyper-methylation at promoter CpG islands typically results in a decreased transcription of downstream genes (BAUER *et al.* 2010). In the 5' region of the alternative *THBS4* isoform, overlapping 20 bp with the alternative exon 1, there is a CpG island. This CpG island spans 550 bp (chr5: 79.322.341–79.322.890) with a 66.9% of cytosines or guanines and contains 48 dinucleotides CpG. We decided to investigate whether the CpG island located upstream the alternative promoter was differentially methylated in the brain between humans and chimpanzees. To do so, one of the most common ways to study the methylation status of a genomic DNA sample is the treatment with sodium bisulfite. This treatment converts cytosine residues into uracils but leaves 5-methylcytosines unaffected.

Thanks to a second short-term exchange scholar fellowship for a 3-month stay at Emory University (Atlanta, USA), we could perform sodium bisulfite genomic DNA conversions from humans and chimpanzee samples, using the EZ DNA Methylation gold kit (Zymo Research). Different tests with varying temperature and time of conversion were performed to optimize the sodium bisulfite treatment. Moreover, due to the high temperatures achieved and the different reagents used, after treatment the genomic DNA integrity was often damaged. Therefore, the initial approach to amplify the complete CpG region in a single PCR fragment was not successful. The 550 bp CpG island was finally amplified in two different overlapping fragments of 425 bp and 383 bp, covering in total a region of 728 bp. A pair of primers for each fragment were used with the sequence modified to amplify sodium bisulfite converted DNA (in the forward primer all the C are modified by T and in the reverse primer al the G are modified by A, see FIGURE 7 at MATERIAL AND METHODS),

In addition, a second set of primers with the right sequence to amplify non-converted DNA was used to control the efficiency of the conversion and ensure that only converted genomic DNA is being amplified. Finally, 400 ng of genomic DNA from five human and five chimpanzee samples was treated with sodium bisulfite, amplified in the two different fragments, and cloned in a pGEM®-T Easy Vector System (Promega). PCR through pUC/M13 universal primers was performed to search for positive colonies. Even that this kind of vectors allowed white/blue screening of the positive and negative colonies, the efficiency of the cloning and transformation process was very low, requiring the analysis of 20-40 colonies for each PCR fragment and sample (around 650 colonies in total), to get enough clones for sequencing. At the end, we could perform the alignment and the comparison of a total of 163 independent sequences (83 from humans and 80 from chimpanzees), including 7 to 10 clones sequences of each fragment from the five humans and five chimpanzee genomic DNA samples.

It is known that methylation usually occurs in CpG dinucleotides. In the region analyzed, there were 51 CpGs sites, 48 of which were described as belonging to the CpG island plus other three located in the flanking regions that were included in the amplified fragment. In a first overview, these CpGs showed no high levels of methylation in general and there did not appear to be different levels of methylation between species. Within the 83 human sequences analyzed, only nine show methylation in at least one CpG (10.8%), with only one clone showing nine methylated positions, meanwhile the others have one CpG methylated. On the contrary, chimpanzee sequences showed methylation in 22 different clones from the 80 that were analyzed (27.5%), the methylation within these clones is more spread than in humans, with 16 clones methylated in only one CpG, five clones methylated in two positions and one clone in three different CpG dinucleotides. Considering the 51 different CpG sites analyzed, no methylation was found in any of the clones analyzed among 34 human CpGs and 38 chimpanzee CpGs (TABLE 17). Only two CpG sites (position 30 and 89) of the analyzed region stood out as having a higher level of DNA methylation in comparison to the non- or low-methylated background (FIGURE 17). Particularly, $CpG_{30}$ presented three clones methylated in humans (freq.=0.075) and nine in chimpanzees (freq.=0.2), and the $CpG_{89}$ presented four clones in humans (freq.=0.1) and

two in chimpanzees (freq.=0.045). No significant differences between species were found when these two particular positions were analyzed with a Student's test ($CpG_{30}$: *P*=0.087 and $CpG_{89}$: *P*=0.341). In addition, considering the number of clones analyzed *versus* the number of clones methylated, the frequency of methylation was too low to consider a significant influence of the $CpG_{30}$ or the $CpG_{89}$ in the promoter control.

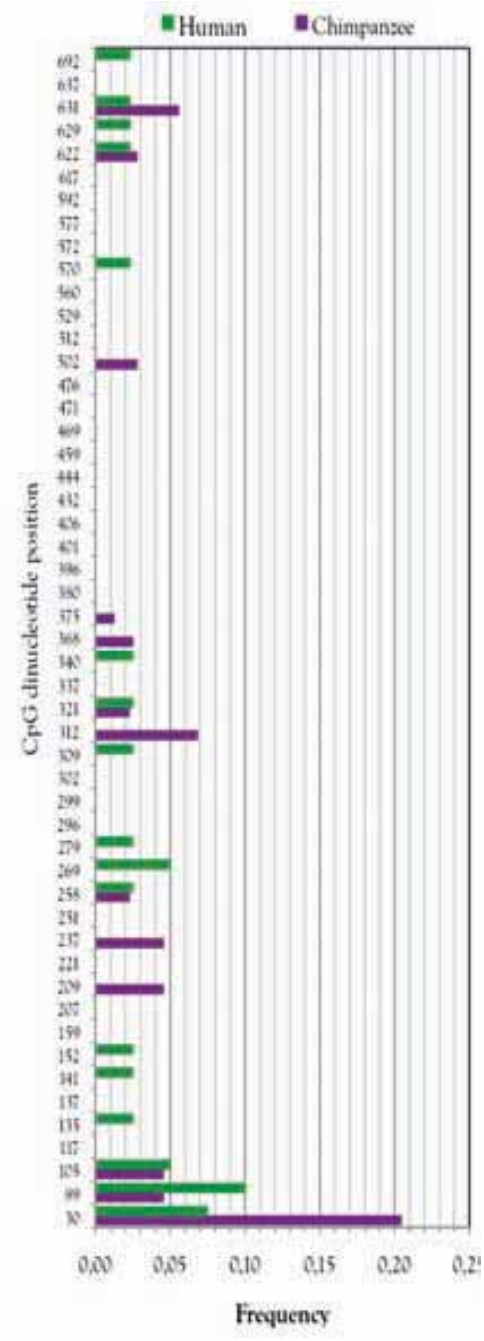



FIGURE 17. Frequency of methylated cytosines of the different CpG positions. Graphical representation of the methylation levels of the 51 CpG sites analyzed from a 550 bp CpG island and its flanking regions upstream the alternative *THBS4* TSS. The sequences of a total of 7 to 10 clones sequences of each fragment from the five humans and five chimpanzee genomic DNA samples were analyzed. Bars indicate the number of clones methylated *versus* the number of clones analyzed (see total values at TABLE 17). Numbers indicate the relative position of the potentially methylated C of the CpGs within the 728 bp region analyzed.

TABLE 17. Methylation levels in the CpG positions analyzed. White numbers indicate the relative position of the potentially methylated C of the CpGs within the 728 bp region analyzed. For each position and sample it is represented the number of clones methylated/total clones analyzed. Total columns summarize the number of clones methylated for each position in human and in chimpanzee samples.

| | | Hs1 | Hs2 | Hs7 | Hs8 | Hs10 | Total | Pt4 | Pt5 | Pt6 | Pt7 | Pt18 | Total |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| CpG FRAGMENT 1 | 30 | 1/8 | 0/8 | 1/7 | 1/7 | 0/10 | 3/40 | 0/8 | 2/10 | 1/9 | 3/7 | 3/10 | 9/44 |
| | 89 | 0/8 | 2/8 | 2/7 | 0/7 | 0/10 | 4/40 | 0/8 | 0/10 | 1/9 | 0/7 | 1/10 | 2/44 |
| | 105 | 1/8 | 0/8 | 1/7 | 0/7 | 0/10 | 2/40 | 1/8 | 0/10 | 1/9 | 0/7 | 0/10 | 2/44 |
| | 117 | 0/8 | 0/8 | 0/7 | 0/7 | 0/10 | 0/40 | 0/8 | 0/10 | 0/9 | 0/7 | 0/10 | 0/44 |
| | 135 | 0/8 | 0/8 | 1/7 | 0/7 | 0/10 | 1/40 | 0/8 | 0/10 | 0/9 | 0/7 | 0/10 | 0/44 |
| | 137 | 0/8 | 0/8 | 0/7 | 0/7 | 1/10 | 1/40 | 0/8 | 0/10 | 0/9 | 0/7 | 0/10 | 0/44 |
| | 141 | 0/8 | 1/8 | 0/7 | 0/7 | 0/10 | 1/40 | 0/8 | 0/10 | 0/9 | 0/7 | 0/10 | 0/44 |
| | 152 | 0/8 | 0/8 | 0/7 | 0/7 | 0/10 | 0/40 | 0/8 | 0/10 | 0/9 | 0/7 | 0/10 | 0/44 |
| | 159 | 0/8 | 0/8 | 0/7 | 0/7 | 0/10 | 0/40 | 0/8 | 0/10 | 0/9 | 0/7 | 0/10 | 0/44 |
| | 207 | 0/8 | 0/8 | 0/7 | 0/7 | 0/10 | 0/40 | 0/8 | 0/10 | 0/9 | 0/7 | 0/10 | 0/44 |
| | 209 | 0/8 | 0/8 | 0/7 | 0/7 | 0/10 | 0/40 | 0/8 | 0/10 | 1/9 | 0/7 | 1/10 | 2/44 |
| | 221 | 0/8 | 0/8 | 0/7 | 0/7 | 0/10 | 0/40 | 0/8 | 0/10 | 0/9 | 0/7 | 0/10 | 0/44 |
| | 237 | 0/8 | 0/8 | 0/7 | 0/7 | 0/10 | 0/40 | 0/8 | 0/10 | 1/9 | 0/7 | 1/10 | 2/44 |
| | 251 | 0/8 | 0/8 | 0/7 | 0/7 | 0/10 | 0/40 | 0/8 | 0/10 | 0/9 | 0/7 | 0/10 | 0/44 |
| | 258 | 0/8 | 0/8 | 1/7 | 0/7 | 0/10 | 1/40 | 0/8 | 0/10 | 1/9 | 0/7 | 0/10 | 1/44 |
| | 269 | 0/8 | 0/8 | 1/7 | 0/7 | 1/10 | 2/40 | 0/8 | 0/10 | 0/9 | 0/7 | 0/10 | 0/44 |
| | 279 | 0/8 | 0/8 | 1/7 | 0/7 | 0/10 | 1/40 | 0/8 | 0/10 | 0/9 | 0/7 | 0/10 | 0/44 |
| | 296 | 0/8 | 0/8 | 0/7 | 0/7 | 0/10 | 0/40 | 0/8 | 0/10 | 0/9 | 0/7 | 0/10 | 0/44 |
| | 299 | 0/8 | 0/8 | 0/7 | 0/7 | 0/10 | 0/40 | 0/8 | 0/10 | 0/9 | 0/7 | 0/10 | 0/44 |
| | 302 | 0/8 | 0/8 | 0/7 | 0/7 | 0/10 | 0/40 | 0/8 | 0/10 | 0/9 | 0/7 | 0/10 | 0/44 |
| | 309 | 0/8 | 0/8 | 1/7 | 0/7 | 0/10 | 1/40 | 0/8 | 0/10 | 0/9 | 0/7 | 0/10 | 0/44 |
| | 312 | 0/8 | 0/8 | 0/7 | 0/7 | 0/10 | 0/40 | 3/8 | 0/10 | 0/9 | 0/7 | 0/10 | 3/44 |
| | 321 | 0/8 | 0/8 | 1/7 | 0/7 | 0/10 | 1/40 | 0/8 | 1/10 | 0/9 | 0/7 | 0/10 | 1/44 |
| | 337 | 0/8 | 0/8 | 0/7 | 0/7 | 0/10 | 0/40 | 0/8 | 0/10 | 0/9 | 0/7 | 0/10 | 0/44 |
| | 340 | 0/8 | 0/8 | 1/7 | 0/7 | 0/10 | 1/40 | 0/8 | 0/10 | 0/9 | 0/7 | 0/10 | 0/44 |
| CpG F1 & F2 | 368 | 0/15 | 0/16 | 0/17 | 0/15 | 0/20 | 0/83 | 0/8 | 1/19 | 0/18 | 1/17 | 0/18 | 2/80 |
| | 375 | 0/15 | 0/16 | 0/17 | 0/15 | 0/20 | 0/83 | 0/8 | 0/19 | 0/18 | 0/17 | 1/18 | 1/80 |
| | 380 | 0/15 | 0/16 | 0/17 | 0/15 | 0/20 | 0/83 | 0/8 | 0/19 | 0/18 | 0/17 | 0/18 | 0/80 |
| | 396 | 0/15 | 0/16 | 0/17 | 0/15 | 0/20 | 0/83 | 0/8 | 0/19 | 0/18 | 0/17 | 0/18 | 0/80 |
| | 401 | 0/15 | 0/16 | 0/17 | 0/15 | 0/20 | 0/83 | 0/8 | 0/19 | 0/18 | 0/17 | 0/18 | 0/80 |
| | 406 | 0/15 | 0/16 | 0/17 | 0/15 | 0/20 | 0/83 | 0/8 | 0/19 | 0/18 | 0/17 | 0/18 | 0/80 |
| CpG FRAGMENT 2 | 432 | 0/7 | 0/8 | 0/10 | 0/8 | 0/10 | 0/43 | 0/0 | 0/9 | 0/9 | 0/10 | 0/8 | 0/36 |
| | 444 | 0/7 | 0/8 | 0/10 | 0/8 | 0/10 | 0/43 | 0/0 | 0/9 | 0/9 | 0/10 | 0/8 | 0/36 |
| | 459 | 0/7 | 0/8 | 0/10 | 0/8 | 0/10 | 0/43 | 0/0 | 0/9 | 0/9 | 0/10 | 0/8 | 0/36 |
| | 469 | 0/7 | 0/8 | 0/10 | 0/8 | 0/10 | 0/43 | 0/0 | 0/9 | 0/9 | 0/10 | 0/8 | 0/36 |
| | 471 | 0/7 | 0/8 | 0/10 | 0/8 | 0/10 | 0/43 | 0/0 | 0/9 | 0/9 | 0/10 | 0/8 | 0/36 |
| | 476 | 0/7 | 0/8 | 0/10 | 0/8 | 0/10 | 0/43 | 0/0 | 0/9 | 0/9 | 0/10 | 0/8 | 0/36 |
| | 502 | 0/7 | 0/8 | 0/10 | 0/8 | 0/10 | 0/43 | 0/0 | 0/9 | 1/9 | 0/10 | 0/8 | 1/36 |
| | 512 | 0/7 | 0/8 | 0/10 | 0/8 | 0/10 | 0/43 | 0/0 | 0/9 | 0/9 | 0/10 | 0/8 | 0/36 |
| | 529 | 0/7 | 0/8 | 0/10 | 0/8 | 0/10 | 0/43 | 0/0 | 0/9 | 0/9 | 0/10 | 0/8 | 0/36 |
| | 560 | 0/7 | 0/8 | 0/10 | 0/8 | 0/10 | 0/43 | 0/0 | 0/9 | 0/9 | 0/10 | 0/8 | 0/36 |
| | 570 | 1/7 | 0/8 | 0/10 | 0/8 | 0/10 | 1/43 | 0/0 | 0/9 | 0/9 | 0/10 | 0/8 | 0/36 |
| | 572 | 0/7 | 0/8 | 0/10 | 0/8 | 0/10 | 0/43 | 0/0 | 0/9 | 0/9 | 0/10 | 0/8 | 0/36 |
| | 577 | 0/7 | 0/8 | 0/10 | 0/8 | 0/10 | 0/43 | 0/0 | 0/9 | 0/9 | 0/10 | 0/8 | 0/36 |
| | 592 | 0/7 | 0/8 | 0/10 | 0/8 | 0/10 | 0/43 | 0/0 | 0/9 | 0/9 | 0/10 | 0/8 | 0/36 |
| | 617 | 0/7 | 0/8 | 0/10 | 0/8 | 0/10 | 0/43 | 0/0 | 0/9 | 0/9 | 0/10 | 0/8 | 0/36 |
| | 622 | 0/7 | 1/8 | 0/10 | 0/8 | 0/10 | 1/43 | 0/0 | 0/9 | 0/9 | 0/10 | 1/8 | 1/36 |
| | 629 | 1/7 | 1/8 | 0/10 | 0/8 | 0/10 | 2/43 | 0/0 | 0/9 | 0/9 | 0/10 | 0/8 | 0/36 |
| | 631 | 0/7 | 0/8 | 0/10 | 0/8 | 0/10 | 0/43 | 0/0 | 1/9 | 1/9 | 0/10 | 0/8 | 2/36 |
| | 637 | 0/7 | 0/8 | 0/10 | 0/8 | 0/10 | 0/43 | 0/0 | 0/9 | 0/9 | 0/10 | 0/8 | 0/36 |
| | 692 | 0/7 | 0/8 | 0/10 | 0/8 | 1/10 | 1/43 | 0/0 | 0/9 | 0/9 | 0/10 | 0/8 | 0/36 |

## 3.4. Search and analysis of enhancers.

Among the *cis*-regulatory elements known to have influence over the expression levels of a gene, enhancers play one of the most important roles. Being linked through activator proteins to transcription factor subunits, enhancers increase the rate of transcription of promoters that can be located quite far away. In addition, many enhancers scattered around the genome can bind different activators and provide a complex variety of responses to diverse cell signals. In fact, they can be localized anywhere within the chromosomes with respect to the genes they regulate, with distances over megabases in some cases (VISEL *et al.* 2009).This feature makes enhancers a tricky element to be detected for a gene regulation study. In this work, we undertook two different approaches for the localization of enhancers combining both computational and experimental techniques.

### 3.4.1. ChIP-Sequencing approach.

The first method used to detect possible enhancers around the *THBS4* gene was based in ChIP-Seq experiments (chromatin immunoprecipitation coupled to high-throughput sequencing) in three different human cell lines expressing *THBS4*: two from neuroblastoma (SH-SY5Y and SK-N-AS) and one from glioblastoma (T98G). We chose to immunoprecipitate the chromatin with the enhancer-associated protein p300. It has been suggested that this protein could bind very specifically to enhancers and thus would be capable to localize their associated activities (VISEL *et al.* 2009). Since much of the published information about binding sites were carried out in HeLa cells, the chromatin of this cell line was also immunoprecipitated as a control, but it was not considered for the final high-throughput sequencing.

Before the immunoprecipitation of the chromatin, the grown cell lines were crosslinked with formaldehyde to ensure that the chromatin structure was preserved during the

isolation and chromatin-immunoprecipitation procedure. To have an optimal immunoprecipitation, chromatin in cell lysates was exposed for 17 minutes to a hydrodynamic shearing to get fragments of 200-500 bp. It is important to note that this fragment size is essential for both sequencing and the specificity of the assay. Longer sizes could pull down in the precipitation process DNA in which the binding site of the protein of interest may be located a significant distance away from it. Finally, a total of 30 µg of chromatin was immunoprecipitated using Dynabeads® Protein A (Invitrogen) coupled with the antibodies p300 and IgG (used as negative control). A chromatin input control (non-immunoprecipitated) was also used as reference control. Two different chromatin immunoprecipitation experiments were carried out for each cell line plus one in HeLa cells that were used as control for the different validations.

The amount of immunoprecipitated DNA recovered is dependent on many factors, including the antibody epitope and the protein binding-site accessibility. To validate the different ChIP experiments, we selected six different ~500 bp segments from regions that had been described to bind p300 in diverse cell types (HEINTZMAN *et al.* 2007, HEINTZMAN *et al.* 2009, VISEL *et al.* 2009) centered approximately around the highest p300 HeLa cells peak (the most common used cell line for these experiments) and the most conserved region between species. In addition, two extra regions that were not reported to have a binding site for p300 were also selected as negative controls. Primers amplifying 50-150 bp in each of these regions were designed to confirm the presence of the target DNA by real-time PCR (TABLE 12 in MATERIALS AND METHODS). These PCRs allowed the quantification of the DNA immunoprecipitated with p300 and IgG antibodies and the quantification of the DNA from the input samples that were not immunoprecipitated. As expected, input samples presented the higher amounts of DNA, since they are amplifying the total genomic DNA and not only the specific inmunoprecipitated regions. On the other hand, amplification levels in the p300 and IgG immunoprecipitated samples were variable between the different regions and cell lines, reaching sometimes higher DNA levels in the negative controls than in the putative positive control regions. However, there was a tendency of the p300 immunoprecipitated DNA to show higher amplification levels than the IgG in positive controls and lower in the negative controls.

Next, DNA from 2 independent ChIP reactions for each cell line were subjected to deep sequencing using the Solexa Genome Analyzer (Illumina) to detect the binding sites for p300. Additionally, for each cell line a homogenate of the input samples of the two experiments was also subjected to deep sequencing. Since input had been isolated from the same cells and had been cross-linked and fragmented under the same conditions as the immunoprecipitated DNA, it provided information about the amount of DNA that was available in the moment before the immunoprecipitation. Around 16-21 million reads were sequenced for each cell line and experiment. Identical read sequences were removed to eliminate duplicates, reducing the number of different sequences, and afterwards, sequences were mapped against the HG19 genome (TABLE 18). In total, between 10 and 13 million different mapped sequences were obtained from each experiment, with the exception of one Sh-SY5Y reaction that had much lower reads (3.3 million). In this case, it is possible that the small number of independent reads was produced by a low amount of DNA, since for this sample only 5.6 ng of DNA were recovered for deep sequencing when the recommended amount is 10 ng and never below 6 ng.

Single reads mapped against the human genome allowed the study and comparison with overlapping genes and regulatory regions. All the estimated regions with coverage equal or higher than ten were initially considered for the detection of p300 binding sites. After that, regions that were also detected in input negative control were discarded. The sequencing of most of the cell lines provided from 200 to 500 clusters of mapped sequences, which could give a reliable localization of binding sites for p300, with the exception of the experiment 1 of SHSY-5Y. Mapped sequences were distributed over TSS regions (-150 bp downstream and +50 bp upstream of known TSS), promoter regions (up to 1 kb upstream of known TSS), and introns or defined genes (TABLE 19).

TABLE 18: ChIP-Seq data. This table summarizes the number of reads obtained for the different ChIP-seq experiments with p300. The different numbers of read sequences correspond with those present before and after removing the repeated ones and only the different non-redundant sequences mapped to the genome (HG19 version).

| | T98G (1) | T98G (2) | SN-N-AS (1) | SN-N-AS (2) | SH-SY5Y (1) | SH-SY5Y (2) |
|---|---|---|---|---|---|---|
| Number of reads | 19,507,382 | 15,986,648 | 18,572,872 | 17,007,444 | 21,072,233 | 18,460,769 |
| Different sequences | 14,323,619 | 15,053,136 | 15,091,026 | 15,626,464 | 6,066,763* | 17,447,990 |
| Mapped diff. seq. | 10,324,256 | 11,169,265 | 11,653,094 | 11,658,841 | 3,384,095* | 13,186,144 |

* ChIP-seq experiments with a much lower number of different and mapped read sequences.

TABLE 19: ChIP-Seq data analysis results. Defined clusters in the diverse ChIP-seq experiments. The different sequences were mapped over TSS regions (TSS -150 bp downstream and +50 bp upstream of known TSS), promoter regions (up to 1 kb upstream of TSS), introns or genes.

| | T98G_1 | T98G_2 | SN-N-AS_1 | SN-N-AS_2 | SH-SY5Y_1 | SH-SY5Y_2 |
|---|---|---|---|---|---|---|
| Clusters | 1,036 | 484 | 347 | 221 | 86,642 | 432 |
| TSS regions | 6 | 1 | 13 | 1 | 2,304 | 2 |
| Promoter regions | 9 | 3 | 1 | 0 | 1,452 | 5 |
| Introns | 411 | 161 | 129 | 68 | 33,599 | 155 |
| Genes | 453 | 189 | 143 | 81 | 39,604 | 192 |

With the information about the localization of the different binding sites for p300 we could specifically check for those located around the *THBS4* gene at chromosome 5. We focused on a region of 50 kb upstream and downstream of the alternative *THBS4* gene, covering a final region of 200 kb. None of the three cell lines used in this study presented a potentially interesting region around *THBS4* replicated in different experiments to be selected for an in-depth experimental analysis. Unfortunately, the ChIP-seq experiments had a lot of background that made very difficult to identify reliable and consistent p300 peaks, and it would be necessary to optimize the protocol to repeat experiments again.

### 3.4.2. Computational prediction of enhancers.

The second approach carried out to determine possible enhancers that could be related with the higher expression of *THBS4* alternative promoter in the human brain, was based on a computational search using a compilation of information of data published from different sources. As starting point, we focused on a region covering 50 kb upstream and downstream of the alternative promoter of *THBS4.* These around 200 kb around the gene were divided in windows of 5 kb for a detailed analysis. We applied two different criteria for the selection of the putative enhancer regions. The first one was based on the ENCODE regulation tracks, which contain information relevant to the regulation of transcription from the ENCODE project, such as the location of modifications of histone proteins that are suggestive of enhancers, chromatin regions hypersensitive to digestion by the DNase enzyme, or sequences bound by proteins responsible for modulating gene transcription using specific antibodies to different transcription factors. The second criterion considered the chromatin state segmentation described in the UCSC track ChromHMM to detect regions of weak or strong enhancers (see 2.11 "Bioinformatic prediction of enhancers" in MATERIALS AND METHODS for additional information). In addition to this, we took advantage of the information about evolutionary conserved regions in the genome available in the ECR Browser on-line interface. This includes the genomes of 13 species that can be aligned to create a conservation profile to identify regions of higher sequence identity than the neutrally evolving background. The different windows were thus aligned in zebrafish, chicken, frog, rat, mouse, opossum, cow, dog, macaque and chimpanzee using the human sequence as reference. An overall idea about the conservation of the different regions and genomic elements were annotated and considered in the selection of the candidate enhancer regions.

The first criterion for selecting an enhancer site was based on three ENCODE regulation tracks about enhancer and promoter associated histone marks, regulatory regions hypersensitive to DNase I, or binding sites of transcription factors detected by ChIP-seq.

After applying the first criterion, a total of 17 regions[2] passed the thresholds for at least two of the three regulation tracks and were considered as possible candidates to act as an enhancer (TABLE 20). When a region was selected, the final 5 kb window was chosen by centering it on the most likely enhancer position. However, as these are too many regions to validate one by one experimentally, we applied the second criterion that provides more information about the selected windows and helps with the screening. This ChromHMM track displayed the chromatin state segmentation for each of nine human cell lines based on the prediction of diverse functional elements by ERNST *et al.* (2011), which divided the genome in fifteen states that were grouped and highlighted at the UCSC browser.

According to the ChromHMM information, from the 17 candidate regions, we discarded three that were labeled as promoter regions (Enhancer region 5, 6 and 10) and one as an insulator (Enhancer region 11). In addition, if the candidate regions were not labeled as enhancer in at least three of the nine cell lines (Enhancer region 2, 8, 9, 12, 13, 14 and 16), they were also discarded for the experimental validation. This removed 11 candidates from the initial list, leaving only six left to perform the experimental analysis. A seventh region upstream of the reference *THBS4* mRNA was selected, even though it did not fulfill the two previous criteria, since it included an ALUY element insertion in the human genome that is not present in other non-human primates, and could thus be regulating the transcriptional activity of the reference or alternative *THBS4* promoter.

[2] Screenshots from the UCSC tracks and conservation between species of the different enhancer regions can be found at APPENDIX I.

TABLE 20: Regions selected as candidate enhancers. Localization of the diverse candidates (final 5 kb centered region) and summary of the criteria used for their selection. The three ENCONDE regulatory tracks included H3K4Me1 ChIP data (Enhancer- and Promoter-Associated Histone Mark), which is fulfilled if the peak has a value of 50 or more over a maximum value of 100, DNaseI HS (Digital DNaseI Hypersensitivity Clusters), and detection of transcription factor binding sites by ChIP-seq (Transcription Factor ChIP-seq), both fulfilled with the presence of at least one cluster. The ChromHMM track displays the functional elements predicted in the regions according to ERNST *et al.* (2011) (in brackets the number of cell lines that present them of the nine in studied is indicated), and the conservation information based in the ECR browser.

| Enhancer region | Position (chr5) | H3K4Me1 peak | DNaseI HS clusters | TFBS by ChIP-seq | ChromHMM classification | Conservation |
|---|---|---|---|---|---|---|
| Enh1 | 79,285,625<br>79,290,624 | ≥80 | 2 | 8 TF most of them detected in Hela cells. | Strong enhancer (2/9)<br>Weak Enhancer (3/9)<br>Repressed (1/9) | Small conservation peak up to opossum around position +1500. |
| Enh2 | 79,291,375<br>79,296,374 | ≥60 | 3 | — | Unclear. Weak enhancer (4/9),weak transcribed (2/9), insulator (1/9) or poised promoter (1/9). | Conservation peaks up to opossum around positions +750 and +4000 |
| Enh3 | 79,300,375<br>79,305,374 | ≥60 | 1 | — | Weak enhancer (5/9)<br>repressed (1/9) | Small conservation peak up to opossum around region +300. Longer intergenic region (+2500) with transposable elements or simple repeats conserved up to opossum. |
| Enh4 | 79,304,501<br>79,309,500 | ≥80 | 4 | 2 TF detected in Hela cells. | Strong enhancer (3/9)<br>weak enhancer (3/9)<br>repressed (1/9) | Peak of conservation up to mouse/rat around region +1750 (intergenic) and +37500 (*MTX* 3 gene UTR) |
| Enh5 | 79,320,126<br>79,325,125 | ≤50 | 3 | 23 TF detected in blood, liver and Hela cells | Strong promoter (8/9)<br>weak enhancer (1/9) | *MTX3* exons and adjacent intronic regions widely conserved. |
| Enh6 | 79,364,251<br>79,369,250 | <50 | 3 | 2 TF detected in blood and embryonic stem cells. | Poised promoter (4/9)<br>repressed (4/9) | UTR and exon 1 of reference *THBS4* conserved up to opossum |

(Continuation TABLE 20)

| Enhancer Region | Position (chr5) | H3K4Me1 peak | DNaseI HS clusters | TFBS by ChIP-seq | ChromHMM classification | Conservation |
|---|---|---|---|---|---|---|
| Enh7 | 79,385,126<br>79,390,125 | ≥60 | 4 | 2 TF detected in blood cells | Weak enhancer (4/9), strong enhancer (1/9), repressed (2/9) | *THBS4* exons well conserved up to opossum |
| Enh8 | 79,399,376<br>79,404,375 | ≤50 | 4 | 6 TF detected in *blood* cells | Weak enhancer (1/9)<br>repressed (1/9) | *THBS4* exons well conserved up to zebrafish |
| Enh9 | 79,407,001<br>79,412,000 | ≤50 | 2 | 4 TF in blood cells | Unclear. Weak enhancer (2/9), weak transcribed (2/9), repressed (1/9) | *THBS4* exons well conserved up to zebrafish. |
| Enh10 | 79,412,626<br>79,417,625 | ≤50 | 4 | 4 TF in blood cells | Unclear. Weak enhancer (8/9), weak promoter (6/9) or insulator (9/9) | *THBS4* exons and UTR well conserved up to chicken. |
| Enh11 | 79,421,501<br>79,426,500 | <50 | 5 | 10 TF in blood, liver and Hela cells | Insulator (8/9)<br>Weak enhancer (1/9) | Low conservation far from primates. Small conservation peak up to mouse around region +2000. |
| Enh12 | 79,427,251<br>79,432,250 | ≤50 | 2 | — | Weak-strong enhancer (1/9), insulator (2/9) weak enhancer (1/9) | Short conservation peaks up to mouse around regions +2250. and +3250 |
| Enh13 | 79,432,001<br>79,437,000 | <50 | 2 | 1 TF in blood cells | Weak enhancer (1/9) repressed (1/9) | Low conservation far from primates. Small conservation peak up to mouse around regions +1500 and +3500. |
| Enh14 | 79,435,376<br>79,440,375 | <50 | 2 | 5 TF in liver cells | Strong enhancer (1/9) repressed (2/9) | Low conservation far from primates. Small conservation peak up to mouse around regions +250 and +4000. |

(Continuation TABLE 20)

| Enhancer Region | Position (chr5) | H3K4Me1 peak | DNaseI HS clusters | TFBS by ChIP-seq | ChromHMM classification | Conservation |
|---|---|---|---|---|---|---|
| Enh15 | 79,440,126<br>79,445,125 | ≤50 | 6 | 6 TF in blood and neuroblastoma cells | Weak enhancer (3/9)<br>repressed (4/9) | More or less conserved up to mouse/rat in a region between +1000 to +2500 |
| Enh16 | 79,451,501<br>79,456,500 | <50 | 1 | 1 TF in blood cells | Unclear. Weak enhancer (1/9), weak transcribed (1/9) or repressed.(2/9) | Low conservation far from primates. Disperse regions conservation up to cow |
| Enh17 | 79,455,751<br>79,460,750 | <50 | 5 | >30 TF in blood and liver cells. | Strong enhancer (2/9), weak Enhancer (4/9) insulators (2/9) | Peak of conservation up to mouse in a region around +3750 |
| EnhALUY | 79,358,501<br>79,363,500 | <50 | 3 | — | Repressed (3/9) | Peak of conservation up to opossum in a region around +2250. Around region +3000 there is a human-specific AluY element not conserved in other species. |

### 3.4.3. Experimental validation of the predicted enhancers.

Around 3 kb of the final seven candidate enhancer regions (Enhancer region 1, 3, 4, 7, 15, 17 and ALUY) were amplified by PCR from DNA of human NA12872 and chimpanzee BAC RP43-41P12. The enhancer candidate inserts were cloned downstream of the reporter gene into a pGL3-Basic vector that contained the previously cloned putative alternative promoter for each species. Unfortunately, this was not accomplished for any species in the Enhancer region 1 and the Enhancer region 17, and for the chimpanzee Enhancer region 3, all of which presented problems during the PCR optimization or the cloning. After transformation, to detect the presence/absence of the insert (enhancer), colonies were screened by PCR with the primers used to obtain the enhancer inserts in the first place (TABLE 9 in MATERIALS AND METHODS). Since the cloning process was carried out by cutting the vector from a single restriction enzyme site (*SalI*), the enhancer insert could be positioned in two directions within the plasmid. Two primers in opposite direction were designed at the extremes of each insert. PCR from each of these primers to the universal RVprimer4 present in the pGL3-basic vector were performed to check the insert's directionality (TABLE 21). Positive clones inserted in the same direction than *THBS4* were sequenced from the forward primer used for the insert amplification and from the universal RVprimer4 of the pGL3-basic vector. These primers cover the major part of the inserted enhancer, confirming the direction on which it was inserted and providing information about the absence of nucleotide changes during PCR.

TABLE 21: Sequence and combination of primers used to determine the clonation direction of the putative enhancers

| Primer name | Sequence (5'-> 3') | Amplicon (bp) | Gene/Region |
|---|---|---|---|
| Enh_3_direc1 | GCTCAGTGCTGGGGTACAAT | 219 (Hs & Pt) | Enhancer Region 3 |
| RvPrimer4 | GACGATAGTCATGCCCCGCG | | |
| Enh_4_direc1 | CTATCTGGGTGAGCAAAAGG | 198 (Hs & Pt) | Enhancer Region 4 |
| RvPrimer4 | GACGATAGTCATGCCCCGCG | | |
| Enh_7_direc1 | CAGAGAAGCAGGGTGTAAGCA | 186 (Hs & Pt) | Enhancer Region 7 |
| RvPrimer4 | GACGATAGTCATGCCCCGCG | | |
| Enh_15_direc1 | CACTGAGATGTTTGGGTGGA | 219 (Hs) 221 (Pt) | Enhancer Region 15 |
| RvPrimer4 | GACGATAGTCATGCCCCGCG | | |
| Enh_17_direc1 | GGTCCCAGGATGGAAAACTT | 173 (Hs) 184 (Pt) | Enhancer Region 17 |
| RvPrimer4 | GACGATAGTCATGCCCCGCG | | |
| Enh_ALU_direc1 | CTTGATCTTGGCTGGCTTTG | 203 (Hs) 202 (Pt) | Enhancer Region ALUY |
| RvPrimer4 | GACGATAGTCATGCCCCGCG | | |
| Enh_3_direc2 | GGCAGAGGGCACACATAGAG | 207 (Hs & Pt) | Enhancer Region 3 |
| RvPrimer4 | GACGATAGTCATGCCCCGCG | | |
| Enh_4_direc2 | GCTAACAAAGCCCTGCTCAA | 200 (Hs) 198 (Pt) | Enhancer Region 4 |
| RvPrimer4 | GACGATAGTCATGCCCCGCG | | |
| Enh_7_direc2 | CAAAAGAGCGATCCTGTCTCA | 345 (Hs) 340 (Pt) | Enhancer Region 7 |
| RvPrimer4 | GACGATAGTCATGCCCCGCG | | |
| Enh_15_direc2 | TTATTCCAGGGCAACCTCCT | 543 (Hs) 541 (Pt) | Enhancer Region 15 |
| RvPrimer4 | GACGATAGTCATGCCCCGCG | | |
| Enh_17_direc2 | TCCTGAGCTTGAGGTTCCAT | 248 (Hs & Pt) | Enhancer Region 17 |
| RvPrimer4 | GACGATAGTCATGCCCCGCG | | |
| Enh_AluY_direc2 | GCCTCAGCCTCCCAAGTAG | 386 (Hs) 387 (Pt) | Enhancer Region ALUY |
| RvPrimer4 | GACGATAGTCATGCCCCGCG | | |

Transcriptional activity of the alternative promoter under the effect of each of these five candidate enhancers from humans and chimpanzees were quantified in two replicate experiments by a luciferase assay in two neuroblastoma cell lines (except for Enhancer region 3, which was only in quantified in humans). Two independent experiments were carried out for each cell line. As a reference, we used the pGL3 with the human or chimpanzee alternative *THBS4* promoter (pMCL-022 and pMCL-025). We did not observe any significant increase in the transcriptional activity of the alternative promoters in the presence of any enhancer' candidate compared to the promoter basal activity in the

corresponding species (FIGURE 18). However, luciferase levels in the human plasmid containing the Enhancer region 15 are significantly lower than the basal activity ($P$=0.004) when assayed in the SH-SY5Y human cell lines. In addition, regarding the differences between species, Enhancer region 4 in the chimpanzee construct presents lower activity than in humans for the SH-SY5Y cell lines ($P$=0.022), but an opposite and not significant result for the SK–N-AS ($P$=0.220) was observed, suggesting that additional experiments should be considered for this specific region. In addition, we found that Enhancer region 15 and the region containing the ALU element (Enhancer region ALUY) showed consistently lower activity in humans compared with chimpanzees in both cell lines, although differences between species were not significant (Enh15: $P_{SHSY\text{-}5Y}$=0.186, $P_{SK\text{-}N\text{-}AS}$=0.114; EnhALUY: $P_{SHSY\text{-}5Y}$=0.201, $P_{SK\text{-}N\text{-}AS}$=0.523). This suggests that none of these regions act as enhancers over *THBS4* alternative isoform, although ideally Enhancer regions 4, 15 and ALUY should be studied in more detail, together with the regions that were not analyzed due to problems during the amplification or cloning.

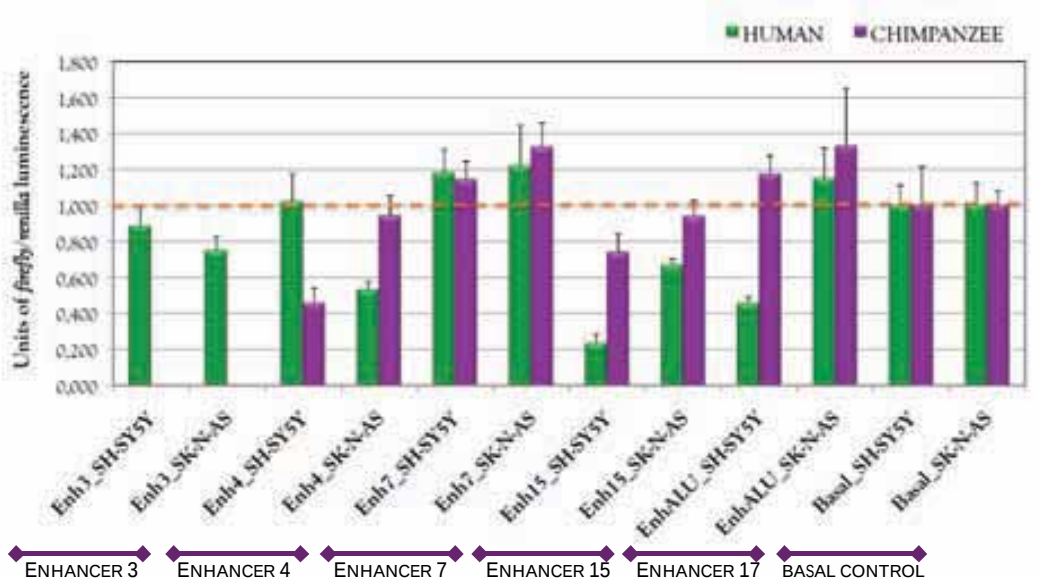



FIGURE 18. Quantification of transcriptional activity quantification of the alternative *THBS4* promoter in presence of different enhancer candidates in humans and in chimpanzees. Luciferase activity was measured from neuroblastoma cell lines transfected with plasmids containing around 3 kb of the alternative *THBS4* promoter of humans or chimpanzees and different regions selected as candidate enhancers. Graphs represent the average ratio between *firefly* luminescence and *renilla* luminescence in the *y* axis, with the standard error indicated by error bars. Orange dashed line represents the basal levels for each cell line of the plasmid including the human *THBS4* alternative promoter without the enhancer region.

## 3.5. *THBS4* expression variation in humans.

As an additional strategy to understand *THBS4* regulation, we checked the nucleotide variation in humans and its possible association to gene expression differences. In order to obtain information about the genetic variants present in *THBS4* in humans, we focused our attention on a region of 100 kb upstream and 100 kb downstream the *THBS4* gene, and used the data available from the International HapMap Project, which has developed a haplotype map of the human genome (INTERNATIONAL HAPMAP 2005). In the HapMap project, three independent populations have been analyzed: YRI (Yoruba), CEU (European origin) and CHB and JPT (Chinese and Japanese). Focusing on our region of interest, it was shown that the three populations presented a considerable linkage disequilibrium corresponding with the upstream region and the 5'-end of both isoforms of *THBS4* and extends through the whole coding sequence, including also the *metaxin-3* gene.


FIGURE 19. HapMap data for the ***THBS4*** gene +/- 100 kb. Representation of the linkage disequilibrium for the YRI (Yoruba, red), CEU (European, blue) and CHB and JPT (Chinese and Japanese, green) populations in a region of 100 kb upstream and 100 kb downstream the *THBS4* gene. According to the authors from the international HapMap consortium, phased genotypes from HapMap Phase II release 22 were used to calculate LD values for all SNP pairs with Haploview. This figure has been modified from the one available at the UCSC genome browser.

As previously mentioned, in the region of the *THBS4* alternative promoter there are several blocks of divergent haplotypes that correspond to the two human variants tested in the luciferase assays. In addition, variation within the human population for the *THBS4* gene was determined by the almost full resequencing of the gene in 24 African American (AA) and 23 European (EU) subjects within the Seattle SNPs Variation Discovery Resource

gene panel (http://pga.gs.washington.edu/) suggesting the presence of two main haplotypes within the gene in humans. These haplotypes showed high linkage disequilibrium and extended for around 16.5 kb at *THBS4* central region including exons 3 to 5 (chr5: 79.374.402–79.390.948). In collaboration with the group of Dr. Manuela Sironi at the Scientific Institut IRCCS E. Medea (Bosisio Parini, Lecco, Italy), an in-depth population genetics study of this region was carried out and it was suggested that both *THBS4* haplotypes were very old and could be subjected to balancing selection (CAGLIANI *et al.* 2013). However, the functional target of this possible balancing selection on *THBS4* was not clear.

None of the nucleotide changes between haplotypes affects the sequence of the protein. Therefore, one possibility was that the presence of these different haplotypes could be related to variation in the *THBS4* expression levels, giving rise to functional differences that could be related to the balancing selection. At *THBS4* exon 3, we could find one of the linked SNPs that define the two haplotypes (rs438042), which produces a synonymous change of an adenine for a thymine. Thanks to this SNP within the *THBS4* transcript, we could investigate whether there is a difference in the gene expression levels in the brain of humans with different haplotypes.

TABLE 22: ***THBS4* genotypes of the 18 frontal cortex samples used.** Results of the sequencing of the samples to genotype the SNP rs438042 at *THBS4* exon 3.

| Individual | Institution number | Genotype | Individual | Institution number | Genotype |
|---|---|---|---|---|---|
| Hs1 | 813 | A/T | Hs13 | OS99-08 | A/A |
| Hs2 | 1029 | A/T | Hs14 | OS03-394 | T/T |
| Hs6 | 1863 | A/T | Hs30 | A02-15 | A/T |
| Hs7 | 1903 | A/T | Hs31 | A02-98 | A/T |
| Hs8 | 1134 | A/A | Hs32 | A03-25 | A/A |
| Hs9 | 1570 | A/A | Hs33 | A03-34 | A/T |
| Hs10 | 1673 | T/T | Hs34 | A03-62 | A/T |
| Hs11 | 1832 | A/T | Hs35 | A04-162 | T/T |
| Hs12 | OS02-35 | A/T | Hs36 | A05-5 | A/T |

Before performing any experiment for the characterization of gene expression variation in *THBS4*, the exon 3 SNP region was amplified and sequenced from frontal cortex cDNA

of 18 independent subjects in order to determine their genotype. Within the 18 human samples, we found a total of five homozygotes A/A, three homozygotes T/T, and ten heterozygotes A/T (TABLE 22). Two independent real-time RT-PCR experiments were performed to measure the total *THBS4* expression levels in these samples using SYBR-green and the previously described primers for the 3' region of the gene. Additionally, as a control, a TaqMan® Gene Expression Assay (Hs00170261_m1 from Life Technologies) was performed to measure *THBS4* total expression, giving similar results. The expression levels in the sample A03-25 were consistently three-fold to six-fold higher than in other individuals and the sample was eliminated from further analysis. The comparison of the different genotypes with the *THBS4* expression levels in frontal suggested that T/T and A/A homozygotes are ~1.3-fold more expressed than A/T heterozygotes (FIGURE 20A). To explore this in more detail, an ANOVA was performed to examine the effect of other variables such as age or sex of the individual (CAGLIANI *et al.* 2013). A significant increase in *THBS4* expression levels with age was observed ($F$ = 10.474, $P$ = 0.007). However, there were no significant differences between genotypes.

To confirm the absence of expression differences between genotypes and to measure the allele-specific gene expression associated to each haplotype, we performed pyrosequencing experiments with these samples. Surprisingly, pyrosequencing results were in the opposite direction in relation with the total mRNA expression results, suggesting that in heterozygotes the A allele is ~1.5-fold ($P$<0.001) more expressed than the T allele (FIGURE 20B). In addition, a small amount of A allele expression was detected in T/T homozygotes, which indicated that there might be some bias in the experiment. Therefore, to clarify these results, a third approach was performed. Specific primers with its 3' end complementary to each SNP allele were designed to perform allele-specific real-time RT-PCR (AS-PCR). The results of this PCR appeared to be more similar to the first expression analysis than to the pyrosequencing results, supporting similar levels of *THBS4* expression among the different genotypes and similar expression levels of each allele within the A/T heterozygote samples ($P$=0.361) (FIGURE 20C). These results are described in a publication (CAGLIANI *et al.* 2013) in collaboration with Dr. Sironi group ( APPENDIX II).

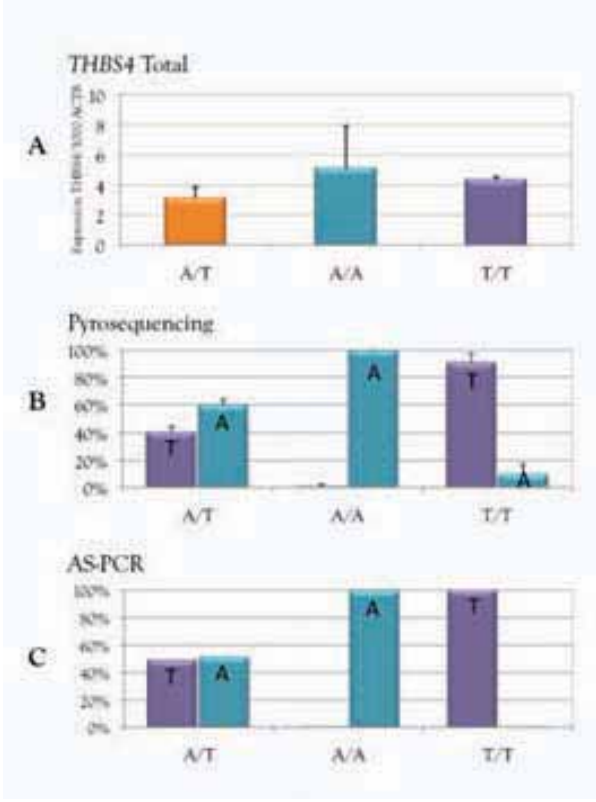



FIGURE 20. Results from the different approaches to investigate the variation of *THBS4* gene expression in the frontal cortices of 17 humans. A. Average of the quantification by real-time RT-PCR of the total *THBS4* mRNA expression levels for each genotype. Graphs denote the average number of *THBS4* molecules for $10^3$ molecules of ß-actin (*ACTB*) in the *y* axis. B. Average percentage of the relative amount of transcript of each allele at SNP rs438042 measured by pyrosequencing. C. Average percentage of the relative amount of transcript of each allele at SNP rs438042 measured by AS-PCR. In all the cases, error bars indicate standard variation between the different samples. Of the 17 samples, 4 are homozygotes A/A, 3 homozygotes T/T, and 10 heterozygotes A/T.

Finally, since the analysis of the putative *THBS4* transcripts indicated that the spliced EST DA759537 contained an additional alternative exon 3, we wanted to investigate whether this alternative splicing introducing an extra exon could be correlated with either of the two *THBS4* haplotypes. In particular, some of the nucleotide changes between the two haplotypes are very close to the 3' end of exon 3 and could affect the splicing regulatory sequences (CAGLIANI *et al.* 2013). A pyrosequencing based allele quantification was performed in the frontal cortices of 11 human individuals (Hs1 to Hs14). To apply this technique it was necessary to have short amplification products (<200), requiring an initial PCR of 472 bp from a specific primer within the extra exon to the exon 5 of *THBS4*, and a

second nested PCR of 79 bp centered in the SNP region (TABLE 23). As happened when this region was also tried to be amplified by real-time RT-PCR, the low levels or the lack of the extra exon in the selected individuals did not allowed to perform the final pyrosequencing experiments. However, the preliminary semi-quantitative PCRs did not suggest any obvious relationship between the different haplotypes and the presence of the alternative exon.

TABLE 23. Sequence and combination of primers for nested PCR of ***THBS4*** alternative exon 3 and pyrosequencing allele quantification.

| Primer name | Sequence (5'-> 3') | Amplicon (bp) | Gene/Region |
|---|---|---|---|
| THBS4FwAE3 | TTTTCTTCTAATCTTCTAGCCATCC | 472 | THBS4 AE3→E6 |
| THBS4-39 | TGTTTCCTTAACCTGCTGTCTC | | |
| THBS4-piro-Fw | CTCCCAGAAACCTGAGACCA | 79 | rs438042 amplification |
| THBS4-piro-Rv | CACCAGCTTCAGCTCTTCCA | | |
| THBS4-piro-S | TCTTCCAAGAAGTCCTG | – | rs438042 genotyping |

## 3.6. Searching for insights about *THBS4* promoters evolution.

The last (but not least), main goal of the thesis was focused on the search for information about which *THBS4* isoform was ancestral and how *THBS4* promoters could have evolved. Looking for any possible information that could provide a clue about which of the *THBS4* isoforms appeared first, we focused in the lack of the signal peptide of the alternative isoform. Considering that the thrombospondin family is known for being extracellular glycoproteins, the existence of a signal peptide might be important since it is required for protein secretion. The amino acid sequences of all thrombospondins in humans (THBS1-THBS4 and COMP) were checked for the detection of their signal peptides with the software SignalP 4.1 (http://www.cbs.dtu.dk/services/SignalP/). We

found that only the alternative isoform of THBS4 is the only one lacking this heterogeneous amino acid sequence.

Considering that the mouse *MTX3* gene is located at the same distance from *THBS4* as in humans, which suggests that the genomic region is well conserved, we centered our study on 10 kb upstream of the *MTX3* gene. This region should include, if conserved, the first and the second exon of the alternative isoform of *THBS4*, and thus was aligned between different mammals species. Sequences obtained from the UCSC genome browser were aligned with the multiple sequence alignment ClustalW2 program, freely available at http://www.ebi.ac.uk. Different alignments (only primate species, primate species plus mouse and rat and the previous plus other mammals like cow and horse) were performed to facilitate the visualization of the region. However, the length of the sequence was too long and the divergence too high to obtain an alignment reliable enough to identify the homologous sequences of the promoter and both alternative *THBS4* UTR exons in other species and provide a clear insight of the conservation of these regions. We then decided to take advantage of online browsers like UCSC browser and ECR browser, which include already pre-computed alignments of multiple sequences, and focused on a region ± 500 bp from each alternative *TBHS4* exon (FIGURE 21). UCSC conservation track shows multiple alignments of up to 44 vertebrate species, with conserved sequences represented in green bars, while pale yellow coloring reflects uncertainty in the relationship between the DNA of the species due to lack of sequence in relevant portions of the aligned species. The ECR browser allows colorful patterning to represent the conservation in different genomic regions: blue areas correspond to positions of coding exons, yellow areas to UTRs, red areas to intergenic sequences, salmon areas to introns, and green areas to transposable elements and simple repeats. The conservation alignment provided by the UCSC browser showed how alternative exon 1 is well conserved among the great apes and rhesus monkeys decreasing the identity between sequences according to the evolutionary distance between species, although in any case conservation seems higher than in the flanking regions. ECR browser showed similar low conservation of this exon, although in this case there was not conservation for the alternative exon 1 in rhesus macaque. We considered this to be an error of the browser's sequence, since the alternative exon 1 has been aligned with other

browsers and experimentally amplified by PCR. Interestingly, the alternative exon 2 is located within a long terminal repetitive element from the family of the mammalian apparent LTR-retrotransposons (MaLR) and according to the browser alignments; it is less conserved still, being found only in a few primate species.

Finally, to provide further supporting information to the conservation alignments, we searched for EST sequences and CAGE tags present in the region upstream from *MTX3* of the mouse genome, which is the only mammal characterized at the functional level to a degree similar to humans. Despite finding several ESTs and CAGE tags consistent with the *THBS4* reference isoform, no evidence of a possible alternative promoter was found. Due to the low conservation, the lack of ESTs or CAGE tags in mouse, and the loss of the signal peptide in the alternative isoform, the reference isoform is likely evolutionarily closer to the rest of the thrombospondin family members. This suggests that the alternative isoform of *THBS4* has appeared after the reference one, probably some time after the divergence of the rodent and primate lineage or between simians and pro-simians, even though it was not possible to determine exactly how or when this change appeared.

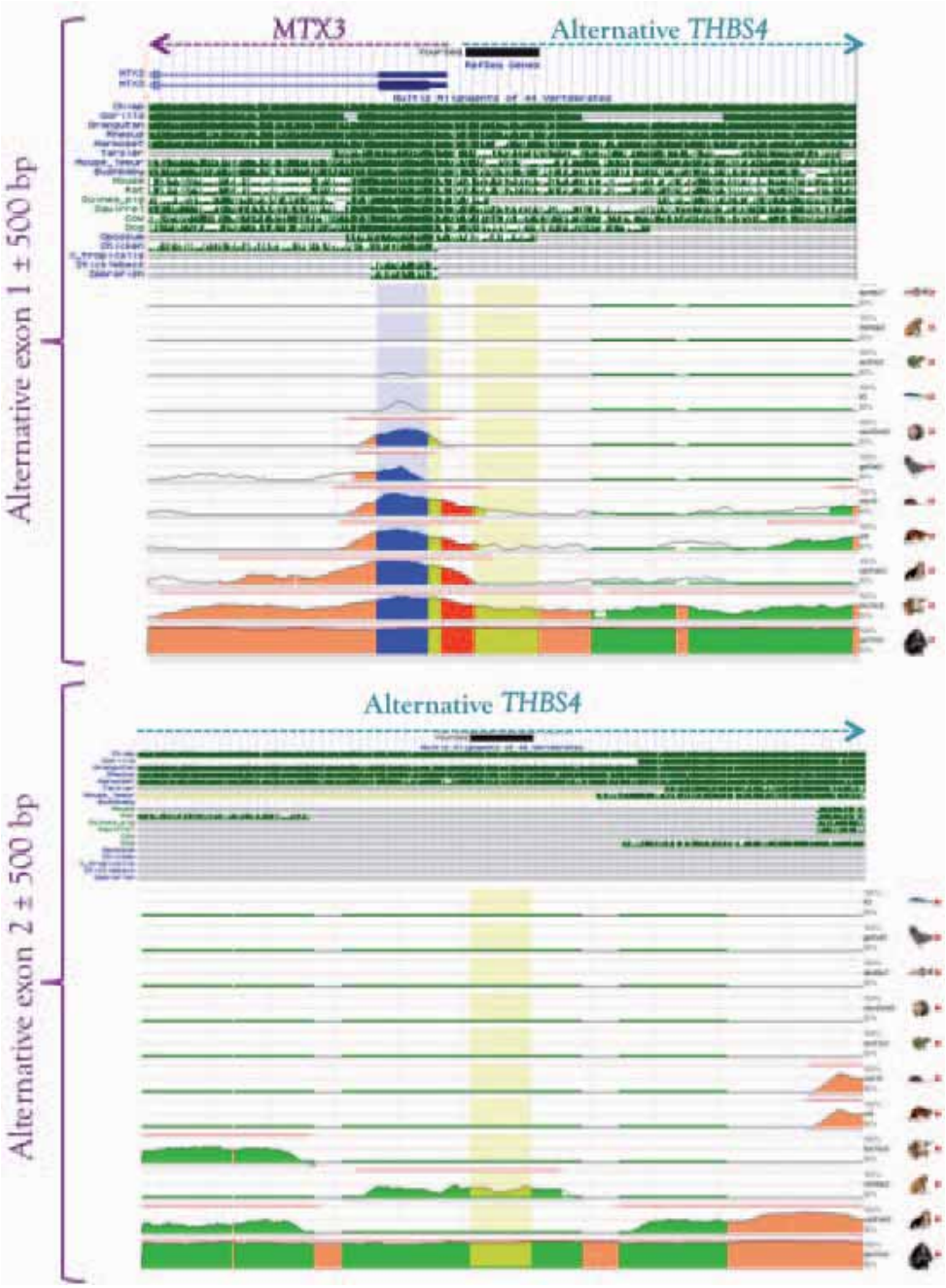



FIGURE 21. Screen shot from UCSC conservation track and ECR browser alignments for ***THBS4*** alternative exons ± 500 bp. The displays are centered in both alternative exons. UCSC track is represented with the aligned species in the left. Black squares represent the alternative exons, green coloring represents conservation between species. Bellow them, the ECR browser alignment is shown with the species represented in the right of the image. Different coloring pattern is used for each genomic region: coding exons in blue, UTRs in yellow, intergenic regions in red, intronic regions in salmon and transposable elements and simple repeats in green. *THBS4* alternative exon 1 and alternative exon 2 are colored in yellow since they are part of the UTR.

# 4

# DISCUSSION

*"To illustrate the potential pitfalls involved in making inferences about the human brain from studies of macaques or other putative model species, [...]. To be sure, we would get some things right. Humans, like macaques, have eyes set together in front of the face rather than on the sides of the head, have dexterous hands with opposable thumbs and digits tipped with nails instead of claws, [...] Unfortunately, we would also make many mistakes. For example, we would conclude that humans walk on all fours, have a tail, and possess a thick coat of fur."*

– TODD M. PREUSS, *The Cognitive Neuroscience* (1995) –

# 4

# DISCUSSION

## 4.1 The study of human brain characteristics.

Part of what it means to be human is how we developed to become human, both individually and as a species. Over the past six million years, as early humans adapted to a changing world, they evolved some characteristics that helped define our species today. There were important milestones along the way, with our physical form and life history of life providing more than enough examples of our peculiarities. Nevertheless, among the traits that distinguish humans from other primates and mammals, there is now a consensus that our brain and its unusual talent for complex thought is the most significant evolutionary characteristic. However, how did we arrive to these assumptions? How do scientists study human evolution?

In all traditional societies, myths have provided different scenarios about the origin of humans and their possible relationship with other animals through earth history. Religion and science have obsessively taken up the challenge and have come up with their own propositions (BOESCH 2007). It was in 1859 when Charles Darwin proposed that all animals are descended from one -or a few- progenitor species and suggested that evolutionary forces (mainly natural selection) shaped the different species, being the ones subjected to more selection the more prevalent at the end. In fact, Darwin was able to provide the necessary evidence to convince scientists of the truth of that insight (DARWIN 1859). About one century after Darwin's theory in the early 1960s, and before the establishment of the currently favored model-animal species (such as rats, mice, or rhesus monkeys), experimental biologists sought to identify features of biological organization shared by a wide variety of species and carried out studies based in field observations of the natural behavior of many animals (HARLOW and HARLOW 1962, MENZEL *et al.* 1963,

SCHRIER *et al.* 1965). The field observations allowed scientists to compare our sense of humanity to that of other primates and to suggest the traits that make us humans. Over the years, scientists have discovered that primates act altruistically, express empathy, find ways to avoid and resolve conflicts mainly through food sharing, and even seem to mourn the loss of deceased friends (FLACK and DE WAAL 2000, PALAGI *et al.* 2004). It has been also shown that primates and humans have the ability to alter the environment for adaptive purposes. Researchers hypothesize that tool use was a game-changer for early humans, since it allowed them to hunt and process different foods. Some researchers even speculate that accessing these new foods led to an increase in brain size and altered group dynamics, which may have changed how early humans communicated with one another (AIELLO and WHEELER 1995).

Current understanding of human brain evolution is based on morphology obtained through cranial fossil remains and by endocranial casts. Some traits, such as endocranial volume, morphology of the frontal lobes or the asymmetry of Broca's regions (HOLLOWAY 1983, FALK 2012), have been considered to be related to the evolution of modern human behavior and can be traced in fossils. However, these traits cannot be plainly evaluated in hominin fossil species due to the incomplete correspondence between endocranial and brain morphology. The study of those traits that do not leave an imprint in the endocranial surface can only be examined by means of comparative studies between humans and other primates (GOMEZ-ROBLES *et al.* 2013). Thus, to understand the evolution of humans, all levels of human brain specialization have to be considered: overall morphology and neuroanatomy, functional connectivity studies by metabolic and fiber imaging, identification of genomic changes, and comparison of gene and protein expression levels. Each of these strategies provide different viewpoints on human specializations and have thus far generated useful information. However, all these tools have important limitations. In particular, it has been shown that in many mammalian species (including primates), 9 out of 11 major brain regions (cerebellum, mesencephalon, diencephalon, olfactory cortex, parahippocampal cortex, hippocampus, neocortex, septum, and striatum) exhibit a robust covariance in size (FINLAY *et al.* 2001). Among these, the frontal lobe stands out with a more than three times larger absolute size in humans compared to the great apes, but it is

not disproportionately enlarged in humans when scaled with total brain size (SEMENDEFERI *et al.* 2002). There are many studies that have examined the relative size of gray and white matter in the frontal lobe or prefrontal cortex; they concluded that when the white matter is considered separately from the gray matter, the human frontal lobe also remains undistinguished from apes in terms of overall relative volume (SCHENKER *et al.* 2005). In spite of this, it has been shown that human brains have a greater distribution of white matter in gyral regions (SCHENKER *et al.* 2005), and as a whole, the human prefrontal cortex is 25% more gyrified than it is in apes (ARMSTRONG *et al.* 1995).

It could be expected that the enlargement of the human brain size was accompanied by the appearance of new cortical areas, and that the most encephalized primates have more cortical areas than other primates, but, as sensible as this seems, there is no good evidence that humans do possess this excess amount of specialized areas (PREUSS 2011). Moreover, given the role of the prefrontal cortex in higher-level cognitive functions, and the effect that its inaccurate development and function can have in several neurological and psychiatric disorders (such as autism), it has been suggested that a phylogenetically reorganization of frontal cortical circuitry could have taken place (TEFFER and SEMENDEFERI 2012). This reorganization could have involved an increase in size of some regions, the decrease of others, and an increased neuronal spacing distance, which may have been critically important for the appearance of human-specific executive and social-emotional functions (TEFFER and SEMENDEFERI 2012).

With the introduction of animal models in research, it has been possible to achieve further progress in the field of human evolution and the study of diseases affecting humans. However, cortical neuroscience studies presented additional challenges. On the one hand, researchers previously lacked the technical resources to explore the human brain in anything like the detail with which they were able to study non-human species. On the other hand, the use of model animals to study the brain could be a source of variability and significant errors. Macaque monkeys, for example, lack some of the features found in the human visual cortex, although they are widely used as models of the human visual system (PREUSS and COLEMAN 2002). Finding all the distinctively human traits about the human

brain will not be easily guaranteed without the ability to compare humans to chimpanzees and other great apes more directly (PREUSS 2011). One of the strangest things about humans, for example, is that for most activities that require the use of a hand, a large majority of the members of our species will instinctively use the right hand, which is mainly under the control of the left hemisphere of our brain (PREUSS 2004). This behavioral lateralization and also brain structural asymmetry were once considered to be exclusively human traits (KANDEL and SQUIRE 2000), owing to the pronounced cerebral hemisphere specialization for some complex cognitive capabilities (CORBALLIS 2009). However, thanks to the enormous improvement of non-invasive imaging techniques over the last decade, there is rising evidence that demonstrate that macrostructural, microstructural and behavioral asymmetries are widespread in other primates too (HOPKINS and PILCHER 2001, KELLER *et al.* 2009). Even that, it is thought that the asymmetry is still more common and/or more pronounced in humans (GOMEZ-ROBLES *et al.* 2013). Additionally, it has been recently suggested that an indication of fluctuating asymmetry may be a hallmark of developmental plasticity in the context of brain evolution and development, being a highly adaptive property that can be recognized at very different levels of brain organization (anatomical, cellular and molecular) (GOMEZ-ROBLES *et al.* 2013). It was shown that different forms of plasticity have a key role in learning and, therefore, in behavior and cognition (FELDMAN 2009). This suggested that exaggerated fluctuating cerebral asymmetries in humans relative to chimpanzees may be indicative of a substantial and sustained developmental plasticity, potentially linked with the unique cognitive abilities which have arisen during human evolution (GOMEZ-ROBLES *et al.* 2013).

Although imaging techniques clearly provide clinically useful data compared to conventional tract-tracing techniques, which involve injections of chemical tracers into the brains of experimental animals, they have certain limitations. Under most circumstances they cannot track fibers into the gray matter and specialized tracers are required to distinguish between anterograde and retrograde connections (PREUSS 2011). Moreover, frequent tracer spill into adjacent regions projecting elsewhere is an inherent curse of the technique. Imaging techniques require the animal to be treated with anesthesia, which excludes the analysis of cognitive functions during the experiment, and as they use scanners

designed for humans, these could be suboptimal for the study of non-human primates due to size differences (DUONG 2010).

Additionally to the imaging studies, the power of histological structural analysis of brain tissue acquired postmortem also reveals human specializations of the cellular and histological organization of the cortex by using the proliferation of antibodies or other ligands to probe the distribution of molecules in the brain. These studies have documented important human features, such as modifications of neuronal and glial phenotypes, which may supply a different cellular substrate for many of the diverse neurological capabilities (OBERHEIM *et al.* 2009), the horizontal spacing of neurons, which may be related to behavioral differences in associative functions including the human capacity for language and decision-making (SEMENDEFERI *et al.* 2011), or the modifications in the afferent innervation organization, which could support the evolution of the human cognition (RAGHANTI *et al.* 2008). Nevertheless, the importance of all these techniques and the value of the postmortem tissue required for them, has grown thanks to the introduction of antifreeze storage solutions preserving tissue for years in a state suitable for immunohistochemistry and *in situ* hybridization, avoiding the degrading effects of overfixation (HOFFMAN and LE 2004).

Another level of information about human brain evolution comes from genomics and comparative molecular biology. The differences in brain structure and function that distinguish humans from other non-human primates and other animals presumably have correlates at genetic level, involving some combination of changes relevant for the gene products. Since the mid-1970s with the publication of KING and WILSON (1975), it was believed that the amino acid sequences of proteins –and the nucleotide sequences of the genes that code for them– are very similar in humans and chimpanzees. In particular, with the completion of the genome sequence of humans, chimpanzees, and other primates (LANDER *et al.* 2001, VENTER *et al.* 2001, CONSORTIUM 2004, CONSORTIUM 2005, PRUFER *et al.* 2012, SCALLY *et al.* 2012), it has been possible to identify a plethora of changes in the human genome that could be related to the evolution of our brain. As mentioned in the introduction, these changes could have different origin and repercussion,

including, among other mechanisms, changes in the expression of genes, structural variants or small sequence changes. For example, after the first comparison of the human and the chimpanzee genome sequences (CONSORTIUM 2005), it was established that the average nucleotide divergence between any two copies of the human and chimpanzee genome is 1.23%, or that some transposable element insertions have been three times more active in humans. Given the enormous similarity between the sequences of these two species, it has been attempted to identify the relationship between gene differences on the one hand and anatomy, physiology, behavior, and overall brain evolution on the other. As the few widely acknowledged specializations of the human brain and cognition are related to encephalization and language, the majority of the studies are focused on the search of genes related to language or brain size, missing an extensive area of genomics research with probably most of the genetic differences between humans and other primates (PREUSS 2011).

In the last decade, high-throughput gene expression analyses have made possible the study of many primate species and tissues, including the brain. By using high-density DNA microarrays, it has been possible to identify multiple genes with expression differences between humans and other primates, getting closer to the actual molecular changes involved in the phenotypic differences between them. Unfortunately, many of these studies were ambiguous to determine what part of the observed expression differences were due to genetic changes and what could be due to environmental factors.

Thanks to the high-throughput sequencing techniques, it has been also possible to identify many different changes and genes that underwent the action of natural selection and that could be related to brain function. For example, these approaches have led to the identification of changes in genes associated with language and speech (*FOXP2* or *PCDH11X/Y*), genes associated with brain size (*MCPH1* or *ASPM*), and genes associated with neuronal functionality (*DRD5*). However, for most of these changes it is very unlikely to know for certain in which tissue type are they acting and what is the phenotypic trait being selected. Moreover, in many occasions, it is not clear if a specific change could be

functionally important, or it would merely be the result of a neutral variation with no consequence for the function of the phenotype.

These different approaches to human evolution allow several levels of comparison between human and non-human primates, but in order to really understand how we became humans, it is necessary to find the genomic variation capable of triggering the evolutionary changes in our brain. Performing this kind of studies is quite difficult, especially when it includes different species, requiring the analysis of multiple possible factors at the risk of not finding at the end the needle in the haystack.

## 4.2 The analysis of gene expression changes in the human brain.

As mentioned above, the observation of interspecific differences between humans and non-human primates in gene-expression levels in the brain or other tissues is inherently difficult to interpret. An individual itself can be influenced by several internal factors, such as hormones or the environment in which the individual is located and develops. When the comparison is additionally performed across species, physiological, morphological and environmental differences are also expected to contribute to variation in gene expression levels (ROMERO *et al.* 2012). It is important to identify and clarify the specific genetic reasons of expression differences and distinguish them from environmental effects. Furthermore, there is a large number of potential sources of variation to be controlled, because they could interfere with the experimental results producing false positives. These can be technical causes, such as variation in sample quality, biological sources, like variation due to sex, age, circadian rhythm, cause or death or post-mortem delay, or incorrect assumptions from the gene expression results, such as increases in mRNA that might not be accompanied by increased protein levels. Finally, gene expression differences could be due to differences in proportion of specific cell types. Since each cell type contributes to a

unique gene expression signature relating to its function, in which only a fraction of its genes are expressed while the rest remains silent, the relative proportions of the cell types affect the overall gene expression profile, especially in heterogeneous tissues like the brain. For example, it is possible that genes that are expressed in brain cells may be silenced in liver cells or heart cells. Accordingly, when comparing between different tissues or the same tissues but in different species, the amount of a specific cell type contained in the specific tissue could implicate having more or less amount of a specific RNA, and that could correlate with differences in gene expression.

When comparing gene expression levels between different species and individuals, it is important to consider that the observed inter- or intra-specific variation of the gene expression in a specific experiment could have been influenced by the effect of natural selection. It could be assumed that high variation in gene-expression levels in different individuals and species would be produced as a general consequence of the difficulty to acquire tissue samples in a comparative study. However, if gene regulation evolves under stabilizing selection, genes are expected to have low variation in expression levels within and between species, being robust to environmental differences. By contrast, if the gene expression changes are seen in a particular lineage, like the higher expression level observed exclusively in humans with little variation among individuals within a species, they are probably under the pressure of directional selection in that particular lineage (GILAD *et al.* 2006). In this case, there are also alternative explanations difficult to exclude, such as lineage-specific relaxation of evolutionary constraints or lineage-specific differences in environment. The identification of the specific genetic mechanism and the regulatory elements underlying the expression changes may help in the detection of different signatures of natural selection. This could allow us to formulate specific and elaborated hypotheses for further experiments; for example, using model organisms for functional experiments inspired by a comparative analysis of gene regulation in primates but evading the limitations of experimenting with them (ENARD *et al.* 2009, MCLEAN *et al.* 2011).

As mentioned before, DNA microarrays studies have identified few hundreds of genes with differences in the expression between humans and other non-human primates. It has

been suggested that this kind of experiment presents methodological difficulties that make necessary the application of additional experimentation to validate the results (PREUSS *et al.* 2004). In addition to the replication of the microarray results with samples from additional individuals, approaches at different molecular levels can be used to validate the resulting candidates (FIGURE 22). For example, at an mRNA level, there are techniques such as cDNA microarrays, which are less sensitive to sequence differences than the high-density oligonucleotide microarrays, and other validation techniques involve examination of expression changes in single genes by northern blotting or quantitative real-time PCR. Furthermore, as mRNA and protein levels are not always correlated (PREUSS *et al.* 2004), other techniques to determine the significance of mRNA-level differences could be focused on changes in protein expression, which can be examined by western blotting. Finally, considering the cellular localization, it is possible to use techniques such as *in situ* hybridization (a labeled complementary DNA or a RNA strand is used to localize a specific DNA or RNA sequence in a section of tissue), or immunohistochemistry (proteins localized in a particular tissues, can be detected by their binding to specific antibodies) to validate the gene expression differences. These different validation techniques could provide a great help to formulate new hypotheses and design different strategies to understand how the different genes are regulated and how they have evolved between species.

But, what is required to go from this amount of genes differentially expressed to the real functional differences? SOMEL *et al.* (2013) proposed two different strategies: the first one is based on the limitation of the investigation to known functional sites and to focus on changes found in the human genome but not in the genomes of any other closely related primate species, including the extinct hominin species. The second proposed strategy is focused on the search of regions within the human genome that contain unusually large numbers of human-specific DNA sequence changes and may thus have appeared as a consequence of an adaptive evolution at such loci.

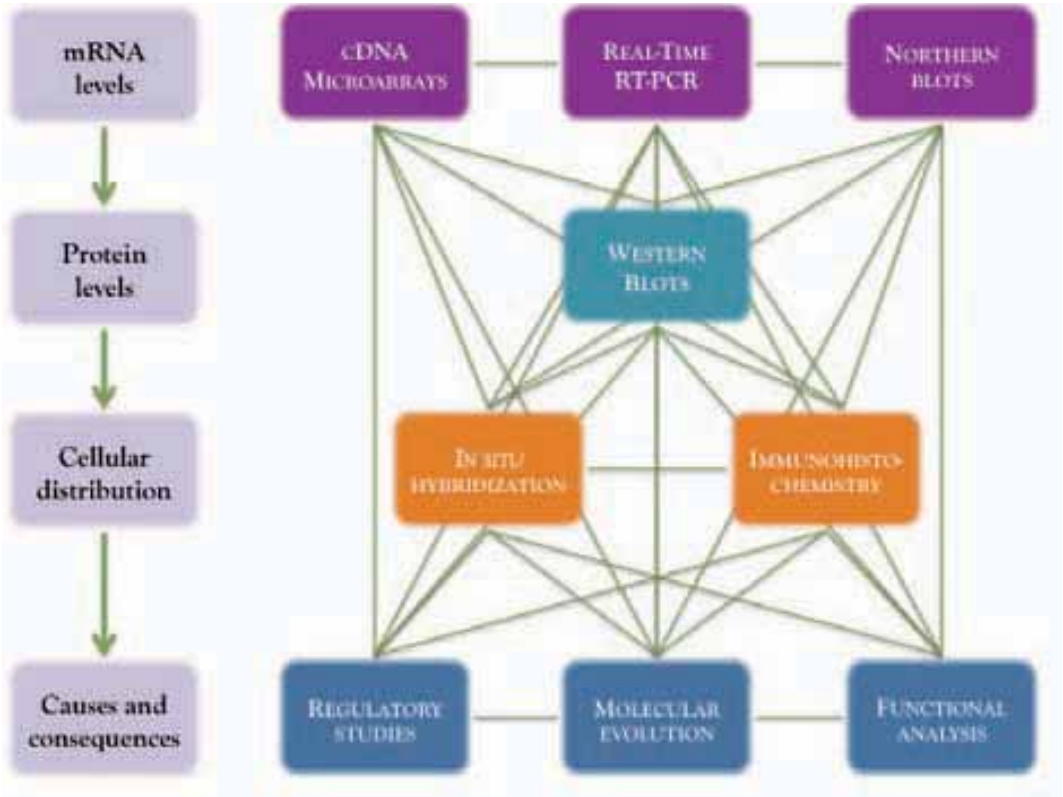



FIGURE 22. Follow-up of gene expression differences. Representation of the diverse mechanisms that can be used to validate the difference of gene expression levels, shown by high-density oligonucleotide microarrays in different comparative studies. At mRNA level, techniques such as cDNA microarrays, northern blotting or quantitative real-time PCR are less sensitive to sequence differences than oligonucleotide microarrays. At protein level, western blots can be performed to ensure that mRNA and protein levels are correlated. Techniques such as *in situ* hybridization or immunohistochemistry can be use to localize within a tissue section the amounts of a specific DNA/RNA or protein. These different techniques of validations could help in the formulation of new hypothesis and design different strategies to understand the causes and consequences of the expression differences.

The strategy that we propose could be considered somehow a fusion technique of the two proposed by SOMEL and colleagues. Once the expression differences are validated by real-time PCR and protein analysis, the main goal is the determination of the causes regulating such expression change. The first step could be the search for possible structural variants that could have appeared during the evolution of the specific gene under investigation. Thanks to different browsers available online, such as UCSC genome browser or the ECR browser, and the amount of published information on structural variants, the genomic region around the gene can be compared between humans and other primate

species, looking for changes in the human lineage. We consider that the next step should be focused on the search of different regulatory elements affecting the expression. Thanks to previous studies using yeast (TIROSH *et al.* 2009) or flies (WITTKOPP *et al.* 2008), it has been shown that changes in *cis* elements appear to be more commonly responsible for inter-species differences in gene expression patterns than changes in *trans*. This finding is rationalized by the likely lower pleiotropy of *cis*-regulatory sequences relative to *trans*-regulatory factors. In particular, the modular structure of *cis*-regulatory sequences allows mutations that affect one module (for example an enhancer controlling expression at a particular time and/or in a particular tissue type) to have little or no impact on expression controlled by other modules. Meanwhile, changes in *trans*-regulatory factors tend to affect many different genes and phenotypes, which may be beneficial in some cases, but the probability that they are beneficial for all of them is extremely low.

According to this, we focus our strategy on the regulation of *cis*-elements, proposing a complete analysis of the gene promoter region based in three independent characterizations. First, the quantification of its transcriptional activity in humans and chimpanzees by reporter assays in several cell lines, which could identify the promoter itself as the cause of the expression differences. Second, the measurement of epigenetic regulation marks, such as different DNA methylation levels around the promoter region between species by bisulfite-sequencing. And third, the computational and experimental search of enhancers that could be regulating the transcriptional activity of the promoter and their posterior validation by reporter assays. However, it is possible that this *cis*-regulatory search does not provide any evidence to directly link them as responsible of the real functional differences. For this reason additional experimentation looking into *trans*-regulatory elements, different post-translational modifications, correlation between gene and protein expression and the application of more complicated and expensive functional analyses, could be also required in order to find the connection between change and fuction. A deep search for the genetic causes of each of these gene expression differences between species could validate unequivocally the differences and enlighten their putative implication in human brain evolution, contributing to the clarification of important characteristics about what makes us humans.

## 4.3 The *THBS4* gene.

In this work, we have focused on the identification of the molecular mechanisms responsible for the up-regulation of the *THBS4* gene in the human brain. This gene was initially selected because it showed the highest expression differences in the adult cortex of humans compared to chimpanzees and macaques in preliminary microarray data (CÁCERES *et al.* 2003), which was subsequently confirmed by gene and protein expression characterization (CÁCERES *et al.* 2007). Elucidating the origin of such high expression differences was considered of paramount importance. Taking into account the role of *THBS4* in the induction of synapse activity via a GABA receptor $\alpha 2\delta$-1 (EROGLU *et al.* 2009), and the recent discovery of its implication in astrogenesis from the subventricular zone after a cortical injury (BENNER *et al.* 2013), the expression differences could be an indication of important evolutionary changes in the synaptic organization and dynamics of the human brain. Moreover, the THBS4 protein is expressed by vascular endothelial and smooth muscle cells (STENINA *et al.* 2003), where it is believed to stimulate leukocytes and induce an inflammatory response by binding to the Mac-1 receptor (PLUSKOTA *et al.* 2005, STENINA *et al.* 2007). Inflammation processes have been closely related to the etiology of many neurological diseases (FUSTER-MATANZO *et al.* 2013). Finally, the THBS4 protein has been located in Aβ-amyloid plaques (CÁCERES *et al.* 2007), which are one of the two histopathological hallmarks in Alzheimer disease (the other principal feature are neurofibrillar Tau tangles), suggesting that THBS4 may be involved in the pathogenesis, or at least in the increase of human vulnerability to this neurodegenerative disorder that appears to be an uniquely human condition. All these characteristics point out the important role of *THBS4* in the human brain and suggest it to be the perfect candidate to apply our proposed strategy to study its possible implication in human brain evolution and in its phenotypic specializations.

### 4.3.1 Causes and consequences of *THBS4* over-expression.

The majority of the comparative studies on gene expression in primates have been focused on understanding the pattern of evolutionary changes and the adaptive phenotypes, questioning what kind of phenotypic differences underlie certain alterations in gene expression levels. Some of these studies based their experimentation on previous genome-wide gene expression studies, using different strategies compared to those suggested above, in order to search for the functional implications of the gene expression differences between species and the most probable mechanisms that are regulating them (PREUSS *et al.* 2004, ROMERO *et al.* 2012). Similar to theses approaches, this work sought to address the genetic causes responsible for the up-regulation of the *THBS4* expression in the human brain.

It is reasonable to understand that, due to the difficulty to obtain samples for the molecular characterization of a gene in primates, where tissue material is often scarce, one way to find what causes gene expression differences is to rely on the genome sequences available. Computational comparison of the human genome sequence in relation with other closely related primate species whose sequence is also available, will detect –if present– the possible structural changes affecting the gene expression between the species (SEGAT and CROVELLA 2011, BEKPEN *et al.* 2012). An exhaustive study of the genome context in which *THBS4* is located in humans and other primate species, allowed in a first instance to discard possible structural changes as a source of the gene expression differences. It was shown that *THBS4* is located inside a chimpanzee-specific inversion (SZAMALEK *et al.* 2005). However, since *THBS4* is around 16 Mb away from the closest inversion breakpoint, it was considered likely that all the possible regulatory elements for gene expression might have the same organization in humans and in chimpanzees. Evidence of different expressed sequence tags and in the accumulation of clusters tagging the region by gene expression cap analysis (CAGE) (KAWAJI *et al.* 2006) permitted the precise identification of a putative alternative promoter isoform located upstream of the reference

transcription start site (TSS) and only a few base pairs away from the next gene *MTX3*. Several amplification experiments supported the hypothesis that the expression of *THBS4* gene involves two different mRNAs, emphasizing the importance of a detailed *THBS4* regulation study.

#### 4.3.1.1 Possible effects acting in *cis*.

Taking into consideration that variations in *cis* regulatory sequences are thought to be the most prevalent cause of phenotypic divergence (CARROLL 2008, STERN and ORGOGOZO 2008, WITTKOPP *et al.* 2008, TIROSH *et al.* 2009) and that the functionality of most *cis* regulatory elements depends on the context relative to the promoter (CAI *et al.* 2001, BEER and TAVAZOIE 2004), the identification of an alternative mRNA for *THBS4* and consequently an extra promoter, open up a long variety of possibilities to explain the up-regulation of the gene in the human brain.

Once we detected the existence of two promoter isoforms for *THBS4,* the genome sequences of humans and chimpanzees (using macaques as an out-group to determine the lineage specificity) were compared in a region of 2 kb upstream and 2 kb downstream the TSSs of both mRNA isoforms. Considering together the single nucleotide changes and the small insertion/deletion events specific of each lineage, around 20 to 30 sequence changes between both species in each 2 kb section were detected. This variation of about 1% of the selected region is encompassed within the expected 1-4% changes described between these two species (VARKI and ALTHEIDE 2005, POLAVARAPU *et al.* 2011).

Real-time RT-PCR was performed to quantify the expression of the specific 5'- end of both *THBS4* mRNAs in several human tissues and cell lines. Our results indicated that each promoter drives transcription in different tissues, allowing a major potential to regulate the gene. The reference *THBS4* isoform presented high expression levels in heart and colon, whereas the levels in the other tissues were, if present at all, lower than those of the alternative isoform. This result may suggest that the reference isoform, but not the

alternative one, is implicated in the described processes of extracellular matrix modifications in cardiac muscle, resulting in fibrosis and modification of myocardial structure and function after ischemic infarction (MUSTONEN *et al.* 2008, FROLOVA *et al.* 2012, MUSTONEN *et al.* 2012). Moreover, the suggested action of *THBS4* as a putative tumor suppressor gene in colorectal cancer (GRECO *et al.* 2010) could also be specifically related to the reference mRNA. In contrast to that, the alternative mRNA isoform was mainly expressed in brain tissues and in sexual organs (both female and male). The fact that *THBS4* expression in the human brain mainly comes from the alternative isoform could be related with the differential up-regulation in humans. To test this hypothesis, a quantification of both isoforms was performed by real-time RT-PCR in the frontal cortex of humans, chimpanzees and macaques. Interestingly, similar to the results observed for total *THBS4* expression (CÁCERES *et al.* 2007), we found around five to six fold higher expression of both the alternative and the reference isoform in humans compared to chimpanzees. However, the alternative isoform of *THBS4* is significantly higher expressed than the reference isoform in both humans and chimpanzees, which suggests that the alternative isoform has a bigger role in the possible functions carried out by *THBS4* in the brain than the reference one.

Finally, the transcriptional activity of both promoter regions of humans and chimpanzees was measured using luciferase reporter assays. The results in all the transfected cell lines used showed no significant transcriptional activity differences in humans and in chimpanzees for both promoter regions, with the exception of the experiments with the HEK293T cell line. In these cells, all the plasmids including sequences upstream the reference isoform with or without a universal enhancer showed a significant increase of the transcriptional activity in chimpanzees compared to humans. This suggests that the reference promoter in chimpanzees might be more active than in humans and is consistent with the *THBS4* expression levels in heart by microarrays and real-time RT-PCR (Cáceres 2007). On the other hand, for both species and all cell lines, significantly higher transcription from the alternative promoter was appreciated compared with the activity of the reference region. These results, in concordance with the higher levels of alternative mRNA expression, suggest the regulation of the alternative promoter is mainly responsible

for *THBS4* upregulation in the human brain. However, according to the reporter assays, *THBS4* expression differences between humans and chimpanzees are not due to sequence differences in the promoter. This makes the characterization of the alternative promoter region essential to further explore other possible regulatory changes, such as epigenetic changes, active enhancers or possible *trans*-acting factors.

#### 4.3.1.2 Potential epigenetic effects.

Several factors could be involved in the regulation of the alternative promoter region and not be specifically related to the sequence itself. One of these mechanisms is the alteration of the DNA methylation at CpG sites near the promoter. The methylation of DNA, in which a methyl group at the C5 position of the cytosine ring is added to generate a 5-methyl-cytosine (BERNSTEIN *et al.* 2007), usually occurs on CpG dinucleotides (cytosine residues that precede a guanosine) (IBRAHIM *et al.* 2006, ZILLER *et al.* 2011). DNA methylation at CpG dinucleotides has long been considered a key mechanism of transcriptional regulation –higher methylation levels are inversely correlated with gene expression levels (JONES and TAKAI 2001)– and a critical factor in embryonic development and in carcinogenesis (REIK 2007, GEIMAN and MUEGGE 2010).

The presence of CpGs in the genome as a whole is relatively atypical; they are often clustered in short DNA regions of 300 to 3000 base pairs called CpG islands (BERNSTEIN *et al.* 2007). The functional consequence of DNA methylation is, therefore, highly dependent on the position of the CpG island within the genome. It has been shown that about 70% of genes have a CpG island within their proximal promoter region (ESTECIO and ISSA 2011). The alternative promoter region of *THBS4* could be considered within this percentage, since it presents a 550 bp CpG island immediately upstream of its TSS. However, these CpG islands near promoter regions are typically unmethylated; it has been suggested that only about 3% of the total CpGs in this position are methylated in normal cells (ESTECIO and ISSA 2011).

The transcriptional activity of genes with a promoter CpG island is suppressed when it is methylated (BIRD 2002). Even though an absence of methylation does not automatically correlate with high expression level, it can be considered a lax state for transcriptional activity (HERMAN and BAYLIN 2003, BERNSTEIN *et al.* 2007). There are also CpG islands present outside of the gene sequences, for example in conserved intergenic regions, although their function has not been well investigated yet. Because their methylation pattern looks like the CpG island within promoters, it has been suspected that they may be part of distal regulatory regions and enhancers (MAUNAKEA *et al.* 2010).

The first genome-wide scan of DNA methylation in human and chimpanzee brain was recently conducted using bisulfite sequencing in three prefrontal cortex samples of each species. Despite high intra-species variation, ZENG *et al.* (2012) found higher CpG methylation levels in chimpanzee compared to the human prefrontal cortex (ZENG *et al.* 2012). Since DNA methylation has been reported to have a crucial role in promoter activity in the brain (HOUSTON *et al.* 2013), and assuming the low probability of finding methylation on the CpG island near the alternative *THBS4* promoter, we proceeded to measure its levels in human and chimpanzee frontal cortex samples. We hypothesized that a possible higher methylation of the chimpanzee alternative promoter could account for a reduction of the *THBS4* transcriptional activity in the chimpanzee and consequently a lower expression relative to humans. Nevertheless, why did we expect transcriptional variation differences related to the promoter methylation when the performed reporter assays did not indicate any? To understand our argument it is necessary to consider the methodology used for the reporter assays. The fragments cloned in the different vectors came from the artificial amplification by PCR, losing all the possible epigenetic marks present in the original sample. Therefore, with the reporter assays the possible hypermethylation of the chimpanzee alternative promoter was not tested.

Rejecting our hypothesis, and in accordance with NUMATA *et al.* (2012), who emphasized that a shorter distance between CpG islands and TSSs decreases CpG hypermethylation, the analysis of the methylated positions within the CpG island upstream the *THBS4* promoter indicates similar low methylation levels in humans and in

chimpanzees. Supporting this result, the whole-genome methylation map of the prefrontal cortex of humans and chimpanzees published just after the realization of our bisulfite-sequencing experiment (ZENG *et al.* 2012), similarly shows no significant methylation differences between species around *THBS4* promoter regions either.

#### 4.3.1.3 Potential effects acting in *cis* at longer distances.

The regulation of eukaryotic gene transcription involves the coordination of many transcription factors and cofactors acting through regulatory DNA sequences (HEINTZMAN *et al.* 2007). The mode of action involves the assembling of several basal factor proteins unto the TSS and generating a full transcription factor complex, finally capable of binding the RNA polymerase. This complex is linked with coactivators and activator proteins late in the process, which in turn bind to other regulatory sequences in the DNA, such as enhancers (LOSOS *et al.* 2008). These enhancers are not always found surrounding the transcriptional start site.

The identification of enhancers has typically relied on two different methodologies. On the one hand, diverse approaches based on computational comparative genomics have been used to evaluate the conservation of the putative enhancers. However, some studies have criticized this evolution-based approach, by emphasizing that a substantial fraction of enhancers displays modest or no conservation across species (BLOW *et al.* 2010, SCHMIDT *et al.* 2010). On the other hand, recent advances in molecular and computational biology have allowed the application of genome-wide tools to the analysis of enhancer structure and function (HEINTZMAN *et al.* 2007). These studies describe enhancers as sequences that may carry epigenetic information in the form of specific histone modifications. Interestingly, some of these histone modifications may serve as marks for future gene expression, whereas others may play a more active part in the transcription activation process (HEINTZMAN *et al.* 2009). Cyclic AMP- responsive element-binding protein (CBP) and p300 are highly similar proteins that have histone acetyltransferase activity and contain a variety of functional domains involved in interactions with other transcription factors or histone modifications

(BEDFORD *et al.* 2010). These two proteins interact with several sequence-specific transcription factors and have been involved in several cell-signaling pathways by activating the transcription of specific genes. Experiments carried out across a 30 Mb region of the human genome have confirmed a correlation between the presence of p300 and enhancer function (CONSORTIUM *et al.* 2007, HEINTZMAN *et al.* 2007). Furthermore, *in vivo* novel enhancers have been identified thanks to the mapping of several thousand p300 binding sites in mouse embryonic forebrain, midbrain, limb and heart, showing different tissue-specific gene expression patterns in transgenic mouse assays (VISEL *et al.* 2009). These results propose that cell type-specific occupancy of enhancers by CBP and/or p300 regulates distinct transcriptional programs in many cell types and, therefore, that these proteins may be a general component of a large class of enhancer elements (HEINTZMAN *et al.* 2009). Putative enhancers, predicted by the presence of distant p300 binding sites, are highly enriched in H3K4me1, H3K4me2 and histone 3 acetylated at lysine 27 (H3K27ac) (HEINTZMAN *et al.* 2007, HEINTZMAN *et al.* 2009). The chromatin state at promoters is mainly invariant across different cell types, whereas histone modifications at enhancers are cell type-specific and are also strongly correlated with gene expression patterns (HEINTZMAN *et al.* 2009).

The main tool for investigating these sequences is chromatin immunoprecipitation (ChIP), in which antibodies are used to select specific proteins or nucleosomes enriching among DNA fragments that are bound to these proteins or nucleosomes. The introduction of microarrays and the developments in next-generation sequencing, allowed the fragments obtained from ChIP to be identified by hybridization to a microarray (ChIP–chip), making possible a genome-scale view of DNA–protein interactions; or instead, to be directly sequenced (ChIP–seq) having a higher resolution, fewer artifacts, greater coverage and a larger dynamic range (PARK 2009).

Considering these possibilities, we carried out two pilot experiments to find possible enhancers regulating *THBS4:* a ChIP analysis coupled with high-throughput sequencing, and a computational search for putative enhancer positions followed by experimental validation. With the first method, we did not find any reliable signal suggesting that the

transcription factor p300 could be binding anywhere near any of the *THBS4* TSSs. In addition, the computational search for enhancers (second method used) resulted in five regions predicted as putative enhancers with a high consistency between the different criteria used for the selection. These criteria were based in some tracks available at the UCSC genome browser: three ENCODE tracks (ENCODE enhancer- and promoter-associated histone mark (H3K4Me1), ENCODE digital DNaseI hypersensitivity clusters, and ENCODE transcription factor ChIP-seq) (CONSORTIUM 2004, CONSORTIUM *et al.* 2007), and other tracks based on the chromatin state segmentation (ChromHMM) (ERNST *et al.* 2011). Additionally, the conservation profile of the different regions and genomic elements was also considered in the search for the candidate enhancer regions, but was not decisive in the final selection since enhancer regions may or may not be conserved between species (BLOW *et al.* 2010, SCHMIDT *et al.* 2010). To validate these candidates we performed reporter assays. These experiments allowed the quantification of the enhancer activity over the transcription of the alternative *THBS4* promoter in humans and chimpanzees. In relation to the possible transcriptional activity between species, no enhanced transcription levels were appreciated between species if the basal transcription of the promoters was considered. However, the chimpanzee Enh4 construct showed lower activity than the human in SH-SY5Y cells (but not in SK-N-AS, in which results were in the opposite direction), suggesting that this region could be differentially regulating both species in SH-SY5Y cell lines and that further characterization of this region might be interesting. It is known that the activity of an enhancer could be restricted to a particular tissue or cell type, a time point in life, or to specific physiological, pathological or environmental conditions (PENNACCHIO *et al.* 2013). Considering this, we cannot strongly conclude that the regions selected do not have enhancer activity. It could only be concluded that in the cell lines and conditions used for the validation, these regions did not regulate the alternative *THBS4* promoter transcription.

#### 4.3.1.4 Limitations of the study.

As it has been mentioned throughout this work, finding the different causes of human brain evolution presents many challenges. In the case of genomic studies to characterize molecularly a specific region like ours, there are also several limitations. Despite the effort to look at the possible regulatory changes in detail, it is not guaranteed to find the genetic explanation to the gene expression differences or to distinguish them from possible environmental effects. Within these studies, functional experiments in humans and other apes are technically restricted to a limited number of immortalized cell-lines or post-mortem tissues. Ideally, subjects would be matched by age, sex, socioeconomic status, cause of death and post-mortem delay (PREUSS *et al.* 2004). But, if finding human normal, control brain tissue with short post-mortem intervals is difficult, getting chimpanzee tissue is even more difficult. There are few research facilities that maintain colonies of chimpanzees; nowadays only five exist in the United States[3]. In addition, since sacrificing chimpanzees is proscribed, chimpanzee tissue (like human tissue) must be acquired post-mortem, being impossible to obtain samples for which the environment has been controlled.

Regarding cell lines as a source of material for experimentation, in this work we had to rely mostly on neuroblastoma cell lines to study the transcriptional activity of the promoters. By definition, a cell line culture is a complex process by which cells are grown under controlled conditions, generally outside of their natural environment. Additionally, in most of the occasions these cell lines have a tumoral origin. Therefore, they naturally do no necessarily replicate what happens in live and healthy tissue, and possible differences in the expression of cell lines do not always correspond with the differences in the expression seen in the normal tissues. In our study, all of the cell lines used for investigating the promoter activity had a human origin; and their genetic background might not correspond

[3] Governmental research facilities in the United States with chimpanzees (not including any other types of facilities housing captive chimpanzees, neither all of the chimpanzees in private labs, since those facilities are not required to divulge such information): Alamogordo Primate Facility (New Mexico), Michale E. Keeling Center for Comparative Medicine and Research (Texas), New Iberia Research Center (Louisiana), Southwest National Primate Research Center (Texas) and Yerkes National Primate Research Center (Georgia).

entirely with that of chimpanzee cells. Moreover, since the set of transcription factors might vary significantly from cell type to cell type (BLANCHETTE 2012), it is also possible that some expression differences affecting any particular regulatory element of *THBS4* promoters may have not been covered in our analysis, because the specific transcription factor required was not expressed in the used cell lines. For example, the gene for β-globin, a protein used in red blood cells for oxygen exchange, and its corresponding upstream regulatory sequences are present in every cell in the human body, but no cell type other than red blood cells expresses β-globin. When REDDY *et al.* (1994) examined the beta globin promoter in different cell types, they found that the two consensus sequences in the β-globin promoter known for binding transcription factors (CCAAT and CACC) were binding the protein in the erythroid cells, but not in the other cell types (REDDY *et al.* 1994). In order to solve these potential problems and in an attempt to cover different regulatory repertoires, we have used as many different neuroblastoma cell lines (and other types of cells expressing *THBS4* such as HEK293T) as were available for us, although the behavior of the promoters in the different cell lines used has been similar.

Despite technical developments, it is still quite difficult to predict transcription factors and enhancers regulating the expression of a particular gene. In addition, due to limitations such as the lack of a specific mark for enhancer identification and enhancer activity/inactivity predictors in a certain cell type or tissue (PENNACCHIO *et al.* 2013), it has been shown that enhancers might not be well conserved in a given tissue (HEINTZMAN *et al.* 2007, SCHMIDT *et al.* 2010). This highlights the importance of studying certain tissues directly from the species under investigation instead of using standard animal models. Moreover, the comparison between the computational prediction methods described to date with experimental validation techniques show that some enhancer regions are missed (false negatives), and other sequences predicted to be active enhancers, such as the five candidates described for *THBS4,* cannot be validated (false positives) (PENNACCHIO *et al.* 2013). It is expected that studies from large groups like the consortium involved in the Encyclopedia of DNA Elements (ENCODE) project, will continue to change their focus from animal models and human cell lines to primary human tissues. This will probably

decrease the number of false negatives and the detection of more different DNA elements computationally. Nevertheless, false positive might be still present due to the inherent problems of the technology by itself. In addition, the next FANTOM (Functional Annotation of the Mammalian genome) project, FANTOM5, was announced to bring a detailed map of active transcriptional enhancers which will be very useful for a better understanding of the long-distance transcriptional regulation (ANDERSSON *et al.* 2012). In particular, a better definition of enhancers not only in humans, but also in chimpanzees could help to identify those that could be involved in differential *THBS4* regulation.

Finally, our study has been limited to look for possible changes acting in *cis,* -setting aside all possible changes in *trans*-. The localization of putative transcription factor binding sites was only carried out as a computational overview of the possible conserved regions between species within 7 kb around the alternative *THBS4* promoter. The comparison between species showed three different putative transcription factor binding sites, one in humans and two in chimpanzees and many others that were conserved between species. However, checking for possible changes in their sequence in different species, their levels of expression and an experimental validation of their effect in the alternative promoter region could be required for further analysis.

Unfortunately, regardless of the different methods used for the molecular characterization of *THBS4* gene, and particularly for the new isoform, one of the main limiatation of our study is that we have not been able to identify the factor(s) responsible for the increase of the gene expression in the human brain (CÁCERES *et al.* 2003, CÁCERES *et al.* 2007). The results obtained in this work, however, point out that the regulation of the alternative isoform is indeed responsible for the aforementioned differences. Considering that any change affecting a *cis*-regulatory element can be expected to affect the affinity to transcription factor binding (either increase or decrease), and consequently being able to modify the efficiency by which regulatory sequences accomplish their function (WITTKOPP and KALAY 2012), we suggest that the differential gene expression might be regulated by a brain-specific enhancer sequence that has so far escaped our scrutiny. In addition, since the differences between humans and chimpanzees are, in ratio, similar in both mRNA isoforms,

the alternative and the reference, we propose that this possible enhancer could be regulating both promoters, being more active in humans than in chimpanzees. Another possible explanation for the similar ratio of expression differences of both promoters could be an increase in the proportion of cells expressing *THBS4* in the human brain, in comparison to the chimpanzee brain. However, given the magnitude of the expression changes (about 5 to 6 times more), and that the visual comparison (even not quantitative) of human, chimpanzee, and macaque *in situ* hybridations performed by CÁCERES *et al.* (2007) did not indicate major differences in the numbers or types of cells expressing *THBS4*, we consider that this explanation is less likely. The expression differences shown between both *THBS4* isoforms could be explained by variation in the regulation of each promoter, such as other enhancer sequences, different transcription factor binding sites, variation in their methylation status, etc.. However, to gain insight into the regulation differences between the two isoforms was not a priority in this thesis project.

### 4.3.2 The role of the alternative isoform.

Among the whole thrombospondin gene family, only *THBS3*, and as has been discovered now, *THBS4*, present more than one described mRNA isoform. At a protein level, all the isoforms of THBS3, as well as, THBS1, THBS2, the reference THBS4 isoform and THBS5/COMP, present a signal peptide at the beginning of their amino acid sequence. Thrombospondins have been described as glycoproteins that undergo transient or longer-term interactions with other extracellular matrix components (ADAMS and LAWLER 2011). To carry out their final function, after translation proteins undergo glycosilation processes in the endoplasmic reticulum. Signal peptides are responsible of the direct import of emerging proteins into the endoplasmic reticulum. In the absence of other trafficking signals, membrane proteins carrying signal peptides are transported to the cell surface, whereas the signal peptides themselves are cleaved off after the successful insertion of the polypeptide into the membrane of the endoplasmic reticulum (GRALLE and PAABO 2011). A functional change in a signal peptide will therefore be reflected in a change of the

efficiency with which mature membrane proteins reach the cell surface. Proteins, like the one produced by the alternative isoform of THBS4 lacking the signal peptide sequence, could avoid the endoplasmic reticulum recognition and thus produce a slightly different protein. This different protein might remain within the cytosol carrying out yet unknown functions.

The increase of alternative *THBS4* expression in the human brain could suggest that the new protein may have a different function in brain tissue activity or in the encephalization process. Considering the possible differences between gene and protein expression levels (PREUSS *et al.* 2004), we strived to quantify both THBS4 proteins levels in different human tissues by western blot analysis, to know whether they correlate with the gene expression differences of both isoforms. Previous western experiments showed the presence of two THBS4 bands in the heart, whereas only one was present in the brain (M. Cáceres, personal communication). These two protein bands could relate to the two mRNA isoforms shown in the gene expression analysis. Even that both heart and brain tissue were expressing both isoforms of the *THBS4* gene, the low levels with which brain tissue expressed the reference isofom could make it undetectable by western blot in brain, but not in heart. Unfortunately, there are not antibodies specific for THBS4 that work very well, and the tests to identify both isoforms at the end of the thesis did not provide useful results.

Based on the first methionine codon in phase with the reading frame, we have speculated that the alternative mRNA of *THBS4* codifies for an 870 amino acid protein. However, the size resolution of the westerns and the variations in the bands due to other possible modifications of the protein structure was not enough to determine whether this size agrees with that of the bands observed. As for further experimentations, it could be interesting to confirm the exact amino acid in which the protein starts by sequencing the N-terminal end. One of the more common methodologies for protein sequencing is based on mass spectrometry, but this method requires an initial big amount of protein that should be inmunoprecipitated from cell line cultures. The second most used method is the Edman degradation reaction, but this method will not work if the N-terminal amino acid of the protein has been chemically modified or if it is concealed within the body of the protein

(DEUTZMANN 2004). A computational prediction of the protein structure could also bring out insights about the localization of the new protein within the cell, its interactions and functions. Finally, in the long term, it would be interesting to check the functional differences between the two isoforms in human cell lines, for example determining the location of the new isoform within different neuroblastoma the cells lines and checking the synaptogenic activity of both isoforms.

### 4.3.3 The likely evolution of *THBS4* promoters.

The identification of homologies, whether morphological, molecular, or genetic, is fundamental to our understanding of biological evolution. The comparison of related genomes has served as an effective way to interpret the genome content. The comparison of the sequenced mammalian genomes revealed hundreds of thousands of conserved DNA regions that have evolved under purifying selection and encompass around 5% of the human genome (LINDBLAD-TOH *et al.* 2011). Surprisingly, only around 30% of these conserved regions (about 1.5% of the whole genome) encodes for protein sequences (LANDER *et al.* 2001). The other 3.5% of the conserved genome relate to non-coding DNA elements, often including *cis*-regulatory elements rarely lost during evolution (HILLER *et al.* 2012).

Thanks to the alignments of the alternative exon sequences through computational genome browsers and additional information about expression analysis in the alternative promoter region, we could suggest that the untranslated region of the alternative isoform is not well conserved in mouse. However, the presence or not of the alternative isoform in mouse and other primate species should be experimentally investigated.

As mentioned before, the analysis of the amino acid sequences of all thrombospondin family members revealed that only the alternative THBS4 protein lacks the signal peptide sequence. This, in addition to the information provided by the alignment conservation,

suggested that the reference isoform of THBS4 is ancestral in relation to the alternative one. Changes in subcellular localization of proteins may be of evolutionary relevance (MARQUES *et al.* 2008). Similar to what happened with the duplication of a gene, the presence of a different isoform of THBS4 could have resulted into a neofunctionalization, where one isoform might have a new function, while the other retains the ancestral function of the progenitor gene. Alternatively, a subfunctionalization could have partitioned the ancestral THBS4 functions between the two isoforms such that their joint levels and patterns of activity are equivalent to the single ancestral gene.

### 4.3.4 Balancing selection in humans: the two *THBS4* haplotypes.

The action of natural selection can cause several different types of changes within a population. How the population changes, depends upon the particular selection pressure the population is subjected to and which traits are favored in the given circumstances. In collaboration with the group of Dr. Sironi from the Scientific Institute IRCCS E.Medea (Bosisio Parini, Lecco, Italy), who performed the different population genetic analysis, it was found that the central region around the exon 3 of the *THBS4* gene displayed unusually low fixation index ($F_{ST}$) (WRIGHT 1950), which extended through a region of roughly 8.4 kb (covering exons 3–5) (CAGLIANI *et al.* 2013). Posterior tests to evaluate deviations of the allele frequency distribution from the expected pattern of neutral variation suggested that this *THBS4* region could have been regulated under balancing selection within the human populations (CAGLIANI *et al.* 2013). This kind of natural selection is an important biological force, under which alleles contributing to an adaptive phenotype are maintained at an equilibrium frequency, which maximizes the mean fitness of the population as long as the selective pressure is acting (CHARLESWORTH 2006, ANDRES *et al.* 2009). The analysis of haplotype genealogy of the *THBS4* region indicated the existence of two major clades with a Time to Most Recent Common Ancestor (TMRCA) estimated in the range of 3 to 4 million years.

The fact that *THBS4* could be a target of balancing selection, implicates that at least one or more variants might have a functional effect on tissue-specific and/or sex-specific activity. The analysis of SNPs located along the major branches of the haplotype genealogy indicated the presence of 65 polymorphisms. Only SNP rs438042 (A/T), which is located at the end of exon 3, affects the *THBS4* coding region, although the two alleles code for the same amino acid. As part of our work, we analyzed the expression of *THBS4* by real-time PCR quantification using mRNA from 17 adult brain samples previously genotyped for exon 3 rs438042 SNP. Accounting for the effect of age, sex and genotype, the results indicated a significant effect of age on the expression of *THBS4,* but no effect of sex or genotype, maybe because of a lack of power due to the small sample size. To confirm that there was not an effect of the genotype on *THBS4* expression, allele-specific real-time PCRs were performed for the same human samples. Supporting the similar levels of *THBS4* expression among the different genotypes, the quantification of the A/T heterozygote samples presented similar expression levels of each allele. Additionally, using existing microarray data to obtain independent confirmation for the observed increase in *THBS4* expression with age, a strong effect of postnatal age on *THBS4* levels was found, based on the data from COLANTUONI *et al.* (2011) and SOMEL *et al.* (2011). However, analysis of data from BERCHTOLD *et al.* (2008) revealed no changes in *THBS4* expression with age in any brain area analyzed (entorhinal cortex, hippocampus, post-central gyrus, and superior frontal gyrus). Interestingly the analysis of the RNA sequencing experiments from SOMEL *et al.* (2011), in line with the previous microarray work of CÁCERES *et al.* (2007), indicated that *THBS4* expression increases during life in human prefrontal cortex but not in chimpanzee and macaque or in the cerebellar cortex of any species. It has been described that developmental changes in gene expression are a specific feature of the human brain (SOMEL *et al.* 2011) suggesting that increased developmental plasticity is selected in evolution.

Finally, to verify whether *THBS4* variants located within the balancing selection region influence the brain expression level of the gene in a larger sample, data from MYERS *et al.* (2007) was reanalyzed. The analysis of the human brain specimens from elderly donors (average age of 81 years) detected an interaction between sex and *THBS4* genotype (mainly mediated by AA homozygous females) in modulating expression levels. These results

however, were not detected when the analysis was performed in lymphoblastoid cell lines. It could be reasonable to assume that since this is a sex-specific effect, it could be mediated by different hormones, producing differences between the primary tissue and the cultured cells which lack a hormone-rich media. On the other hand, it could be also possible that the discrepant results might be due to different tissue specificity, as has been proposed for balancing selection at regulatory variants, resulting from the ability of distinct alleles/haplotypes to confer preferential expression in a specific tissue or cell-type, or to modulate transcription in response to diverse stimulus (LOISEL *et al.* 2006, CAGLIANI *et al.* 2008). The analysis of 81 patients with Alzheimer's disease previously genotyped for rs438042 showed a significant interaction between sex and genotype in gray matter and peripheral gray matter tissue, also mediated in its majority by female AD patients with an AA genotype. Although the age average between female patients (mean age = 76.2) was the same as the male patients with the same genotype (mean age = 76.8), the volume of the gray matter and peripheral gray matter were markedly reduced in these females. These data could suggest a role of *THBS4* variants in the brain volume of Alzheimer's patients. However, it is still not very clear what the specific targets of the balancing selection are or if they relate to the increase of expression levels in human evolution. It could be tempting to speculate that increased *THBS4* expression levels might be somehow related to this.

# 5

# CONCLUSIONS

# 5

# CONCLUSIONS

The following conclusions can be drawn from this work:

1. The genomic context affecting *THBS4* expression is equivalent between humans and chimpanzees with few structural differences. The *THBS4* gene is included within a chimpanzee-specific inversion, which has its closest breakpoint 16 Mb away and does not appear to affect the possible regulatory elements controlling *THBS4* gene expression

2. Analysis of available expressed sequences indicated the presence of a previously unknown alternative *THBS4* mRNA beginning 44 kb upstream from the known *THBS4* transcription start site, which was confirmed by PCR amplification and sequencing. This new mRNA has a length of 3115 bp and presents 23 exons, two of them specific for the alternative isoform and 21 shared with the reference isoform of the gene, corresponding with exons 2 to 22 of the reference mRNA.

3. *THBS4* reference mRNA presents the first ATG codon encoding for the initial methionine of the protein and the beginning of the 26 aa signal peptide sequence in exon 1. The alternative mRNA of *THBS4* lacks this exon 1, losing the signal peptide and probably shifting the start of translation until the exon 2, where the next ATG is found.

4. In humans, *THBS4* is expressed in heart, brain, thymus, skeletal muscle, colon, testis and ovary. Quantitative RT-PCR to compare expression levels of both mRNAs revealed that colon and heart express primarily the reference mRNA and sexual organs and brain the alternative mRNA.

5. The alternative *THBS4* mRNA isoform is not human specific and is also expressed in the brain of non-human primates, but it is expressed around 5 times higher in humans than in chimpanzee frontal cortex.

6. Evaluation of the transcriptional activity of both *THBS4* promoter regions from humans and chimpanzees using reporter assays showed significant differences between both promoters, but not between species, in the neuroblastoma cell lines assayed. This is consistent with the search of transcription factor binding sites in the alternative promoter region, which only found three TFBS differentially predicted between both species.

7. Comparison of the DNA methylation levels of a CpG island upstream the alternative transcription start site in five humans and five chimpanzees detected similar low methylation levels between species. Two CpG dinucleotides stood out from the unmethylated background, however the number of clones methylated in comparison to the number of clones analyzed was to low to consider an implication in the promoter regulation.

8. The search for a putative enhancer region controlling the *THBS4* alternative promoter did not provide any reliable candidate. However, taking into account all the results we consider that the increased *THBS4* expression is probably regulated by a brain-specific enhancer sequence that has so far escaped our scrutiny.

9. The expression analysis of SNP rs438042, which had been associated to two *THBS4* haplotypes that are maintained by balancing selection, indicated no effect of the genotype over *THBS4* gene expression.

10. Alignments of the alternative exon sequences in different vertebrate species suggested that the untranslated region of the alternative isoform is not well conserved and no evidence of expression of this alternative mRNA was found in mouse. In addition, the lack of the signal peptide in the alternative THBS4 protein, which is conserved in the

rest of the family members, suggests that the reference isoform of *THBS4* is ancestral than the alternative one.

# APPENDIX

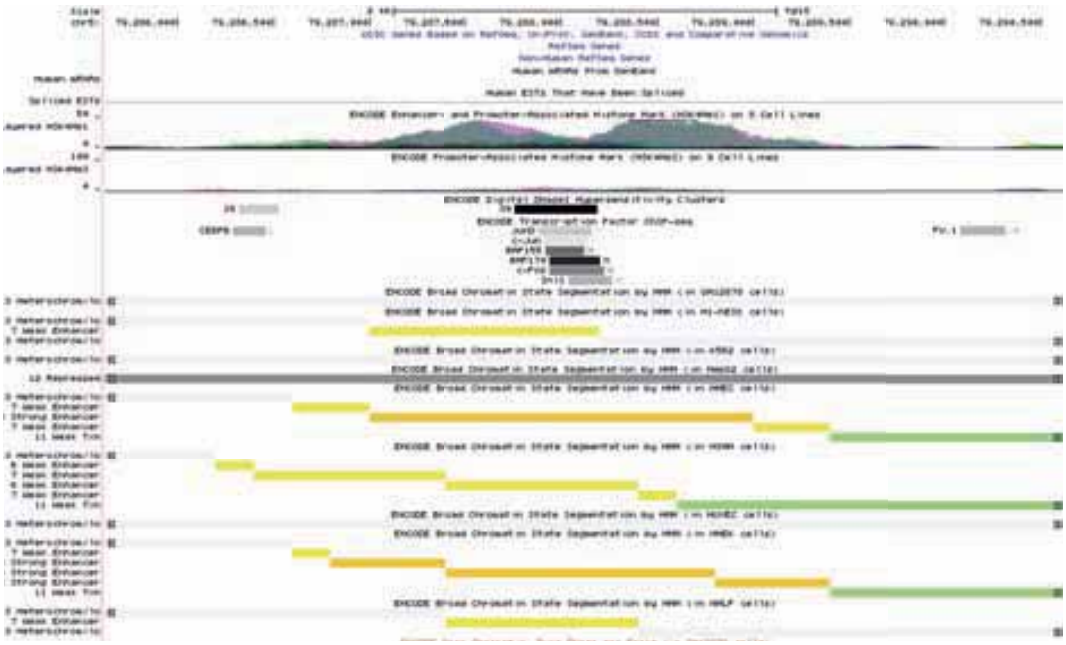
# I

# APPENDIX I

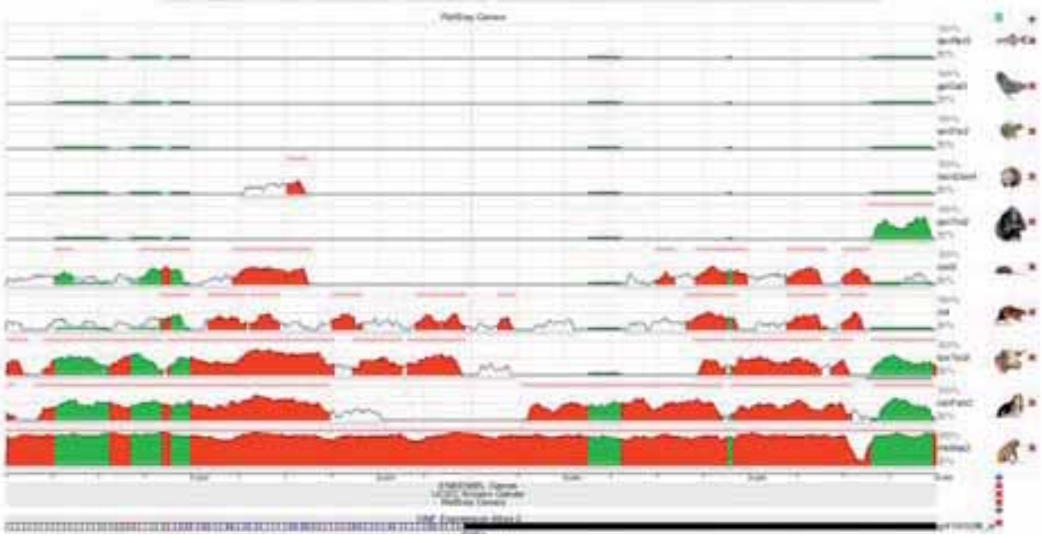
Enhancer region 1

Position: Chr5: 79,285,625–79,290,624

UCSC ENCODE tracks:



Evolutionary conserved regions:

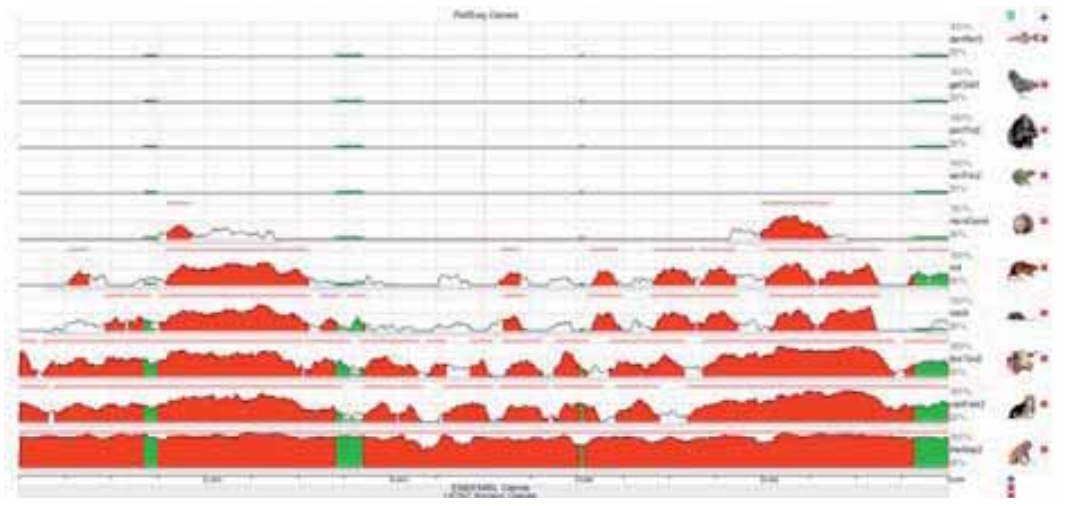


Enhancer region 2

Position: Chr5: 79,291,375–79,296,374

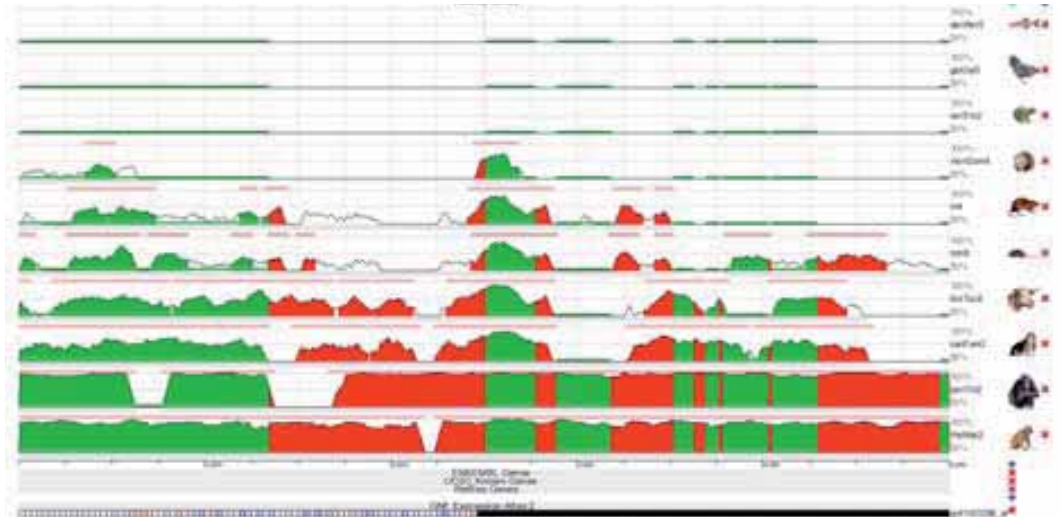
UCSC ENCODE tracks:

Evolutionary conserved regions:



Enhancer region 3
Position: Chr5: 79,300,375–79,305,374
UCSC ENCODE tracks:

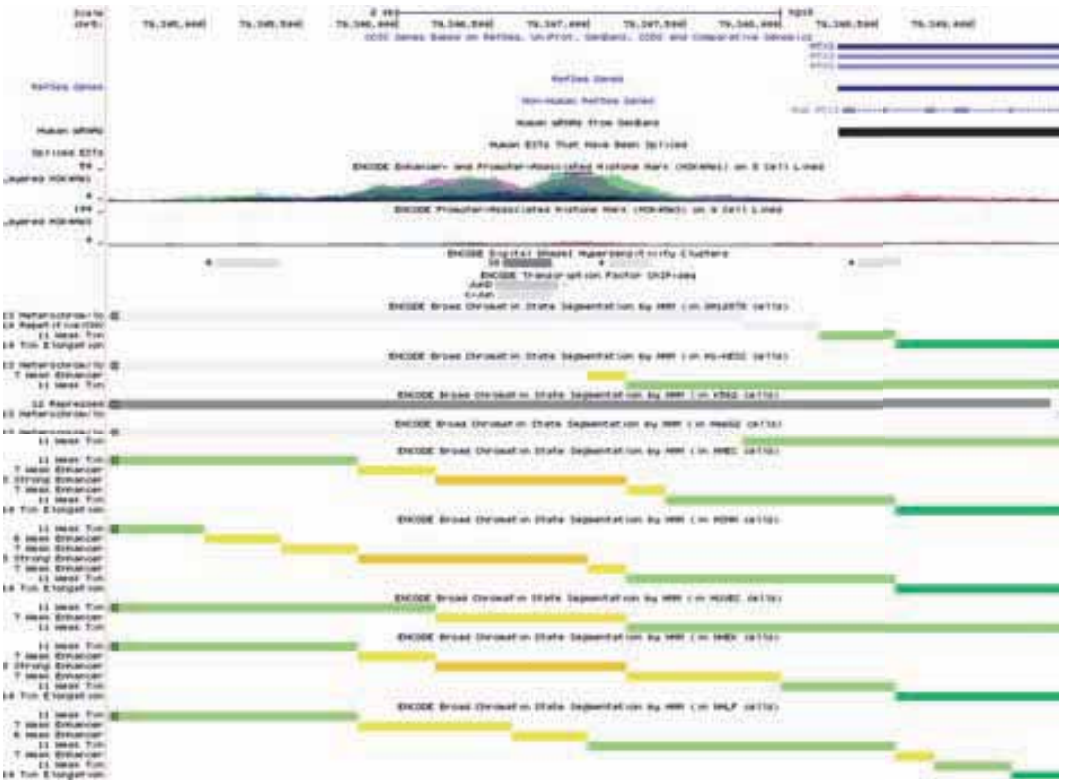
Evolutionary conserved regions:



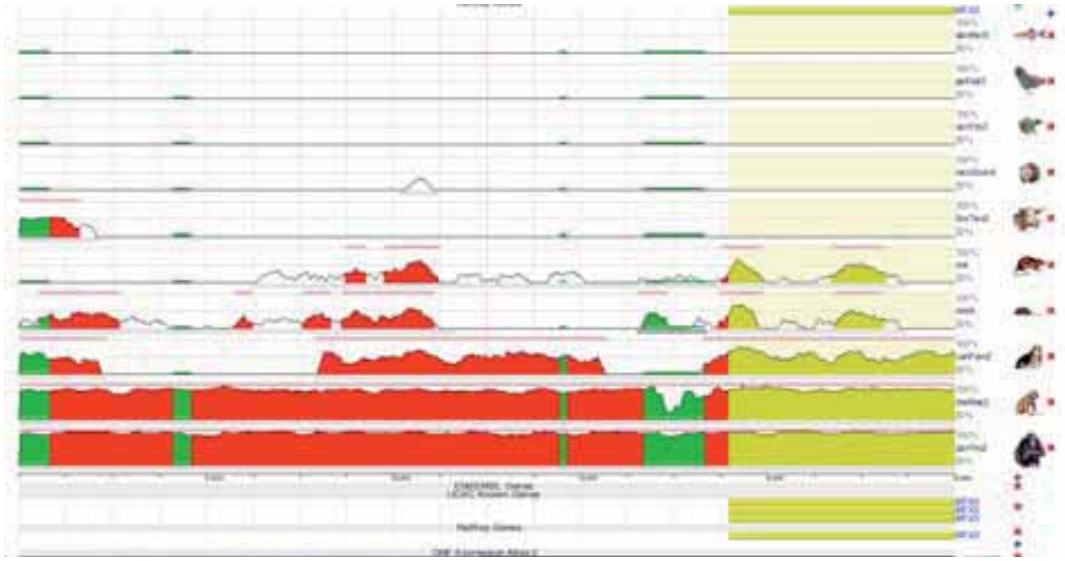
Enhancer region 4

Position: Chr5: 79,304,501 –79,309,500

UCSC ENCODE tracks:



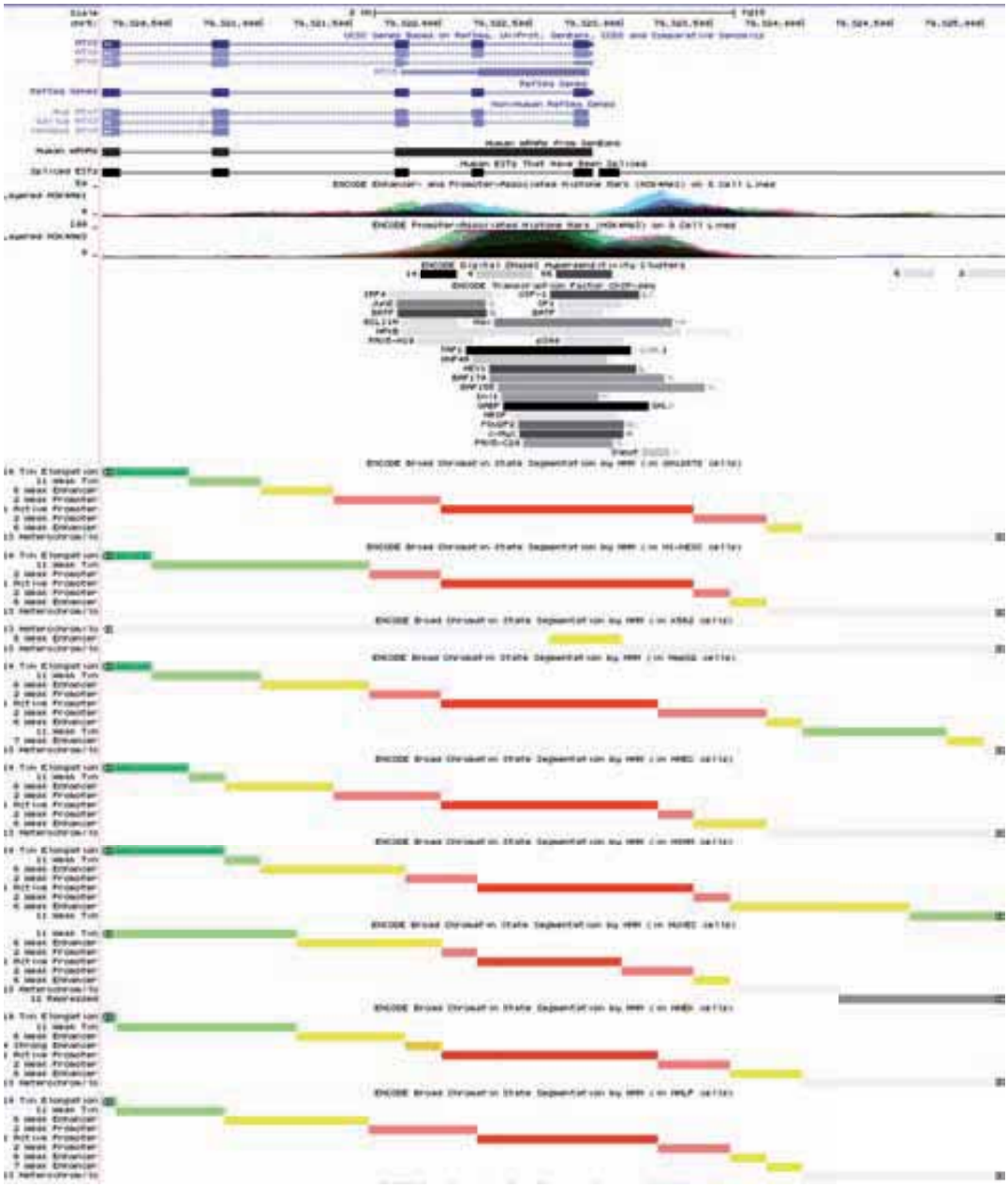
Evolutionary conserved regions:



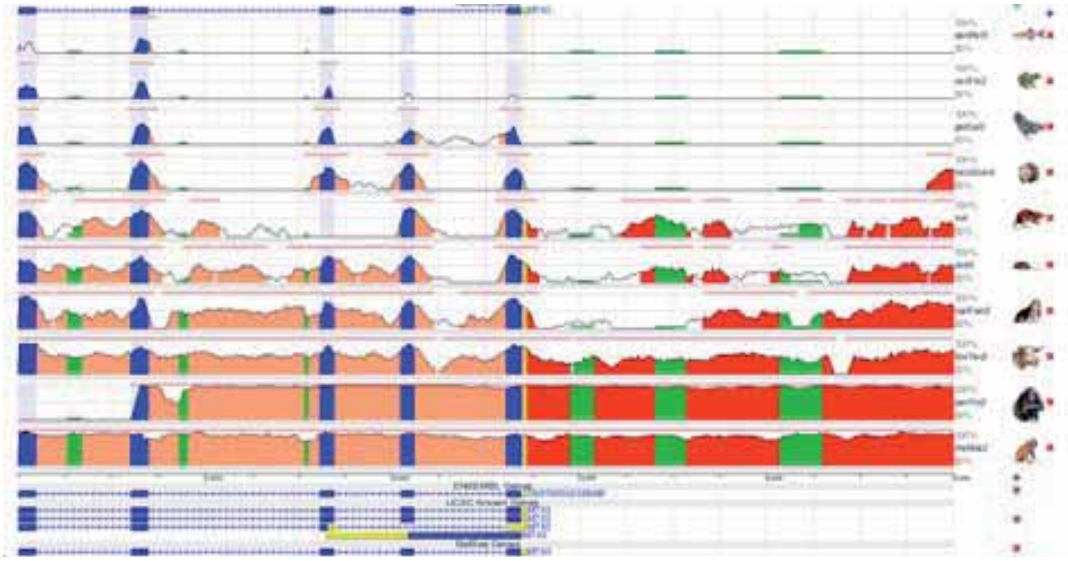
Enhancer region 5

Position: Chr5: 79,320,126–79,325,125

UCSC ENCODE tracks:

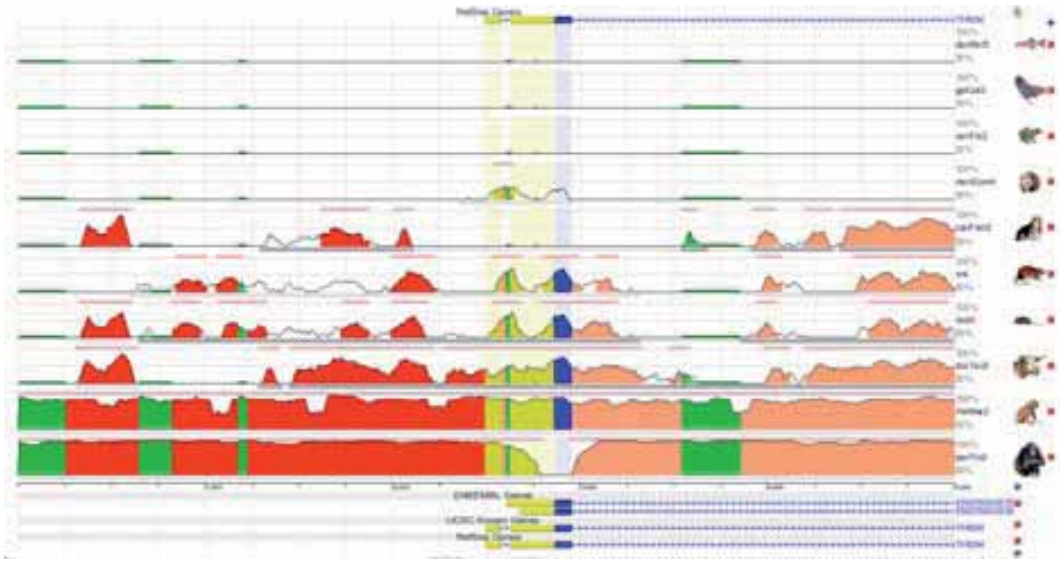


Evolutionary conserved regions:



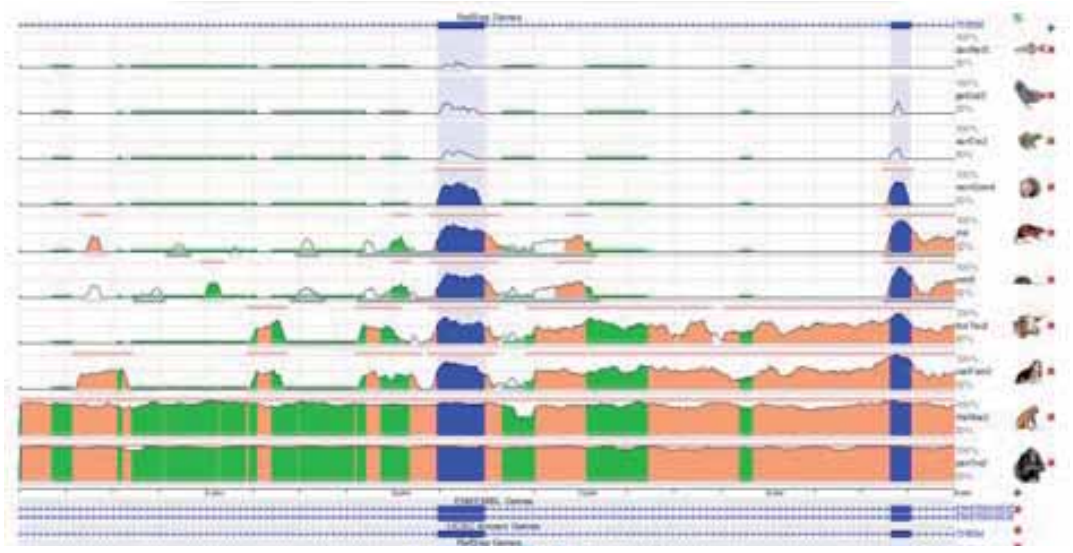
Enhancer region 6
Position: Chr5: 79,364,251–79,369,250
UCSC ENCODE tracks:

Evolutionary conserved regions:

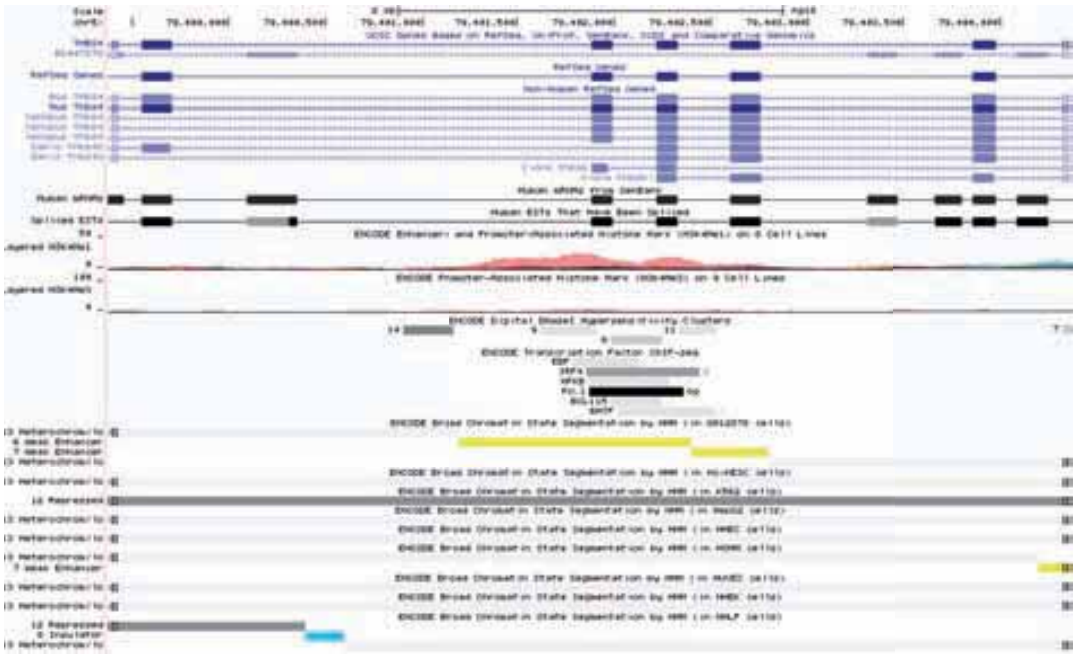


Enhancer region 7
Position: Chr5: 79,385,126–79,390,125
UCSC ENCODE tracks:

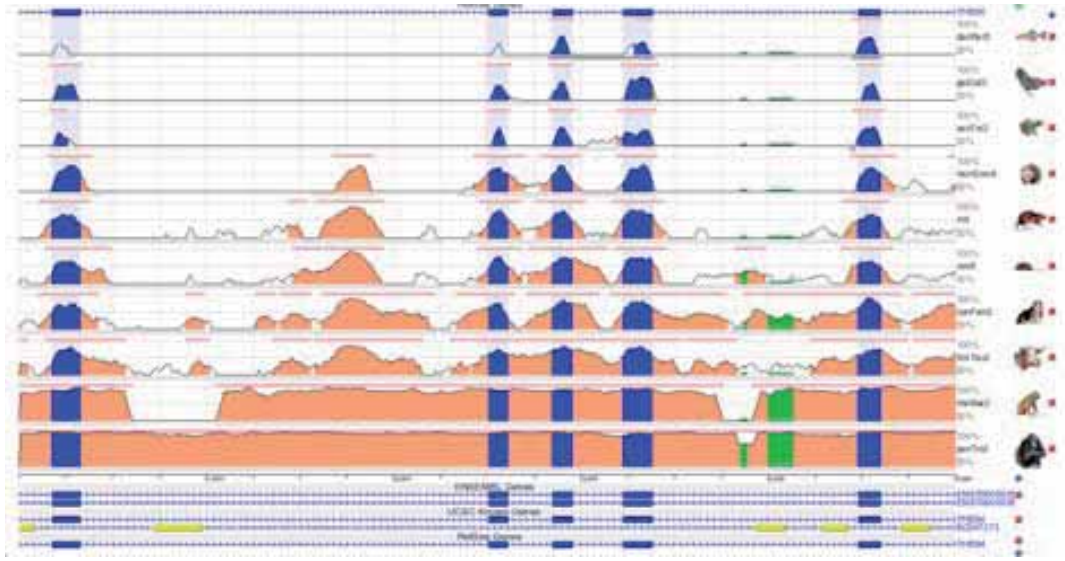
Evolutionary conserved regions:



Enhancer region 8
Position: Chr5: 79,399,376 –79,404,375
UCSC ENCODE tracks:



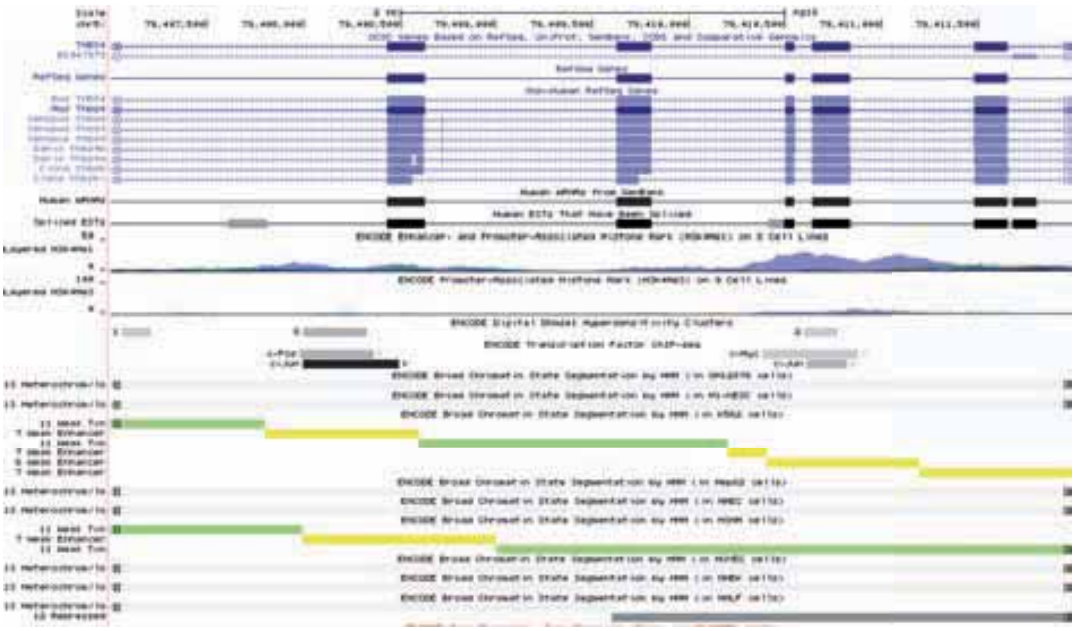
Evolutionary conserved regions:



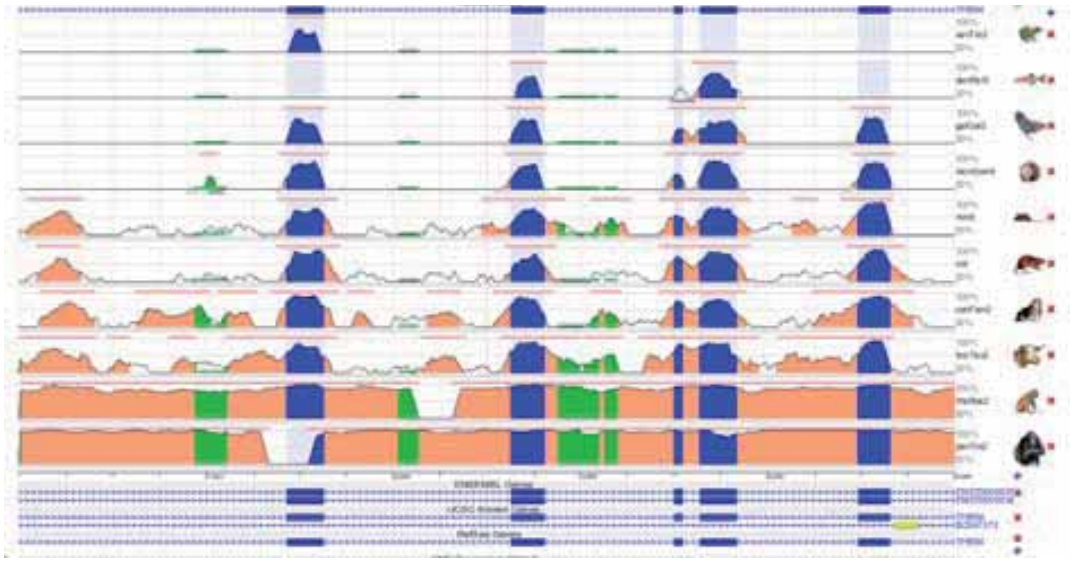
Enhancer region 9

Position: Chr5: 79,407,001 –79,412,000

UCSC ENCODE tracks:



Evolutionary conserved regions:

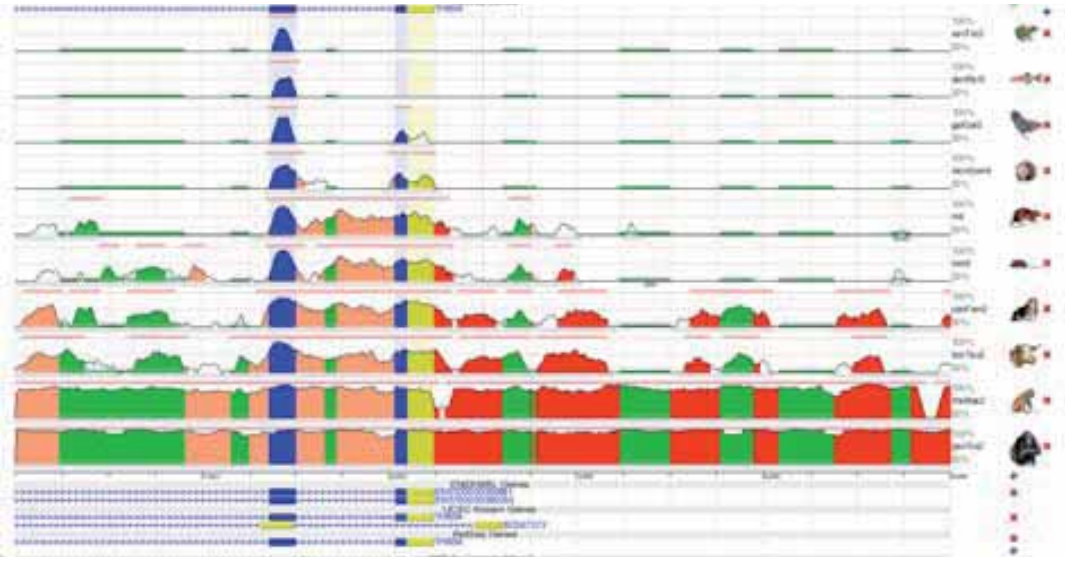


Enhancer region 10

Position: Chr5: 79,412,626–79,417,625

UCSC ENCODE tracks:

Evolutionary conserved regions:

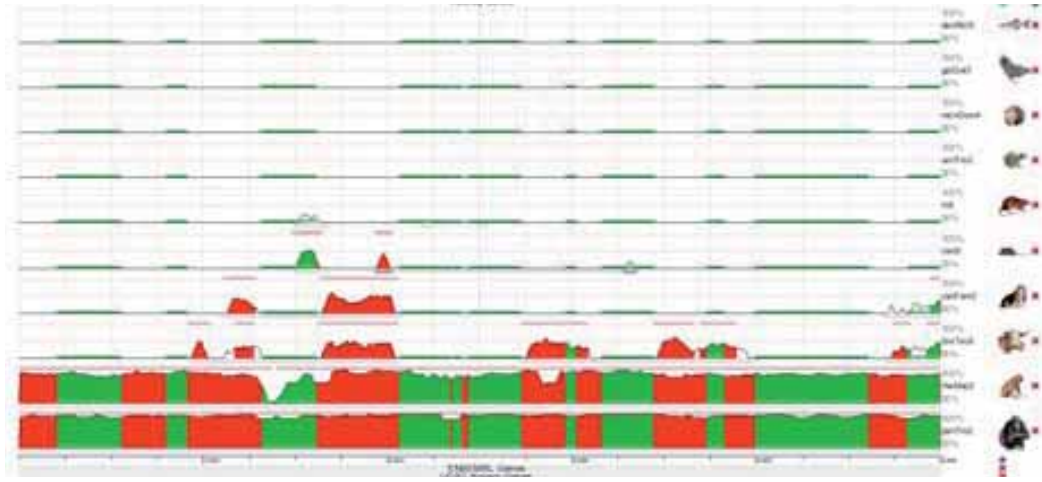


Enhancer region 11

Position: Chr5: 79,421,501 –79,426,500

UCSC ENCODE tracks:

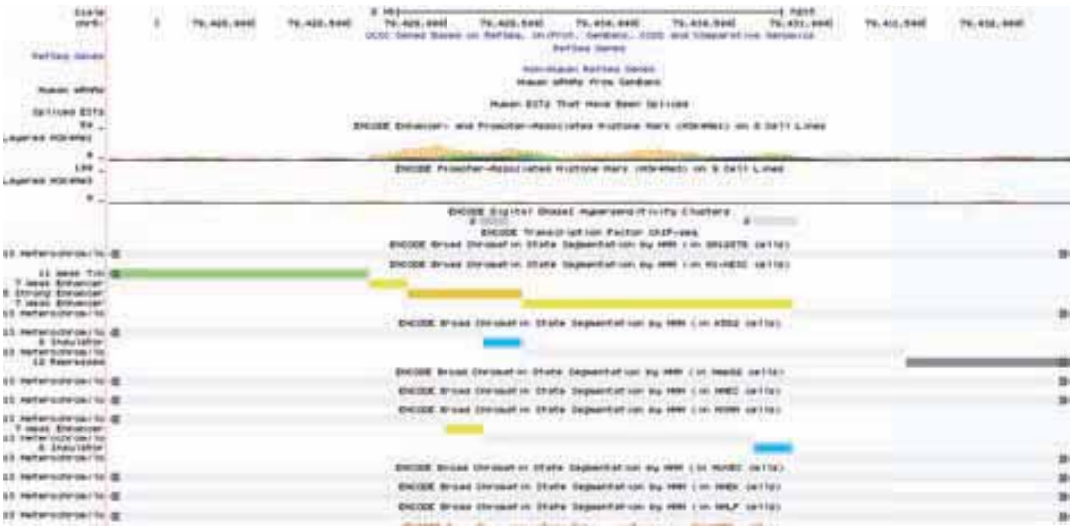
Evolutionary conserved regions:



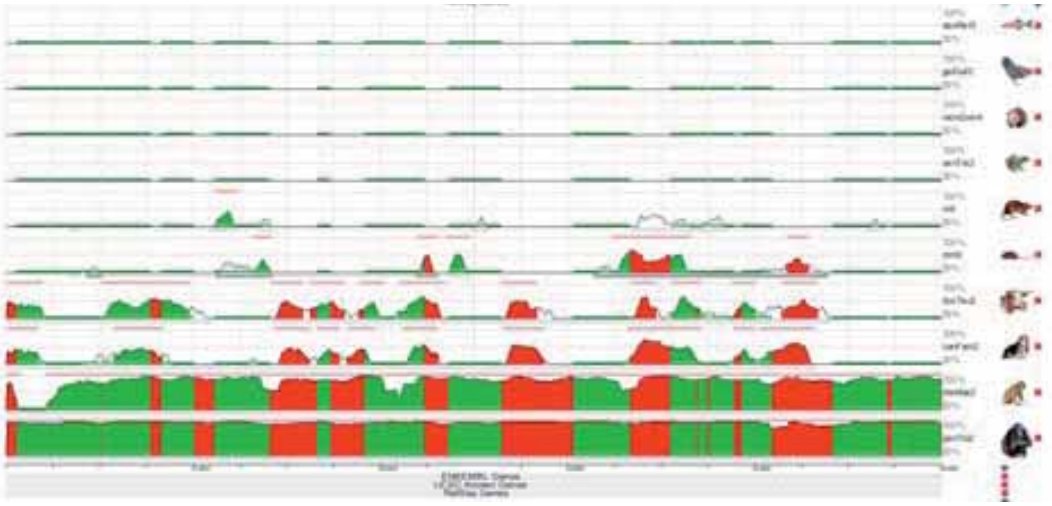
Enhancer region 12

Position: Chr5: 79,427,251–79,432,250

UCSC ENCODE tracks:

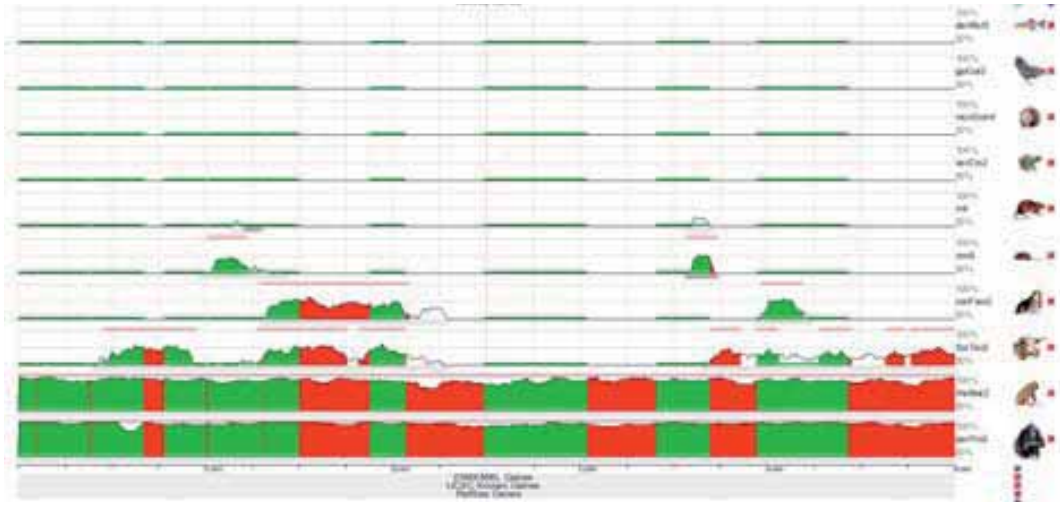


Evolutionary conserved regions:



Enhancer region 13
Position: Chr5: 79,432,001–79,437,000
UCSC ENCODE tracks:

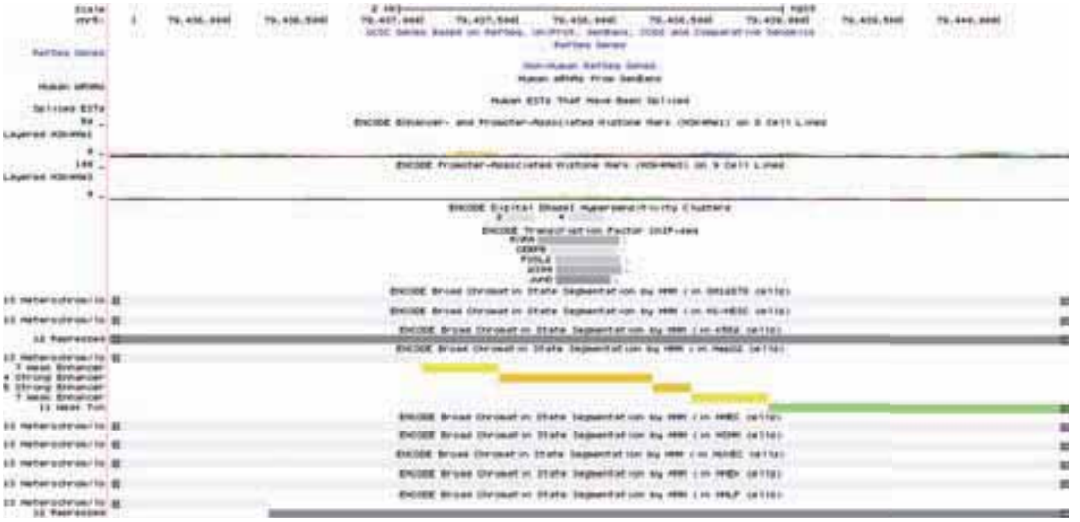
Evolutionary conserved regions:



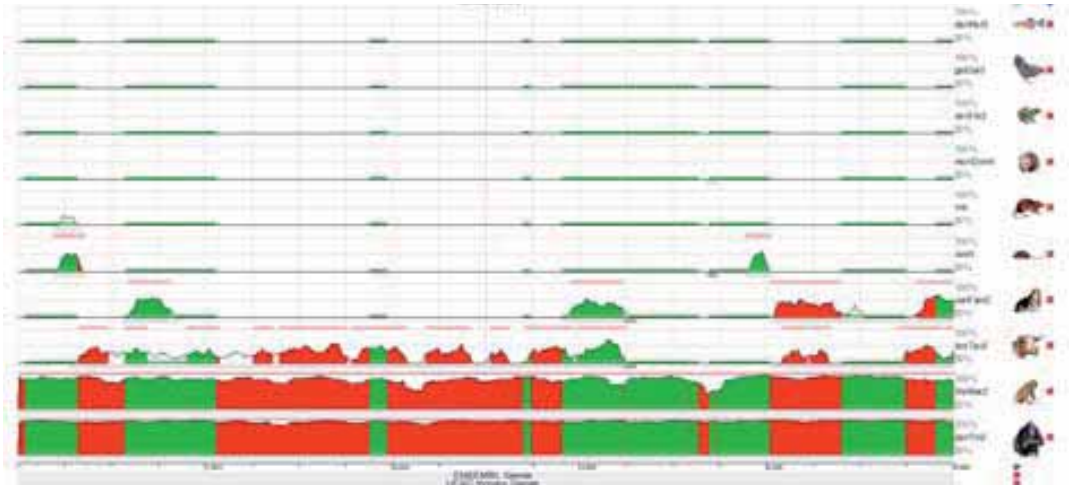
Enhancer region 14

Position: Chr5: 79,435,376–79,440,375

UCSC ENCODE tracks:

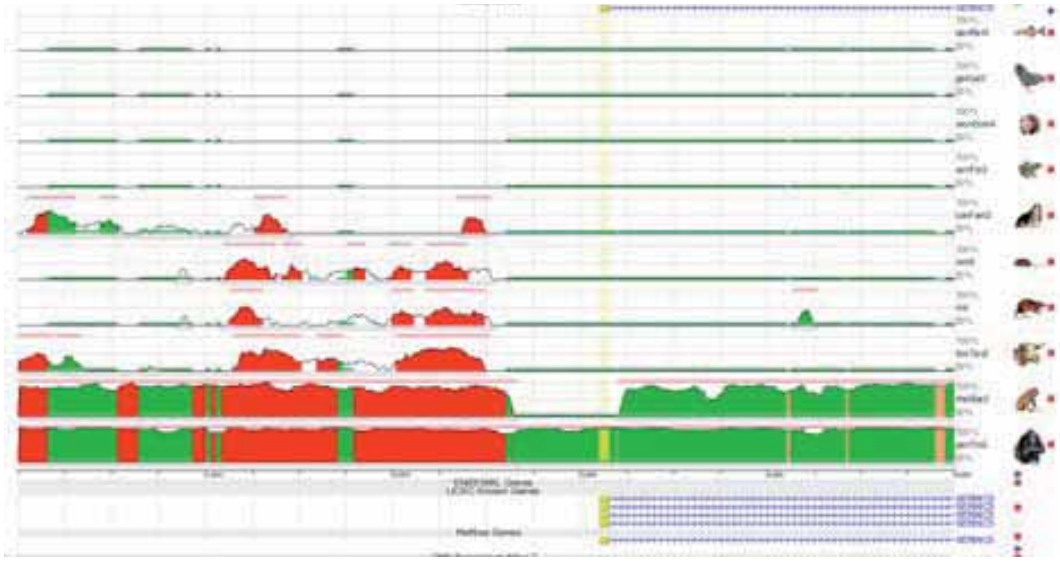


Evolutionary conserved regions:



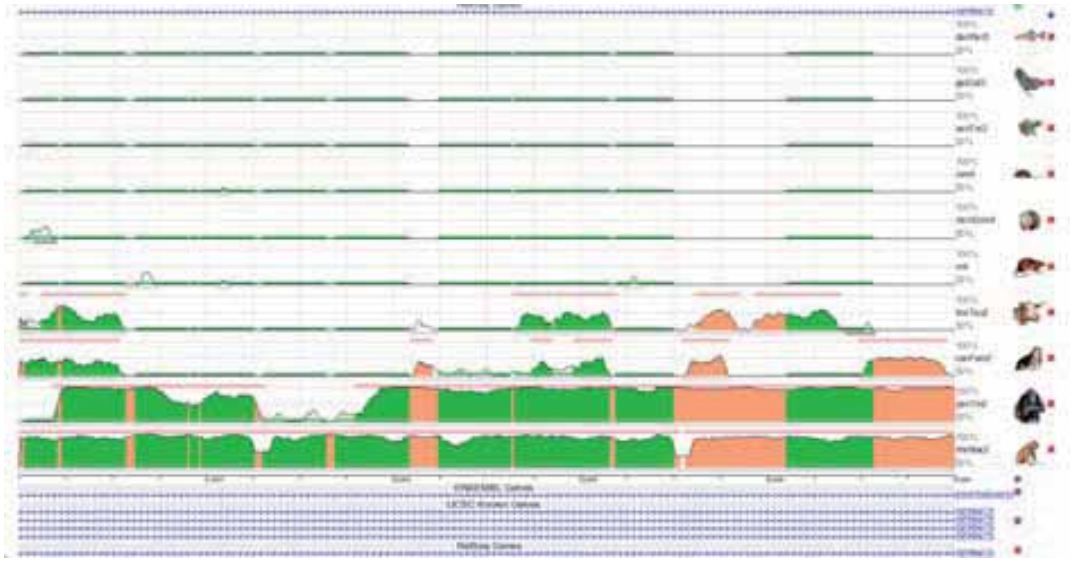
Enhancer region 15
Position: Chr5: 79,440,126–79,445,125
UCSC ENCODE tracks:

Evolutionary conserved regions:



Enhancer region 16
Position: Chr5: 79,440,126–79,445,125
UCSC ENCODE tracks:

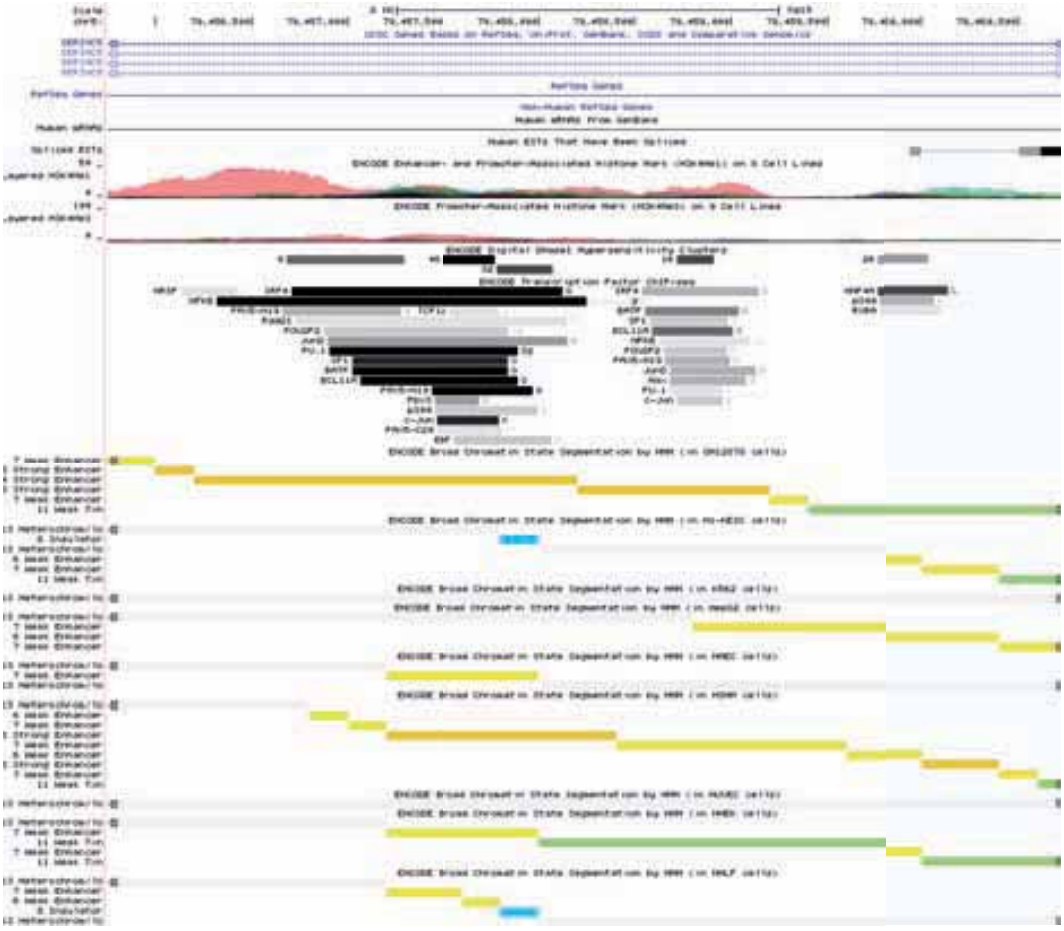
Evolutionary conserved regions:



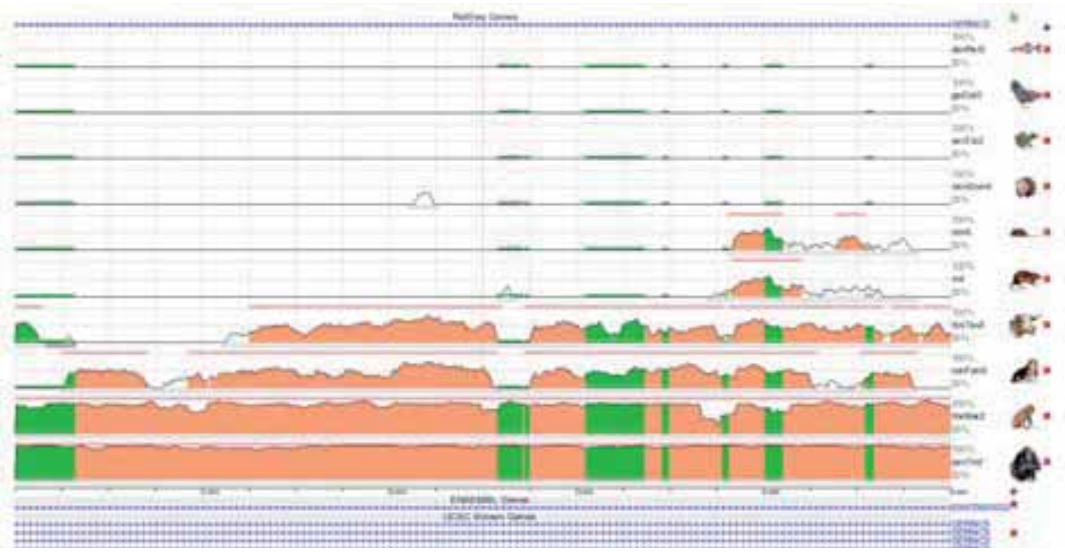
Enhancer region 17

Position: Chr5: 79,455,751–79,460,750

UCSC ENCODE tracks:

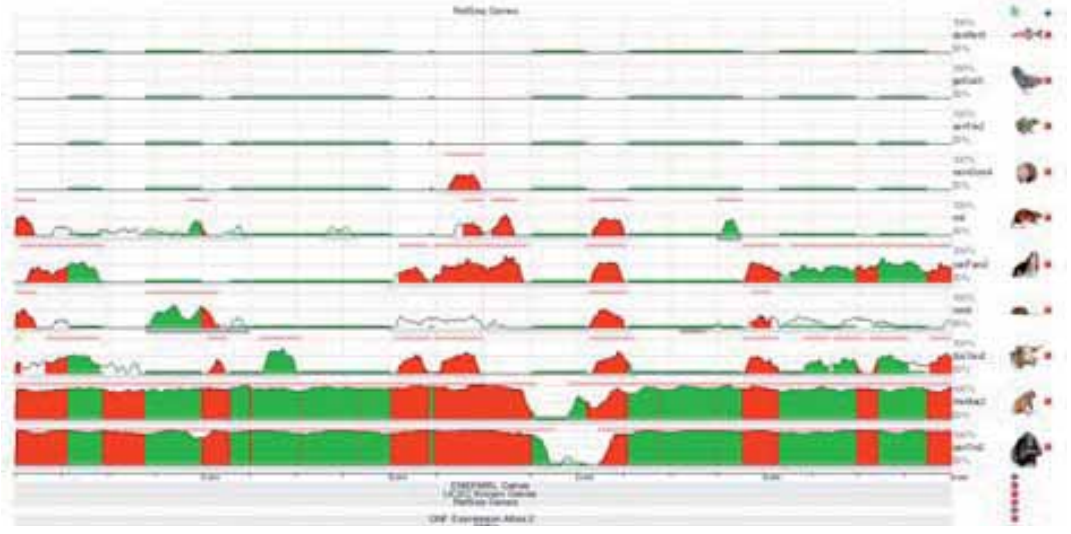


Evolutionary conserved regions:



Enhancer region AluY
Position: Chr5: 79,358,501–79,363,500
UCSC ENCODE tracks:

Evolutionary conserved regions:



# APPENDIX

# II

RESEARCH ARTICLE

Human Mutation

OFFICIAL JOURNAL
HGVS
HUMAN GENOME VARIATION SOCIETY
www.hgvs.org

# Long-Standing Balancing Selection in the *THBS4* Gene: Influence on Sex-Specific Brain Expression and Gray Matter Volumes in Alzheimer Disease

Rachele Cagliani,[1†] Franca R. Guerini,[2†] Raquel Rubio-Acero,[3] Francesca Baglio,[2] Diego Forni,[1] Cristina Agliardi,[2] Ludovica Griffanti,[2] Matteo Fumagalli,[1] Uberto Pozzoli,[1] Stefania Riva,[1] Elena Calabrese,[2] Martin Sikora,[4‡] Ferran Casals,[4] Giacomo P. Comi,[5] Nereo Bresolin,[1,5] Mario Cáceres,[3,6] Mario Clerici,[2,7] and Manuela Sironi[1]*

[1]*Scientific Institute IRCCS E. Medea, Bosisio Parini (LC), Italy;* [2]*Don C. Gnocchi Foundation ONLUS IRCCS, Milan, Italy;* [3]*Institut de Biotecnologia i de Biomedicina Universitat Autònoma de Barcelona, Bellaterra, Barcelona, Spain;* [4]*Institute of Evolutionary Biology (UPF-CSIC) Universitat Pompeu Fabra, Barcelona, Catalonia, Spain;* [5]*Dino Ferrari Centre, Department of Physiopathology and Transplantation, University of Milan, Fondazione Ca' Granda IRCCS Ospedale Maggiore Policlinico, Milan, Italy;* [6]*Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Spain;* [7]*Chair of Immunology, Department of Physiopathology and Transplantation, University of Milan, Milano, Italy*

Communicated by Christine Van Broeckhoven

Received 13 November 2012; accepted revised manuscript 1 February 2013.
Published online 19 February 2013 in Wiley Online Library (www.wiley.com/humanmutation). DOI: 10.1002/humu.22301

**ABSTRACT:** The *THBS4* gene encodes a glycoprotein involved in inflammatory responses and synaptogenesis. *THBS4* is expressed at higher levels in the brain of humans compared with nonhuman primates, and the protein accumulates in β-amyloid plaques. We analyzed *THBS4* genetic variability in humans and show that two haplotypes (hap1 and hap2) are maintained by balancing selection and modulate *THBS4* expression in lymphocytes. Indeed, the balancing selection region covers a predicted transcriptional enhancer. In humans, but not in macaques and chimpanzees, *THBS4* brain expression increases with age, and variants in the balancing selection region interact with sex in influencing *THBS4* expression ($p_{interaction} = 0.038$), with hap1 homozygous females showing lowest expression. In Alzheimer disease (AD) patients, significant interactions between sex and *THBS4* genotype were detected for peripheral gray matter ($p_{interaction} = 0.014$) and total gray matter ($p_{interaction} = 0.012$) volumes. Similarly to the gene expression results, the interaction is mainly mediated by hap1 homozygous AD females, who show reduced volumes. Thus, the balancing selection target in *THBS4* is likely represented by one or more variants that regulate tissue-specific and sex-specific gene expression. The selection signature associated with *THBS4* might not be related to AD pathogenesis, but rather to inflammatory responses.
Hum Mutat 00:1–11, 2013. © 2013 Wiley Periodicals, Inc.

**KEY WORDS:** *THBS4*; balancing selection; transcriptional regulation; brain; Alzheimer disease

## Introduction

Thrombospondins (THBS) are large extracellular-matrix glycoproteins involved in different biological processes including cell adhesion, wound healing, angiogenesis, vessel wall biology, and connective tissue organization [Adams and Lawler, 2004; Adams, 2004; Elzie and Murphy-Ullrich, 2004]. In the human genome, five *THBS* genes (*THBS1–4*, and *COMP/THBS5*) code for modular protein products with synaptogenic potential [Christopherson et al., 2005; Eroglu et al., 2009], suggesting that they play a role in central nervous system development.

From an evolutionary perspective, *THBS4* (MIM #600715) may be considered the most interesting thrombospondin gene, as previous studies indicated a sixfold higher expression of the mRNA and protein in the adult cortex of humans compared with chimpanzees and macaques [Caceres et al., 2003; 2007]. *THBS4* is expressed by various cell types in the frontal cortex of humans (and, at lower levels, of nonhuman primates) [Caceres et al., 2007], and protein expression is particularly strong in the synapse-rich neuropil in humans [Caceres et al., 2007]. Importantly, *THBS4* protein was also shown to accumulate in β-amyloid plaques [Caceres et al., 2007], which are abundant in Alzheimer disease (AD), a neurodegenerative condition associated with inflammation. This observation, together with the reported proinflammatory effects of THBS4, suggest that this protein may be involved in the pathogenesis of AD. Indeed, *THBS4* is expressed by vascular endothelial and smooth muscle cells [Stenina et al., 2007], where it stimulates leukocyte adherence and induction of inflammatory responses as a consequence of its ability to bind integrin $\alpha_M\beta_2$ (also known as Mac-1) [Pluskota et al., 2005]. Interestingly, Mac-1 is also expressed by glial cells and is central in mediating β-amyloid peptide-induced microglial activation [Zhang et al., 2011].

Studies of *THBS4* genetic diversity in humans have mostly focused on a nonsynonymous polymorphism (p.Ala387Pro, rs1866389:G>C) located in exon 9. In fact, the two protein variants display different functional properties with respect to endothelial cell adhesion and proliferation [Stenina et al., 2007], as well as induction of neutrophil inflammatory responses, with the 387Pro allele possibly representing a proatherogenic variant [Pluskota et al., 2005]. However, very little is known about the effects of *THBS4* variants on brain-related phenotypes.

Additional Supporting Information may be found in the online version of this article.

†These authors equally contributed to this work.

‡Present address: Department of Genetics, Stanford University, Stanford, CA 94305, USA.

*Correspondence to: Manuela Sironi, Scientific Institute IRCCS E. Medea, Bioinformatic Lab, Via don L. Monza 20, 23842 Bosisio Parini (LC), Italy. E-mail: manuela.sironi@bp.lnf.it

Contract grant sponsors: Ministerio de Educación y Ciencia, Spain (BFU2007-60930); Italian Ministry of Health; Fondazione CARIPLO.

## Materials and Methods

### Sequencing, Population Genetics, and Statistical Analysis

Human genomic DNA from Yoruba (YRI) and Asian (AS) individuals was obtained from the Coriell Institute for Medical Research. A 8.4 kb region spanning *THBS4* (NM_003248.4) exons 3–5 (chr5:79,382,598–79,390,997, hg18 Assembly) was PCR amplified and directly sequenced (primer sequences are available as Supp. Table S1). PCR products were treated with ExoSAP-IT (USB Corporation, Cleveland, OH), directly sequenced on both strands with a Big Dye Terminator sequencing Kit (v3.1, Applied Biosystems, Monza, Italy) and run on an Applied Biosystems ABI 3130 XL Genetic Analyzer (Applied Biosystems). Sequences were assembled using AutoAssembler version 1.4.0 (Applied Biosystems), inspected manually by two distinct operators; singletons were reamplified and resequenced. Genotype data for African Americans (AAs) and Europeans (EUs) were retrieved from the SeattleSNPs Website (http://pga.mbt.washington.edu). Haplotypes were inferred using PHASE version 2.1 [Stephens et al., 2001; Stephens and Scheet, 2005].

The $F_{ST}$ statistic [Wright, 1950] estimates genetic differentiation among populations and was calculated as previously proposed [Hudson et al., 1992]. For the sliding window analysis of *THBS4*, windows of 5 kb moving with a step of 150 bp were used, and the same procedure was applied to all genes resequenced by the SeattleSNPs program. Values deriving from sliding windows obtained from all genes resequenced in SeattleSNPs panel 1 (i.e., the DNA panel including AA and EU subjects) were used to identify the 2.5th percentile. It is worth noting that negative $F_{ST}$ should be interpreted as 0 and the 2.5th percentile value of $F_{ST}$ from SeattleSNPs gene-sliding windows resulted extremely close to 0.

The maximum-likelihood-ratio HKA (MLHKA) test was performed as previously described [Fumagalli et al., 2009] using the MLHKA software [Wright and Charlesworth, 2004] and multilocus data of 16 selected gene regions (reference regions). *Pan troglodytes* (NCBI panTro2) was used as an out-group.

Tajima's $D$ [Tajima, 1989], Fu and Li's $D^*$ and $F^*$ [Fu and Li, 1993] statistics, as well as diversity parameters $\theta_W$ [Watterson, 1975] and $\pi$ [Nei and Li, 1979] were calculated using *libsequence* [Thornton, 2003]. Calibrated coalescent simulations were performed using the *cosi* package [Schaffner et al., 2005] and its best-fit parameters for YRI, AA, EU, and AS populations with 10,000 iterations. As for the empirical comparison, genotype data for 238 resequenced human genes were derived from the NIEHS SNPs Program Website (http://egp.gs.washington.edu/). In particular, we selected genes that had been resequenced in populations of defined ethnicity including AA, EU, YRI, and AS (NIEHS panel 2). Reference windows (5 kb) were randomly selected for each gene resequenced by the NIEHS program; the only requirement was that windows did not contain any long (>500 bp) resequencing gap; if the gene did not fulfill this requirement it was discarded, as were 5 kb regions displaying less than five SNPs. The number of analyzed windows for AA, YRI, EU, and AS were 209, 203, 177, and 172, respectively.

Estimates of the population recombination rate parameter $\rho$ were obtained from resequencing data with the use of the Web application MAXDIP (http://genapps.uchicago.edu/maxdip/) [Hudson, 2001] and converted to cM/Mb.

The association between *THBS4* expression or brain volumes with genotype was analyzed by general linear models with continuous and factorial variables. Similarly, the effect of age on *THBS4* expression was evaluated by linear models, as specified in the text. As suggested by other authors [Lu et al., 2001; Somel et al., 2011; Shupe et al., 2006] age was $\log_2$ transformed to assure accurate modeling if the rate of change in expression decreases with increasing age. Analyses were implemented in the R environment (www.r-project.org).

### Haplotype Analysis and TMRCA Calculation

The median-joining network to infer haplotype genealogy was constructed using NETWORK 4.5 [Bandelt et al., 1999]. Estimate of the time to the most common ancestor (TMRCA) was obtained using a phylogeny-based approach implemented in NETWORK 4.5 using a mutation rate based on the number of fixed differences between chimpanzee and humans. An additional TMRCA estimate derived from application of a maximum-likelihood coalescent method implemented in GENETREE [Griffiths and Tavare, 1995; 1994]. Again, the mutation rate $\mu$ was obtained on the basis of the divergence between human and chimpanzee and under the assumption both that the species separation occurred 6 MY ago [Glazko and Nei, 2003] and of a generation time of 25 years. The migration matrix was derived from previous estimated migration rates [Schaffner et al., 2005]. A maximum-likelihood estimate of $\theta$ equal to 7.4 was obtained, resulting in an effective population size of 16,818, a value comparable to most figures reported in the literature [Tishkoff and Verrelli, 2003]. With these assumptions, the coalescence time, scaled in $2N_e$ units, was converted into years. A third TMRCA estimate was obtained by applying a previously described method [Evans et al., 2005] that calculates the average pairwise difference between all chromosomes and the MRCA: this value was converted into years on the basis of mutation rate retrieved as above.

### Brain Expression Analysis

A total of 18 human frontal cortex tissue samples obtained postmortem from different sources were used [Caceres et al., 2007] (Supp. Table S2). Total RNA was isolated from approximately 100 mg of frozen tissue by homogenization with TRIzol (Invitrogen, Monza, Italy) and was purified with the RNeasy kit (Qiagen, Milano, Italy). Complementary DNA (cDNA) was synthesized by reverse transcription of approximately 1 $\mu$g of DNase I-treated RNA from each sample with the SuperScript First Strand Synthesis kit (Invitrogen). To determine the rs438042:A>T genotype of each individual, a 478 bp fragment encompassing *THBS4* exons 3–6 was PCR amplified and the two strands were sequenced directly with the same primers used for the amplification. Total or allele-specific *THBS4* mRNA expression levels of each individual were measured in triplicate by real-time RT-PCR using the housekeeping gene $\beta$-actin to control for differences in initial cDNA, as previously described [Caceres et al., 2007]. Two different real-time quantification experiments were carried out from independently synthesized cDNAs from the same RNAs using the iTaq SYBR Green Supermix with Rox (BioRad, Segrate, Italy) and the ABI Prism 7900HD Sequence Detection System (Applied Biosystems) or the LightCycler 480 SYBR Green I Master (Roche, Basel, Switzerland) and the LightCycler 480 Real-Time PCR System (Roche). Results were analyzed with the manufacturer recommended software and protocol, and the amplification levels of *THBS4* were normalized by dividing them by the $\beta$-actin amplification level for each sample. The number of cDNA molecules was calculated by comparison with a standard curve of known amounts of the corresponding PCR products, which were quantified with Qubit dsDNA HS assay (Invitrogen).

Expression levels in one sample were consistently threefold to sixfold higher than other individuals and was eliminated from further analysis. Primers used for expression analysis are reported in Supp. Table S3.

Brian expression data from previous large-scale analyses [Somel et al., 2011; Berchtold et al., 2008; Colantuoni et al., 2011; Myers et al., 2007] were downloaded and analyzed using general linear models, as specified in the text. RNA sequencing data from human, chimpanzee, and macaque brains from Somel et al. (2011) were obtained from this work without further elaboration.

### Transcript Analysis in Lymphoblastoid Cell Lines

To verify whether the rs438042:A>T variant, located near the exon 3 donor splice site, causes aberrant splicing events, we performed RT-PCR analysis. Total RNA from 20 lymphoblastoid cell lines from HapMap individuals carrying different genotypes for rs438042 was extracted with TRIzol reagent (Invitrogen), following the manufacturer's protocols. The purified RNA (1 $\mu$g) was reverse transcribed using random hexamers and Ready-To-Go® You-Prime First-Strand Beads (GE Healthcare, Milano, Italy). RT-PCR was performed with high fidelity polymerase (Pfu DNA Polymerase; Promega, Milano, Italy) and forward and reverse primers located in exon 2 and 6. The PCR products were tested on 2% agarose gel. Quantitative RT-PCR reactions were carried out using specific TaqMan® Gene Expression Assays on an ABI 7900 Real-Time PCR System (Applied Biosystems). Reactions were performed in 25 $\mu$l reaction mixture volumes with cycling conditions set to the recommended protocols of the manufacturer, and using two TaqMan® Gene Expression Assays: Hs00170261_m1 that recognizes *THBS4* transcripts and Hs99999903_m1 that detects $\beta$-actin. Triplicate single-round reactions were carried out for each sample.

### AD Patients

Eighty-one patients (45 females and 36 males) with a clinical diagnosis of AD according to the Recommendations from the National Institute on Aging-Alzheimer Association workgroups on diagnostic guidelines for AD [McKhann et al., 2011] were recruited for this study at the Neurology Department of the Don Gnocchi Foundation in Milano, Italy. The mean age of AD patients was 76.0 years (age range 65–89 years). All patients underwent complete medical and neurological evaluation, laboratory analysis, CT scan or MRI, and other investigations—when necessary (e.g., EEG, SPET scan, CSF examination, etc.)—to exclude reversible causes of dementia. All patients were Italian of EU origin.

### Conventional MRI Acquisition and Brain Volumes Computation

AD patients were investigated using conventional MRI with a 1.5 Tesla system (Avanto, Siemens, Erlangen, Germany). In a single session, volume measurements on 3D-T1-weighted MR images (acquisition parameters were as follows: TR 1,900 msec; TE 3.37 msec, number of slices 176, slice thickness 1 mm with no gap, in-plane resolution 1 × 1 mm$^2$) were obtained from all the patients without moving them from the scanner. Routine $T_2$-weighted MRI and FLAIR were performed to rule out vascular lesions as contributory factor to memory loss or cognitive decline. Brain tissue volume, normalized for subject head size, was estimated on 3D-$T_1$-weighted MR images, with SIENAX [Smith et al., 2001; Smith et al., 2002; Smith, 2002], which is part of FSL (www.fmrib.ox.ac.uk/fsl; [Smith et al., 2004]). This method has been described in detail elsewhere [Smith et al., 2001; 2002; Smith, 2002]. Briefly, in SIENAX, brain and skull images are extracted, and the brain image is affine-registered to a MNI standard template using the skull image to determine registration scaling, to be used as a normalization for head size. Next, tissue-type segmentation with partial volume estimation is carried out [Zhang et al., 2001] to calculate total volume of brain tissue (normalized brain volume—NBV), also including separate estimates of volumes of gray matter (GM), white matter (WM), peripheral gray matter (pGM), and ventricular cerebral spinal fluid (vCSF). Intracranial volume was calculated by adding the volumes of CSF, total GM, and total WM together.

The present study conforms to the principles outlined in the Declaration of Helsinki. The study protocol was approved by the Institutional Review Board of the Fondazione Don Gnocchi, Milano. All the subjects, or their relatives if so required, provided written informed consent before admission to the study.

## Results

### Population Genetic Differentiation and Nucleotide Diversity

The *THBS4* gene has been almost fully resequenced within the SeattleSNPs Variation Discovery Resource gene panel (http://pga.gs.washington.edu/) in 24 AA and 23 EU subjects, with only small sequencing gaps scattered along the introns. We exploited the availability of these data to perform a sliding window analysis of population genetic differentiation along *THBS4*, measured as $F_{ST}$ [Wright, 1950]. Under the assumption of neutrality, $F_{ST}$ is determined by demographic history (i.e., genetic drift and gene flow), which affects all loci similarly. Therefore, as a reference, we calculated the 2.5th percentile in the distribution of $F_{ST}$ values obtained for sliding windows across all resequenced SeattleSNPs genes (see *Materials and Methods* for details). As shown in Figure 1A, an extended region of the *THBS4* gene displays unusually low $F_{ST}$ values, ranking below the 2.5th percentile. The region spans roughly 8.4 kb and encompasses exons 3–5 (chr5:79,382,598–79,390,997, hg18 Assembly). To gain further information, this region was fully resequenced in two additional populations, namely YRI and East ASs. We used these data to calculate $F_{ST}$ among the three major nonadmixed ethnic groups (i.e., YRI, EU, and AS) and obtained a value of 0.035; this corresponds to a percentile of 0.04 in the distribution of $F_{ST}$ calculated among the same populations for all genes resequenced by the NIEHS SNPs Program.

Nucleotide diversity in this region was next assessed using two indexes: $\theta_W$ [Watterson, 1975], an estimate of the expected per site heterozygosity, and $\pi$ [Nei and Li, 1979], the average number of pairwise sequence nucleotide differences. Again, as a control for demographic effects, both indexes were calculated for 5 kb windows deriving from 238 NIEHS genes. As reported in Table 1, the *THBS4* study region displays extremely high nucleotide diversity levels in all analyzed populations. Under neutral evolution, the amount of within-species diversity is predicted to correlate with levels of between-species divergence, as both depend on the neutral mutation rate [Kimura, 1983]. The HKA test [Hudson et al., 1987] is commonly used to verify this expectation. We applied a MLHKA test [Wright and Charlesworth, 2004] by comparing polymorphism and divergence levels at the *THBS4* study

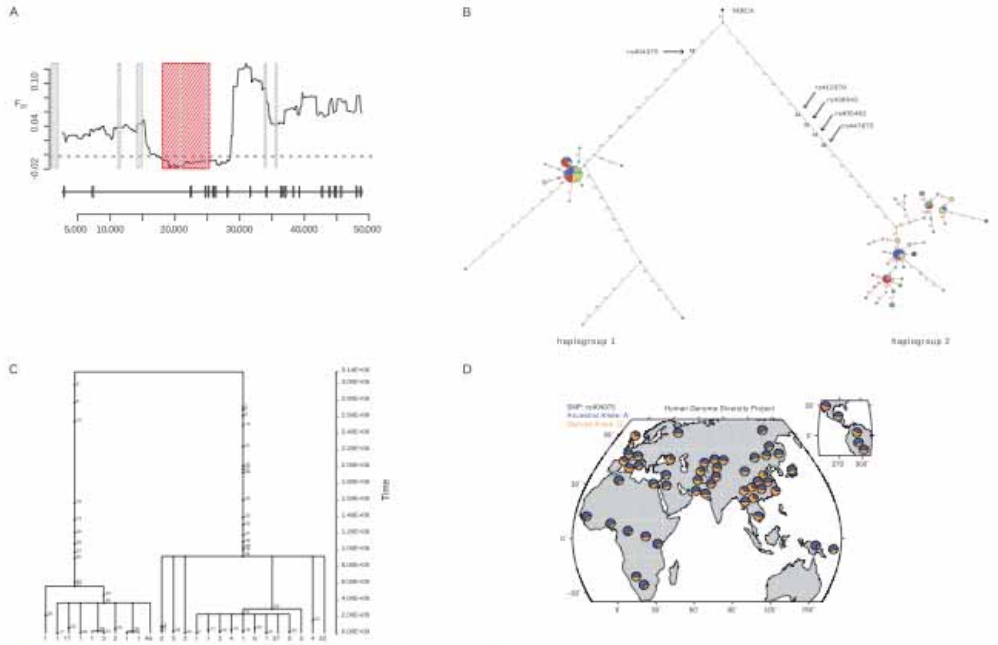



**Figure 1.** **A**: Sliding window analysis of $F_{ST}$ along the *THBS4* gene (NM_003248.4). The exon–intron gene structure is shown. The red shading identifies the region we analyzed (encompassing exons 3–5). Gray shadings indicate resequencing gaps. We used windows of 5 kb sliding with a step of 150 bp. The hatched line represents the 2.5th percentile in the distribution of $F_{ST}$ values calculated for SeattleSNPs genes. **B**: Haplotype genealogy for the *THBS4* gene region we analyzed. The genealogy was reconstructed through a median-joining network and each node represents a different haplotype, with the size of the circle proportional to the haplotype frequency. Nucleotide differences between haplotypes are indicated on the branches of the network. Circles are color-coded according to population (yellow: AA, blue: EU, red: AS, green: YRI). The most recent common ancestor (MRCA) is also shown (black circle). The relative position of mutations along a branch is arbitrary. **C**: GENETREE analysis. Mutations are represented as black dots and named for their physical position along the regions. The absolute frequency of each haplotype is also reported. **D**: Worldwide allele frequency distribution for rs404375:A>G. Each pie represents one HGDP-CEPH population. The ancestral A allele is shown in blue, the derived G allele in yellow. The image was generated using data from the Human Genome Diversity Project [Cann et al., 2002; Li et al., 2008].

**Table 1. Nucleotide Diversity, Neutrality Tests, and MLHKA Test for the Analyzed *THBS4* Region**

| | | | $\Pi$ ($\times 10^{-4}$) | | $\Theta_W$ ($\times 10^{-4}$) | | Tajima's *D* | | Fu and Li's *D** | | Fu and Li's *F** | | MLHKA test | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Population | $N^a$ | $S^b$ | Value | Rank[c] | Value | Rank[c] | Value (*P*)[d] | Rank[c] | Value (*P*)[d] | Rank[c] | Value (*P*)[d] | Rank[c] | $k^e$ | *P* |
| YRI | 44 | 50 | 23.19 | 0.99 | 14.13 | 0.89 | 2.22 (<0.001) | >0.99 | 1.61 (<0.001) | >0.99 | 2.17 (<0.001) | >0.99 | 2.75 | 0.0084 |
| AA | 48 | 51 | 24.31 | 0.99 | 14.20 | 0.88 | 2.47 (<0.001) | >0.99 | 0.96 (0.018) | 0.97 | 1.80 (<0.001) | >0.99 | 2.43 | 0.0093 |
| EU | 46 | 45 | 23.72 | 0.99 | 12.64 | 0.95 | 3.03 (<0.001) | >0.99 | 1.56 (0.006) | >0.99 | 2.48 (<0.001) | >0.99 | 3.08 | 0.0027 |
| AS | 40 | 44 | 24.71 | 0.98 | 12.78 | 0.97 | 3.30 (<0.001) | >0.99 | 1.53 (0.009) | 0.98 | 2.54 (<0.001) | >0.99 | 3.54 | 0.0005 |

[a] Sample size (chromosomes).
[b] Number of segregating sites.
[c] Percentile rank relative to a distribution of 238 5 kb windows from NIEHS genes.
[d] *P* value obtained by coalescent simulations.
[e] Selection parameter.

region with 16 NIEHS genes resequenced in all populations (see *Materials and Methods*). Results, summarized in Table 1, indicate that a significant excess of nucleotide diversity versus interspecies divergence is detectable in all populations for the *THBS4* study region.

It has recently been shown that biased gene conversion (BGC) affects neutral substitution patterns [reviewed in Duret and Galtier (2009)] with particularly strong effect in subtelomeric regions and in regions with high male-specific recombination rates [Dreszer et al., 2007; Duret and Arndt, 2008; Webster et al., 2005]. *THBS4* is not subtelomeric and male-specific recombination rate (as derived from the deCODE data for male recombination rate) is very low; the same applies to recombination rate calculated from population genetic data in the region we analyzed (0.15 cM/Mb). Moreover, analysis of polymorphic positions in the region indicated that only 18 out of 65 are A/T → G/C mutations. Thus, these data suggest that

BGC does not play a major role in shaping nucleotide variability at *THBS4*.

### Neutrality Tests and Haplotype Analysis

We verified whether the neutral model could be rejected for the *THBS4* gene region encompassing exons 3–5. Tajima's $D$ [Tajima, 1989] and Fu and Li's $D^*$ and $F^*$ [Fu and Li, 1993] are commonly used statistics that evaluate departures of the allelic frequency distributions from the expected patterns of neutral variation by comparing different estimates of nucleotide variation. Positive values of Tajima's $D$ and of Fu and Li's $D^*$ and $F^*$ indicate an excess of intermediate frequency variants and are a hallmark of balancing selection. Because population history, in addition to selective processes, is known to affect frequency spectra and all related statistics, we performed coalescent simulations using a calibrated population genetics model that incorporates demographic scenarios [Schaffner et al., 2005]. Also, we calculated test statistics for 5 kb windows deriving from NIEHS genes. The null hypothesis of neutrality is rejected by all statistics in all the populations (Table 1). Consistently, most values calculated for the *THBS4* study region display values higher than the 99th percentile in the distribution of 5 kb reference windows (Table 1).

To examine the genealogy of *THBS4* haplotypes, we built a median-joining network. The topology (Fig. 1B) revealed a few recurrent mutations/gene conversion events. Nonetheless, two major clades (haplogroups 1 and 2) separated by long-branch lengths are evident, each containing common haplotypes. To estimate the TMRCA of the two haplotype clades, we applied a phylogeny-based method [Bandelt et al., 1999] and obtained a TMRCA of 3.71 MY (SD: 573 KY) (see *Materials and Methods*). To obtain a second estimate, we used GENETREE [Griffiths and Tavare, 1995]: the TMRCA amounted to 3.14 MY (SD: 288 KY) (Fig. 1C). A third TMRCA estimate of 3.98 MY was obtained by applying a previously described method [Evans et al., 2005] (see *Materials and Methods*). As expected from the estimated TMRCAs, analysis of polymorphic positions in chimpanzees [Auton et al., 2012; Hvilsom et al., 2012] indicated that no variant is shared with humans (i.e., no *trans*-specific polymorphisms are observed).

Analysis of SNPs along the major branches of the phylogeny (Fig. 1B) indicated that rs404375:A>G has been genotyped in the Human Genome Diversity Cell Line Panel: as shown in Figure 1D, this variant has intermediate frequency in all populations.

### Prediction of Functional Differences

The results reported above strongly suggest that the *THBS4* region we analyzed represents a balancing selection target and, therefore, that it carries one or more variants with a functional effect. Only SNP rs438042:A>T (on the major branch of the haplotype phylogeny, Fig. 1B), which is located at the end of exon 3, affects the *THBS4* coding region, although the two alleles code for the same amino acid. This SNP, as well as rs405482:A>G and rs447875:G>A (Fig. 2A), is very close to the exon's donor splice site, within the region that pairs with the U1 snRNA. Interestingly, the G allele for rs447875:G>A (hap1) is widely conserved in mammals and coincides with a position of perfect complementarity to the U1 snRNA [Roca et al., 2008] (Fig. 2A). Therefore, the G>A change might be important for splicing. To clarify this issue, we analyzed *THBS4* transcripts from 20 lymphoblastoid cell lines from HapMap individuals carrying different genotypes for rs438042:A>T (four hap1 homozygotes, seven heterozygotes, nine hap2 homozygotes) and from 11 brain frontal cortex samples (three hap1 homozygotes, six heterozygotes, two hap2 homozygotes). Transcript analysis revealed no product corresponding to the skipping of exon 3 (not shown). Nonetheless, it should be noted that exclusion of exon 3 from the transcript would result in a shift of the reading frame, suggesting that the resulting transcript encoding a truncated protein might undergo nonsense-mediated decay (NMD).

We also analyzed the balancing selection region for the presence of possible modulators of gene expression. Analysis of chromatin state [Ernst and Kellis, 2010] along the *THBS4* gene region indicated that an enhancer is predicted in a region encompassing exon 3 (Fig. 2B). Consistently, the region shows histone marks (H3K4Me1) associated with transcriptional enhancers (Fig. 2B), suggesting that the selection target(s) might be represented by one or more variants that regulate *THBS4* expression.

### *THBS4* Expression Analysis

To investigate whether variants in the balancing selection region modulate *THBS4* transcript levels, we first exploited available RNA sequencing data [Montgomery et al., 2010] to analyze expression in HapMap lymphoblastoid cell lines from EU subjects carrying different genotype for rs412379:G>A; this variant is located on the branch separating the two major haplotype clades (variant 32 in Fig. 1B) and falls within the predicted transcriptional enhancer (Fig. 2B). By fitting a general linear model that included sex, we observed a significant effect of *THBS4* genotype on expression ($F = 3.38$, $P = 0.041$) with hap1 conferring higher expression (Fig. 2C). No sex-specific effect was noted. We confirmed this observation by real-time PCR on hap1 ($n = 7$) and hap2 ($n = 4$) homozygous lymphoblastoid lines ($F = 10.43$, $P = 0.013$). A similar result was obtained by Zeller et al. (2010) in an analysis of expression quantitative trait loci in human monocytes: in these cells, hap1 confers significantly higher *THBS4* expression (ANOVA, $P = 9.3 \times 10^{-16}$) (Fig. 2D).

We next examined expression of *THBS4* mRNA by real-time PCR in the postmortem frontal cortex samples of 17 adults (age range: 30–87 years) genotyped for exon 3 rs438042:A>T, which is located on the branch separating the two major haplotype clades (Fig. 1B) and in linkage disequilibrium with rs412379:G>A (Supp. Fig. S1). Two independent expression experiments were performed to assure reliability (see *Materials and Methods*). To analyze the effect of age, genotype, and sex on *THBS4* expression, we applied a mixed model ANOVA (between-within design to account for repeated measures for expression). A significant effect of age on *THBS4* expression was evident ($F = 10.474$, $P = 0.007$) with increased expression at older ages, whereas the differences between experiments were not significant ($F = 1.55$, $P = 0.232$), nor were the effects of sex and genotype. To explore this further, we checked allele-specific expression levels in heterozygotes. Allele-specific real-time PCRs were performed for the same human samples, and, supporting the similar levels of *THBS4* expression among the different genotypes, similar expression levels of each allele were also quantified in A/T heterozygotes.

To obtain independent confirmation for the observed increase in *THBS4* expression with age, and to test whether this effect is specific to the brain, we retrieved gene expression data from large-scale studies that analyzed gene expression changes in different cerebral areas [Somel et al., 2011; Berchtold et al., 2008; Colantuoni et al., 2011] or in nonneural tissues [Welle et al., 2003; Rodwell et al., 2004]. The effect of age on *THBS4* expression was evaluated by fitting linear models that included sex. Colantuoni et al. (2011) analyzed 269 subjects in a wide age range and with different ethnic origin. A strong effect of postnatal age on *THBS4* levels was noted

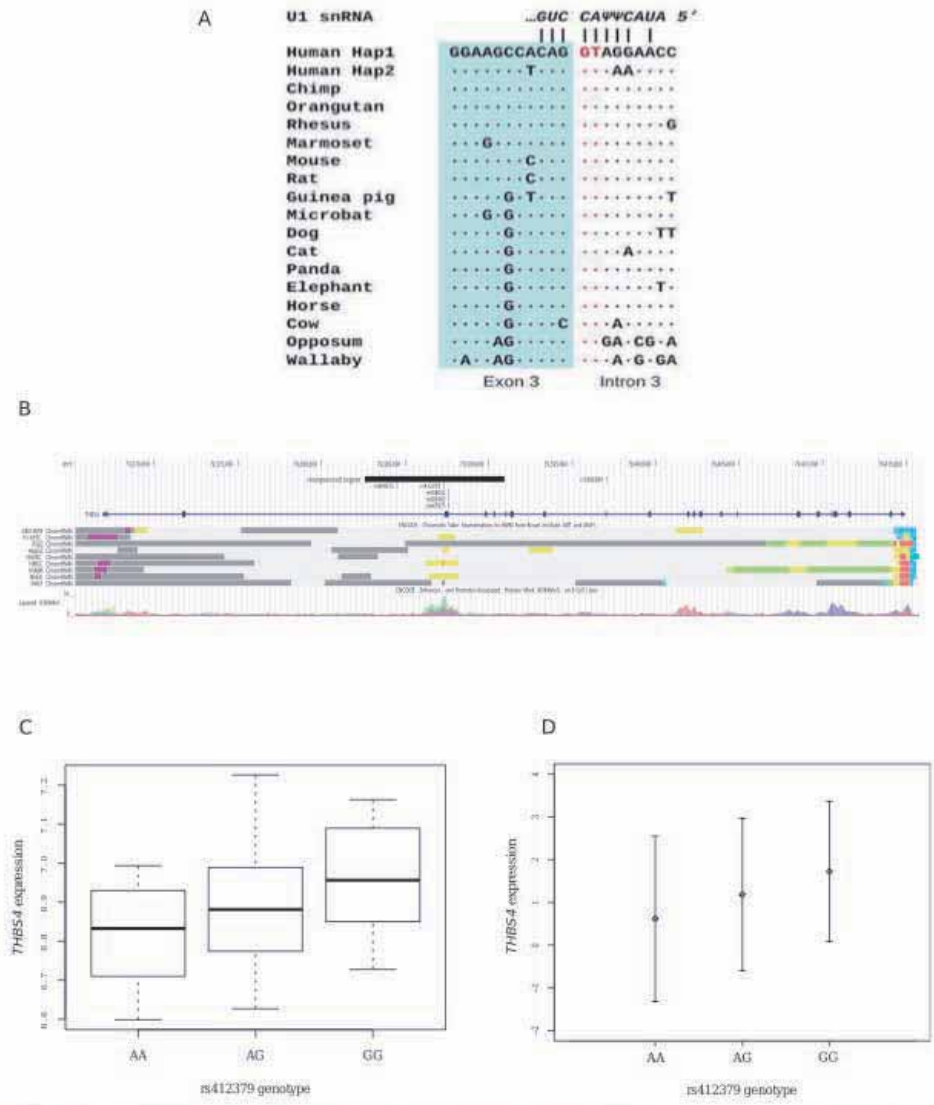



**Figure 2.** **A**: Analysis of the 5′ splice site of *THBS4* exon 3. Multiple sequence alignment of the two main human haplotypes (showing variable positions rs438042:A>T, rs405482:A>G, and rs447875:G>A) with multiple mammals. The first two invariable bases of the intron are shown in red. Base pairing with the 5′ end of the U1 snRNA is shown on top. Ψ represents pseudo-uridine, which is a modified uridine nucleotide capable of base pairing to both A and G nucleotides. Multiple alignment was done with MUSCLE [Edgar, 2004] and other species THBS4 sequences were obtained from the genome reference sequences (chimpanzee, panTro3; orangutan, ponAbe2; rhesus, rheMac2; marmoset, calJac3; mouse, mm9; rat, rn4; guinea pig, cavPor3; microbat, myoLuc2; dog, canFam2; cat, felCat4; panda, ailMel1; elephant, loxAfr3; horse, equCab2; cow, bosTau6; opossum, monDom5; wallaby, macEug2). **B**: Chromatin state segmentation and H3K4Me1 histone marks for nine different human cell types derived from the UCSC Genome Browser track. Different chromatin states are colored to highlight predicted functional elements. Purple: inactive/poised promoter; orange: strong enhancer; yellow: weak/poised enhancer; blue: insulator; light green: weak transcribed; gray: polycomb-repressed; light gray: heterochromatin, low signal. The region we resequenced is shown in black. **C**: Analysis of *THBS4* expression in lymphoblastoid cell lines deriving from 60 HapMap individuals of EU ancestry carrying different genotypes at rs412379:G>A (the G and A alleles tag hap1 and 2, respectively). Data were obtained from a previous work [Montgomery et al., 2010]. **D**: Analysis of *THBS4* expression in monocytes from 1,490 individuals recruited in Germany as a function of rs412379:G>A genotype (the G and A alleles tag hap1 and 2, respectively). Data were obtained from a previous work [Zeller et al., 2010].

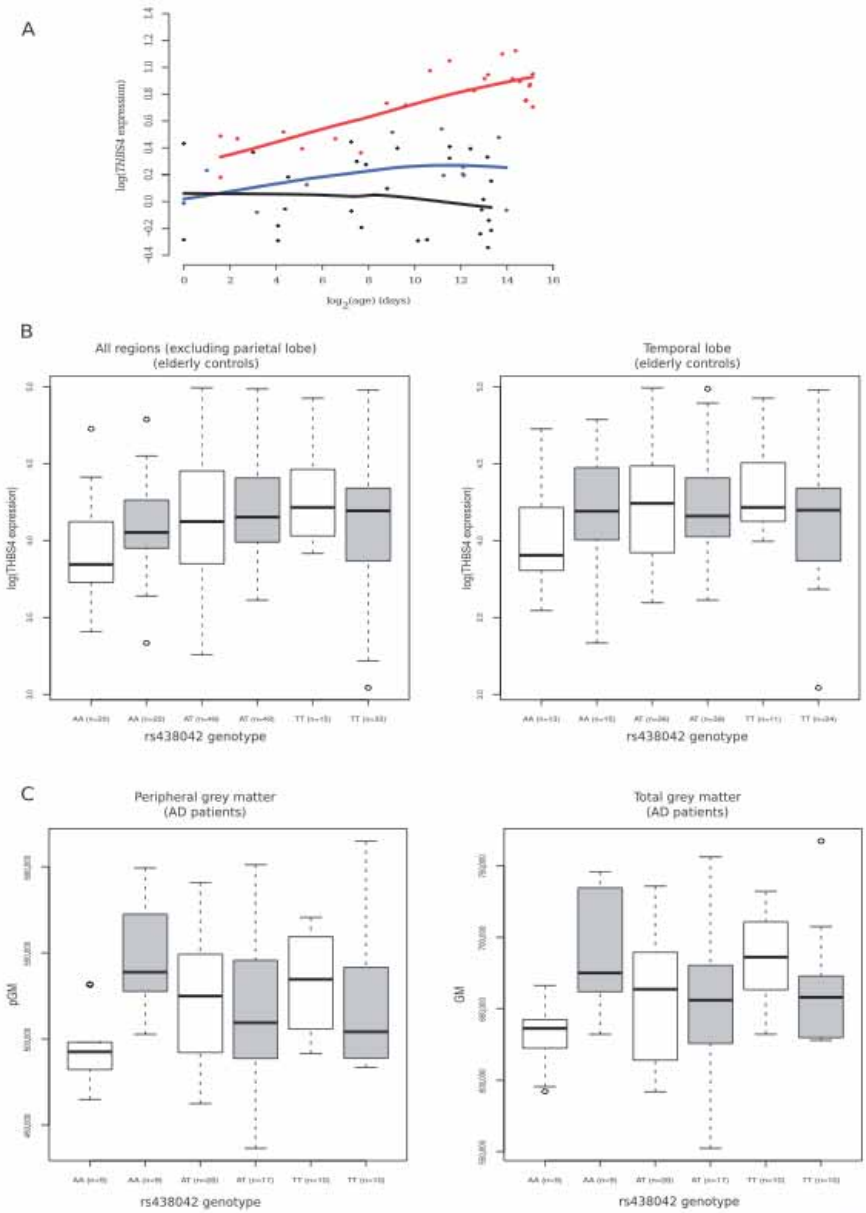



**Figure 3.** **A**: Expression profile of *THBS4* in the prefrontal cortex at different age points. Data are derived from Somel et al. (2011) and refer to humans (red), chimpanzees (blue), and macaques (black). Each point represents an individual, lines show lowest fittings [Cleveland, 1981]. **B**: Box-and-whisker plot of *THBS4* expression level in brain samples from elderly healthy subjects [Myers et al., 2007]. The expression level of *THBS4* quantified by DNA microarray (log transformed) is plotted as a function of sex (white, females; gray, males) and rs438042:A>T genotype (the A and T alleles tag hap1 and 2, respectively). Data are shown for all brain regions with the exclusion of three samples from parietal lobes or for specimens from temporal lobes only. **C**: Box-and-whisker plot of brain volumes in 81 AD patients. Volumes of peripheral gray matter and total gray matter are plotted as a function of sex (white, females; gray, males) and rs438042:A>T genotype (the A and T alleles tag hap1 and 2, respectively). Mean age per group is as follows: AA females, 76.2; AA males, 76.9; AT females, 77.3; AT males, 75.5; TT females, 73.7; TT males 74.6.

in both AAs ($n = 147$, $\beta = 0.19$, $P = 2.3 \times 10^{-10}$) and in subjects of EU origin ($n = 112$, $\beta = 0.22$, $P = 2 \times 10^{-16}$). Results suggesting increased *THBS4* expression with age were also obtained using microarray and RNA sequencing data from Somel et al. (2011); RNA sequencing experiments indicated that in human prefrontal cortex *THBS4* expression increases during life (expression: newborn, −0.434; young adult, 0.677; old adult 2.02); this trend is not observed in chimpanzee (expression: newborn, −0.58; young adult, −0.52) and macaque (expression: newborn, −0.63; young adult, −0.53) or in the cerebellar cortex of any species (not shown). This trend is also evident from the microarray data: in line with previous work [Caceres et al., 2007], at any age *THBS4* expression is higher in human brain cortex than in chimpanzee and macaque. Moreover, a clear increase with age ($\beta = 0.08$, $P = 7.6 \times 10^{-6}$) is evident in human brain but not in the other two species (Fig. 3A). Conversely, analysis of data from a third work [Berchtold et al., 2008] revealed no changes in *THBS4* expression with age in any brain area analyzed (entorhinal cortex, hippocampus, postcentral gyrus, and superior frontal gyrus). Sex-specific differences were not detected in any study. Finally, we analyzed *THBS4* expression levels at different ages in muscle [Welle et al., 2003] and kidney [Rodwell et al., 2004]. No expression change was observed in the 15 muscle samples analyzed by Welle et al. (2003) ($\beta = -0.018$, $P = 0.89$), whereas in kidney specimens, a trend toward decreasing expression at older ages was noted (the model included ethnicity as an additional factor; for the medulla, $\beta = -0.56$, $P = 0.001$; for the renal cortex, $\beta = -0.09$, $P = 0.52$). Thus, most studies support the observation that *THBS4* expression increases with age in the human brain, but not in other tissues.

Finally, we verified whether *THBS4* variants located within the balancing selection region influence the brain expression level of the gene in a larger sample. We thus retrieved genotyping and gene expression data from a previous whole-genome analysis on 193 human brains [Myers et al., 2007]. The genotype status at rs438042:A>T was imputed from available data of rs412379:G>A (variant 32 in the haplotype phylogeny, Fig. 1B), which is in full LD with the previous SNP (Supp. Fig. S1), and *THBS4* expression level was estimated from probe GI_40549419-S. These brain samples were derived from healthy subjects with an average age of 81 years (SD: 9.18), and were collected in different brain areas. Analysis of *THBS4* expression in these specimens indicated that the three samples from parietal lobes displayed a much higher expression of *THBS4* (Supp. Fig. S2); these samples were therefore removed from further analyses. The effect of rs438042:A>T genotype on *THBS4* expression was analyzed using a general linear model that included age, sex, brain region, and postmortem interval as independent factors. A weak effect of genotype on expression was detected ($P = 0.057$); notably though, a significant interaction between sex and genotype ($P = 0.038$), as well as a major effect of the sampled brain region ($P = 0.0031$) were detected (Supp. Table S4). The interaction effect between sex and genotype was mainly mediated by AA homozygous females, in whom the lowest expression level among all sex–genotype combinations was observed (Fig. 3B). Similar results were obtained when analysis was restricted to specimens sampled from temporal lobes ($n = 140$). Again, a significant interaction was observed between rs438042:A>T genotype and sex ($P = 0.025$, Supp. Table S4), with samples from AA homozygous women showing lower *THBS4* expression (Fig. 3B).

### *THBS4* in AD

Overall, the results above strongly suggest that *THBS4* expression in human brain increases with age, and that in elderly healthy individuals transcript levels are modulated by different factors, including an interaction between sex and the genotype status at variants located in the balancing selection region. These results and *THBS4* accumulation in $\beta$-amyloid plaques [Caceres et al., 2007] led us to investigate the role of *THBS4* variants in AD. Thus, the following measures were obtained for 81 AD patients: total volume of brain tissue (NBV), volumes of GM and WM, pGM, and vCSF (see *Materials and Methods* for details). These patients were genotyped for rs438042:A>T and *APOE* alleles. The effect of *THBS4* genotype on MRI measures was evaluated by fitting a general linear model including sex, age, and presence/absence of the *APOE ε4* allele. The main effect of genotype was not significant for any of the analyzed volumes, whereas, as expected, a significant effect of age was observed for most measures (Supp. Table S5). A significant interaction between sex and genotype was observed both for pGM ($P = 0.014$) and GM ($P = 0.012$) (Supp. Table S5). Also in this case, the interaction was mainly mediated by female AD patients with an AA genotype. Thus, although being on average the same age (mean age = 76.2) as male patients with the same genotype (mean age = 76.8), markedly reduced pGM and GM volumes were seen in these females (Fig. 3C). Notably, AA homozygous AD females display the lowest volumes of GM and pGM compared with all sex–genotype combinations (Fig. 3C). These results parallel those obtained from the analysis of *THBS4* brain expression levels (Fig. 3B).

Finally, all AD subjects were genotyped for rs1866389:G>C (p.Ala387Pro) and no effect of this variants on brain volumes was noted, either per se or in interaction with other factors (not shown).

## Discussion

Gene regulatory elements represent potential targets of adaptive evolution, and several studies have indicated that *cis*-regulators have been more frequently subject to positive selection compared with *trans*-acting elements [reviewed in Lappalainen and Dermitzakis, (2010)]. Thus, we made use of resequencing data to analyze genetic diversity across *THBS4*, a gene whose expression has been strongly upregulated in the brain during human evolution [Caceres et al., 2007; 2003]. Data herein indicate that a region centered around exon 3 has been a target of balancing selection. Analysis of haplotype genealogy indicated the presence of two major clades with *TMRCA* estimates in the range of 3–4 MY. Such deep coalescence times are extremely rare in the human genome [Tishkoff and Verrelli, 2003; Garrigan and Hammer, 2006] and are not consistent with evolutionary neutrality. Nonetheless, an alternative explanation for the deep TMRCA and high nucleotide diversity we observed may in principle be the result of demographic effects, namely archaic admixture. A hallmark of ancient population structure is the presence of highly differentiated haplotypes with very little evidence of recombination between lineages [Wall, 2000], as observed for the *THBS4* study region. Yet, several evidences suggest that the signature we observed at *THBS4* is not merely the result of archaic introgression: first, unlike *THBS4*, most previously described introgressed neutral alleles are specific or almost specific to one single population [Garrigan et al., 2005; Barreiro et al., 2005; Cox et al., 2008; Hardy et al., 2005]; second, a significant skew toward intermediate frequency alleles is not consistent under a model of archaic admixture, as a neutral introgressed allele is expected to segregate at low frequency in modern populations [Hawks et al., 2008]; third, the low $F_{ST}$ we observed in the region is not explained by archaic admixture. Nonetheless, introgression and selection are not mutually exclusive, as a haplotype introduced by admixture may have higher chances to be detected in modern populations when subjected to

a selective event (that opposes its chances of being lost by drift). This has been shown to be the case for other genes involved in brain development, such as *MCPH1* and *ASAH1* [Evans et al., 2006; Kim and Satta, 2008].

Finally, with respect to the limited evidence of recombination in the region, this might also be explained by the fact that signatures of long-standing balancing selection are eroded over time by both mutation and recombination. Thus, Bubb et al. (2006) suggested that, to be detectable, long-standing balancing selection must involve at least two physically linked loci that are both selection targets (i.e., a balanced haplotype). This would allow accumulation of neutral variability over longer regions because of the selection against most recombination events in the interval. Therefore, one possible explanation for this observation is the presence of more than one selection target in the region we analyzed, although the precise location of these variants remains to be determined.

Analysis of SNPs located along the major branches of the haplotype genealogy, where the selected variant(s) is expected to be located, indicated that three polymorphisms possibly affect positions relevant for splicing. In lymphoblastoid cell lines, we were unable to detect any transcript lacking exon 3, although these might be degraded through NMD. In line with this possibility, expression of *THBS4* in these cells was found to be significantly associated with the presence of hap1, which carries the ancestral conserved splice site variants. However, another possibility is that the balancing selection region carries elements important for gene transcription, as the chromatin profiling data suggest the presence, in different cell types, of a transcriptional enhancer in the region encompassing exon 3. Thus, one or more polymorphisms in the region might affect tissue-specific (and sex-specific) transcription regulatory elements. In fact, in a relatively large sample of human brain specimens from elderly donors, we detected an interaction between sex and *THBS4* genotype in modulating expression levels, an effect mainly mediated by hap1 homozygous females, who showed reduced expression. These results were not recapitulated in lymphoblastoid cell lines, where hap1 conferred higher expression without any detectable effect of gender. The discrepancy might be due to different reasons: the sex-specific effect may be secondary to hormonal differences (which are not reproduced in cultured cells) or may be tissue specific (e.g., may derive from differential binding by a tissue-specific *trans*-acting factor). Indeed, balancing selection at regulatory variants has previously been proposed to result from the ability of distinct alleles/haplotypes to confer preferential expression in a tissue- or cell-type-specificity manner, or to modulate transcription in response to distinct stimuli [Loisel et al., 2006; Cagliani et al., 2008]. Also, we detected no interaction between sex and genotype in the brain sample we analyzed by real-time PCR, possibly because of a lack of power due to the small size and to the confounding effect of age (e.g., the sample included one single hap1 homozygous female aged 74).

Sex-specific differences in gene expression are common in the human brain [Reinius et al., 2008]. Notably, a recent study indicated that noncoding variants in three out of four analyzed genes display sex-specific effects on brain morphometric traits [Rimol et al., 2010], suggesting that sex-specific differences in gene expression might influence common variance in brain structure. Interestingly, results herein indicate a sex-specific recessive effect of the A allele of rs438042:A>T (hap1) in conferring both lower *THBS4* expression and significantly reduced total GM and pGM in AD subjects.

Although these observations and the increased expression of *THBS4* with age might suggest that the protein has a neuroprotective function, further analyses will be required to validate this possibility. On the one hand, despite the observation whereby the upregulation of genes involved in immune response and inflammation represents a general feature of aging [de Magalhaes et al., 2009], analysis of gene expression in muscle and kidney revealed no increase in *THBS4* expression during lifetime, suggesting that this effect is specific to the brain and might represent a compensatory mechanism of neuroprotection [de Magalhaes et al., 2009]. On the other hand, Liang et al. (2008, 2010) described a significant reduction of *THBS4* expression in nondemented individuals with intermediate AD (NDAD) versus controls in brain areas severely affected in AD (entorhinal cortex, fold change = −2.4; hippocampus, fold change = −2.1; superior frontal gyrus, fold change = −2.2; middle temporal gyrus, fold change = −3.1; in this latter area, *THBS4* expression was also significantly reduced in late-onset AD compared with controls). Therefore, if *THBS4* is neuroprotective, it remains to be evaluated why it is downregulated in AD brains. Nonetheless, it is worth mentioning that Liang et al. (2008, 2010) analyzed gene expression in laser-capture microdissected neurons rather than brain homogenates; therefore, their data might not recapitulate the typical expression profile changes associated with aging and neurodegeneration (e.g., upregulation of the inflammatory and immune response pathways) [de Magalhaes et al., 2009]. Overall, additional experimental analyses will be required to clarify these issues, and the role of *THBS4* on brain volumes in AD subjects should be interpreted with caution because of the relatively small sample size.

Interestingly, analysis of data from a previous work [Somel et al., 2011] suggested that the increase of *THBS4* levels with age is specific to humans, as it is not observed in chimpanzees and macaques. In the same work, the authors more generally showed that developmental changes in gene expression are a specific feature of the human brain [Somel et al., 2011], suggesting that increased developmental plasticity is evolutionary selected. In general, variants that affect phenotypes with postreproductive onset, including AD, are considered to be rarely targeted by selection. Thus, although our data suggest a role of *THBS4* variants in AD, the selective pressure responsible for maintaining variability in the gene may not relate to the development of this disease. A recent study showed that haplotypes carrying AD susceptibility alleles have undergone recent selective sweeps in human populations [Raj et al., 2012], and indicated that, as these genes have a role in immune responses, the selective pressure underlying these events might relate to pathogen exposure. Similarly, the protective *APOE ε3* allele is thought to have spread in human populations as a result of natural selection [Fullerton et al., 2000; Drenos and Kirkwood, 2010]. Still, APOE functions in different systems including immune response and lipid homeostasis that may have been targeted by selection. This might also be the case for *THBS4*, as the gene has been involved in inflammatory processes, and it regulates macrophage adhesion to endothelial surfaces and neutrophil activation [Pluskota et al., 2005; Stenina et al., 2003]. Moreover, *THBS4* induction has been observed in response to spinal injury and contributes to sensitization and neuropathic pain [Kim et al., 2012]. In this context, *THBS4* is mainly expressed by astrocytes, suggesting that the gene plays a role in multiple systems, which potentially represent targets of natural selection.

In summary, results herein indicate that in our species, but not in chimpanzee and macaque, *THBS4* brain expression increases with age. Application of different population genetic tests revealed that a region within the *THBS4* transcription unit has been a target of long standing balancing selection in human populations; this region encompasses a predicted transcriptional enhancer and polymorphisms tagging the two major haplotypes modulate *THBS4* expression in white blood cells, suggesting that the selection target(s) is accounted for by a regulatory polymorphism. Although hap1 confers higher expression in lymphocytes, this haplotype is associated with a

sex-specific decrease in gene expression in human brain, and may determine reduced grey matter volumes in AD patients. Our results warrant further investigation to determine whether the changes in *THBS4* expression profile during human lifetime are affected by variants in the balancing selection region, and to study the effect of *THBS4* polymorphisms in other tissues and cell types. Also, the role of *THBS4* in affecting brain volumes of AD patients will need replication in an independent larger sample.

## Acknowledgments

We thank the following individuals and institutions for their assistance in obtaining the human brain RNAs used in this study: Todd Preuss, Carolyn Suwyn, Isidre Ferrer, Elena Miñones-Moyano, Eulalia Martí, the Brain and Tissue Bank for Developmental Disorders at the University of Maryland, the Emory University Alzheimer Disease Research Center, and the Institute of Neuropathology Brain Bank of the University Hospital of Bellvitge. We also thank Carles Arribas, Xavier Estivill, Juan Valcárcel, and Patricia Wittkopp for technical assistance and advice on gene-expression analysis.

## References

Adams JC. 2004. Functions of the conserved thrombospondin carboxy-terminal cassette in cell-extracellular matrix interactions and signaling. Int J Biochem Cell Biol 36:1102–1114.

Adams JC, Lawler J. 2004. The thrombospondins. Int J Biochem Cell Biol 36:961–968.

Auton A, Fledel-Alon A, Pfeifer S, Venn O, Segurel L, Street T, Leffler EM, Bowden R, Aneas I, Broxholme J, Humburg P, Iqbal Z, et al. 2012. A fine-scale chimpanzee genetic map from population sequencing. Science 336:193–198.

Bandelt HJ, Forster P, Rohl A. 1999. Median-joining networks for inferring intraspecific phylogenies. Mol Biol Evol 16:37–48.

Barreiro LB, Patin E, Neyrolles O, Cann HM, Gicquel B, Quintana-Murci L. 2005. The heritage of pathogen pressures and ancient demography in the human innate-immunity CD209/CD209L region. Am J Hum Genet 77:869–886.

Berchtold NC, Cribbs DH, Coleman PD, Rogers J, Head E, Kim R, Beach T, Miller C, Troncoso J, Trojanowski JQ, Zielke HR, Cotman CW. 2008. Gene expression changes in the course of normal brain aging are sexually dimorphic. Proc Natl Acad Sci USA 105:15605–15610.

Bubb KL, Bovee D, Buckley D, Haugen E, Kibukawa M, Paddock M, Palmieri A, Subramanian S, Zhou Y, Kaul R, Green P, Olson MV. 2006. Scan of human genome reveals no new loci under ancient balancing selection. Genetics 173:2165–2177.

Caceres M, Lachuer J, Zapala MA, Redmond JC, Kudo L, Geschwind DH, Lockhart DJ, Preuss TM, Barlow C. 2003. Elevated gene expression levels distinguish human from non-human primate brains. Proc Natl Acad Sci USA 100:13030–13035.

Caceres M, Suwyn C, Maddox M, Thomas JW, Preuss TM. 2007. Increased cortical expression of two synaptogenic thrombospondins in human brain evolution. Cereb Cortex 17:2312–2321.

Cagliani R, Fumagalli M, Riva S, Pozzoli U, Comi GP, Menozzi G, Bresolin N, Sironi M. 2008. The signature of long-standing balancing selection at the human defensin beta-1 promoter. Genome Biol 9:R143.

Cann HM, de Toma C, Cazes L, Legrand MF, Morel V, Piouffre L, Bodmer J, Bodmer WF, Bonne-Tamir B, Cambon-Thomsen A, Chen Z, Chu J, et al. 2002. A human genome diversity cell line panel. Science 296:261–262.

Christopherson KS, Ullian EM, Stokes CC, Mullowney CE, Hell JW, Agah A, Lawler J, Mosher DF, Bornstein P, Barres BA. 2005. Thrombospondins are astrocyte-secreted proteins that promote CNS synaptogenesis. Cell 120:421–433.

Cleveland WS. 1981. LOWESS: a program for smoothing scatterplots by robust locally weighted regression. Am Statistic 35:54.

Colantuoni C, Lipska BK, Ye T, Hyde TM, Tao R, Leek JT, Colantuoni EA, Elkahloun AG, Herman MM, Weinberger DR, Kleinman JE. 2011. Temporal dynamics and genetic control of transcription in the human prefrontal cortex. Nature 478:519–523.

Cox MP, Mendez FL, Karafet TM, Pilkington MM, Kingan SB, Destro-Bisol G, Strassmann BI, Hammer MF. 2008. Testing for archaic hominin admixture on the X chromosome: model likelihoods for the modern human RRM2P4 region from summaries of genealogical topology under the structured coalescent. Genetics 178:427–437.

de Magalhaes JP, Curado J, Church GM. 2009. Meta-analysis of age-related gene expression profiles identifies common signatures of aging. Bioinformatics 25:875–881.

Drenos F, Kirkwood TB. 2010. Selection on alleles affecting human longevity and late-life disease: the example of apolipoprotein E. PLoS ONE 5:e10022.

Dreszer TR, Wall GD, Haussler D, Pollard KS. 2007. Biased clustered substitutions in the human genome: the footprints of male-driven biased gene conversion. Genome Res 17:1420–1430.

Duret L, Arndt PF. 2008. The impact of recombination on nucleotide substitutions in the human genome. PLoS Genet 4:e1000071.

Duret L, Galtier N. 2009. Biased gene conversion and the evolution of mammalian genomic landscapes. Annu Rev Genomics Hum Genet 10:285–311.

Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32:1792–1797.

Elzie CA, Murphy-Ullrich JE. 2004. The N-terminus of thrombospondin: The domain stands apart. Int J Biochem Cell Biol 36:1090–1101.

Ernst J, Kellis M. 2010. Discovery and characterization of chromatin states for systematic annotation of the human genome. Nat Biotechnol 28:817–825.

Eroglu C, Allen NJ, Susman MW, O'Rourke NA, Park CY, Ozkan E, Chakraborty C, Mulinyawe SB, Annis DS, Huberman AD, Green EM, Lawler J, et al. 2009. Gabapentin receptor alpha2delta-1 is a neuronal thrombospondin receptor responsible for excitatory CNS synaptogenesis. Cell 139:380–392.

Evans PD, Gilbert SL, Mekel-Bobrov N, Vallender EJ, Anderson JR, Vaez-Azizi LM, Tishkoff SA, Hudson RR, Lahn BT. 2005. Microcephalin, a gene regulating brain size, continues to evolve adaptively in humans. Science 309:1717–1720.

Evans PD, Mekel-Bobrov N, Vallender EJ, Hudson RR, Lahn BT. 2006. Evidence that the adaptive allele of the brain size gene microcephalin introgressed into Homo sapiens from an archaic Homo lineage. Proc Natl Acad Sci USA 103:18178–18183.

Fu YX, Li WH. 1993. Statistical tests of neutrality of mutations. Genetics 133:693–709.

Fullerton SM, Clark AG, Weiss KM, Nickerson DA, Taylor SL, Stengard JH, Salomaa V, Vartiainen E, Perola M, Boerwinkle E, Sing CF. 2000. Apolipoprotein E variation at the sequence haplotype level: implications for the origin and maintenance of a major human polymorphism. Am J Hum Genet 67:881–900.

Fumagalli M, Cagliani R, Pozzoli U, Riva S, Comi GP, Menozzi G, Bresolin N, Sironi M. 2009. Widespread balancing selection and pathogen-driven selection at blood group antigen genes. Genome Res 19:199–212.

Garrigan D, Hammer MF. 2006. Reconstructing human origins in the genomic era. Nat Rev Genet 7:669–680.

Garrigan D, Mobasher Z, Kingan SB, Wilder JA, Hammer MF. 2005. Deep haplotype divergence and long-range linkage disequilibrium at xp21.1 provide evidence that humans descend from a structured ancestral population. Genetics 170:1849–1856.

Glazko GV, Nei M. 2003. Estimation of divergence times for major lineages of primate species. Mol Biol Evol 20:424–434.

Griffiths RC, Tavare S. 1994. Sampling theory for neutral alleles in a varying environment. Philos Trans R Soc Lond B Biol Sci 344:403–410.

Griffiths RC, Tavare S. 1995. Unrooted genealogical tree probabilities in the infinitely-many-sites model. Math Biosci 127:77–98.

Hardy J, Pittman A, Myers A, Gwinn-Hardy K, Fung HC, de Silva R, Hutton M, Duckworth J. 2005. Evidence suggesting that homo neanderthalensis contributed the H2 MAPT haplotype to homo sapiens. Biochem Soc Trans 33:582–585.

Hawks J, Cochran G, Harpending HC, Lahn BT. 2008. A genetic legacy from archaic homo. Trends Genet 24:19–23.

Hudson RR. 2001. Two-locus sampling distributions and their application. Genetics 159:1805–1817.

Hudson RR, Boos DD, Kaplan NL. 1992. A statistical test for detecting geographic subdivision. Mol Biol Evol 9:138–151.

Hudson RR, Kreitman M, Aguade M. 1987. A test of neutral molecular evolution based on nucleotide data. Genetics 116:153–159.

Hvilsom C, Qian Y, Bataillon T, Li Y, Mailund T, Salle B, Carlsen F, Li R, Zheng H, Jiang T, Jiang H, Jin X, et al. 2012. Extensive X-linked adaptive evolution in central chimpanzees. Proc Natl Acad Sci USA 109:2054–2059.

Kim DS, Li KW, Boroujerdi A, Peter Yu Y, Zhou CY, Deng P, Park J, Zhang X, Lee J, Corpe M, Sharp K, Steward O, et al. 2012. Thrombospondin-4 contributes to spinal sensitization and neuropathic pain states. J Neurosci 32:8977–8987.

Kim HL, Satta Y. 2008. Population genetic analysis of the N-acylsphingosine amidohydrolase gene associated with mental activity in humans. Genetics 178:1505–1515.

Kimura M. 1983. The neutral theory of molecular evolution. Cambridge: Cambridge University Press.

Lappalainen T, Dermitzakis ET. 2010. Evolutionary history of regulatory variation in human populations. Hum Mol Genet 19:R197–R203.

Li JZ, Absher DM, Tang H, Southwick AM, Casto AM, Ramachandran S, Cann HM, Barsh GS, Feldman M, Cavalli-Sforza LL, Myers RM. 2008. Worldwide human relationships inferred from genome-wide patterns of variation. Science 319:1100–1104.

Liang WS, Dunckley T, Beach TG, Grover A, Mastroeni D, Ramsey K, Caselli RJ, Kukull WA, McKeel D, Morris JC, Hulette CM, Schmechel D, et al. 2008. Altered neuronal gene expression in brain regions differentially affected by Alzheimer's disease: a reference data set. Physiol Genomics 33:240–256.

Liang WS, Dunckley T, Beach TG, Grover A, Mastroeni D, Ramsey K, Caselli RJ, Kukull WA, McKeel D, Morris JC, Hulette CM, Schmechel D, et al. 2010. Neuronal gene expression in non-demented individuals with intermediate Alzheimer's disease neuropathology. Neurobiol Aging 31:549–566.

Loisel DA, Rockman MV, Wray GA, Altmann J, Alberts SC. 2006. Ancient polymorphism and functional variation in the primate MHC-DQA1 5' dis-regulatory region. Proc Natl Acad Sci USA 103:16331–16336.

Lu L, Airey DC, Williams RW. 2001. Complex trait analysis of the hippocampus: mapping and biometric analysis of two novel gene loci with specific effects on hippocampal structure in mice. J Neurosci 21:3503–3514.

McKhann GM, Knopman DS, Chertkow H, Hyman BT, Jack CR, Jr, Kawas CH, Klunk WE, Koroshetz WJ, Manly JJ, Mayeux R, Mohs RC, Morris JC, et al. 2011. The diagnosis of dementia due to Alzheimer's disease: recommendations from the national institute on aging-Alzheimer's association workgroups on diagnostic guidelines for Alzheimer's disease. Alzheimers Dement 7:263–269.

Montgomery SB, Sammeth M, Gutierrez-Arcelus M, Lach RP, Ingle C, Nisbett J, Guigo R, Dermitzakis ET. 2010. Transcriptome genetics using second generation sequencing in a caucasian population. Nature 464:773–777.

Myers AJ, Gibbs JR, Webster JA, Rohrer K, Zhao A, Marlowe L, Kaleem M, Leung D, Bryden L, Nath P, Zismann VL, Joshipura K, et al. 2007. A survey of genetic human cortical gene expression. Nat Genet 39:1494–1499.

Nei M, Li WH. 1979. Mathematical model for studying genetic variation in terms of restriction endonucleases. Proc Natl Acad Sci USA 76:5269–5273.

Pluskota E, Stenina OI, Krukovets I, Szpak D, Topol EJ, Plow EF. 2005. Mechanism and effect of thrombospondin-4 polymorphisms on neutrophil function. Blood 106:3970–3978.

Raj T, Shulman JM, Keenan BT, Chibnik LB, Evans DA, Bennett DA, Stranger BE, De Jager PL. 2012. Alzheimer disease susceptibility loci: evidence for a protein network under natural selection. Am J Hum Genet 90:720–726.

Reinius B, Saetre P, Leonard JA, Blekhman R, Merino-Martinez R, Gilad Y, Jazin E. 2008. An evolutionarily conserved sexual signature in the primate brain. PLoS Genet 4:e1000100.

Rimol LM, Agartz I, Djurovic S, Brown AA, Roddey JC, Kahler AK, Mattingsdal M, Athanasiu L, Joyner AH, Schork NJ, Halgren E, Sundet K, et al. 2010. Sex-dependent association of common variants of microcephaly genes with brain structure. Proc Natl Acad Sci USA 107:384–388.

Roca X, Olson AJ, Rao AR, Enerly E, Kristensen VN, Borresen-Dale AL, Andresen BS, Krainer AR, Sachidanandam R. 2008. Features of 5'-splice-site efficiency derived from disease-causing mutations and comparative genomics. Genome Res 18:77–87.

Rodwell GE, Sonu R, Zahn JM, Lund J, Wilhelmy J, Wang L, Xiao W, Mindrinos M, Crane E, Segal E, Myers BD, Brooks JD, et al. 2004. A transcriptional profile of aging in the human kidney. PLoS Biol 2:e427.

Schaffner SF, Foo C, Gabriel S, Reich D, Daly MJ, Altshuler D. 2005. Calibrating a coalescent simulation of human genome sequence variation. Genome Res 15:1576–1583.

Shupe JM, Kristan DM, Austad SN, Stenkamp DL. 2006. The eye of the laboratory mouse remains anatomically adapted for natural conditions. Brain Behav Evol 67:39–52.

Smith SM. 2002. Fast robust automated brain extraction. Hum Brain Mapp 17:143–155.

Smith SM, De Stefano N, Jenkinson M, Matthews PM. 2001. Normalized accurate measurement of longitudinal brain change. J Comput Assist Tomogr 25:466–475.

Smith SM, Jenkinson M, Woolrich MW, Beckmann CF, Behrens TE, Johansen-Berg H, Bannister PR, De Luca M, Drobnjak I, Flitney DE, Niazy RK, Saunders J, et al. 2004. Advances in functional and structural MR image analysis and implementation as FSL. Neuroimage 23 Suppl 1:S208–S219.

Smith SM, Zhang Y, Jenkinson M, Chen J, Matthews PM, Federico A, De Stefano N. 2002. Accurate, robust, and automated longitudinal and cross-sectional brain change analysis. Neuroimage 17:479–489.

Somel M, Liu X, Tang L, Yan Z, Hu H, Guo S, Jiang X, Zhang X, Xu G, Xie G, Li N, Hu Y, et al. 2011. MicroRNA-driven developmental remodeling in the brain distinguishes humans from other primates. PLoS Biol 9:e1001214.

Stenina OI, Desai SY, Krukovets I, Kight K, Janigro D, Topol EJ, Plow EF. 2003. Thrombospondin-4 and its variants: expression and differential effects on endothelial cells. Circulation 108:1514–1519.

Stenina OI, Topol EJ, Plow EF. 2007. Thrombospondins, their polymorphisms, and cardiovascular disease. Arterioscler Thromb Vasc Biol 27:1886–1894.

Stephens M, Scheet P. 2005. Accounting for decay of linkage disequilibrium in haplotype inference and missing-data imputation. Am J Hum Genet 76:449–462.

Stephens M, Smith NJ, Donnelly P. 2001. A new statistical method for haplotype reconstruction from population data. Am J Hum Genet 68:978–989.

Tajima F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 123:585–595.

Thornton K. 2003. Libsequence: a C++ class library for evolutionary genetic analysis. Bioinformatics 19:2325–2327.

Tishkoff SA, Verrelli BC. 2003. Patterns of human genetic diversity: implications for human evolutionary history and disease. Annu Rev Genomics Hum Genet 4:293–340.

Wall JD. 2000. Detecting ancient admixture in humans using sequence polymorphism data. Genetics 154:1271–1279.

Watterson GA. 1975. On the number of segregating sites in genetical models without recombination. Theor Popul Biol 7:256–276.

Webster MT, Smith NG, Hultin-Rosenberg L, Arndt PF, Ellegren H. 2005. Male-driven biased gene conversion governs the evolution of base composition in human alu repeats. Mol Biol Evol 22:1468–1474.

Welle S, Brooks AI, Delehanty JM, Needler N, Thornton CA. 2003. Gene expression profile of aging in human muscle. Physiol Genomics 14:149–159.

Wright S. 1950. Genetical structure of populations. Nature 166:247–249.

Wright SI, Charlesworth B. 2004. The HKA test revisited: a maximum-likelihood-ratio test of the standard neutral model. Genetics 168:1071–1076.

Zeller T, Wild P, Szymczak S, Rotival M, Schillert A, Castagne R, Maouche S, Germain M, Lackner K, Rossmann H, Eleftheriadis M, Sinning CR, et al. 2010. Genetics and beyond—the transcriptome of human monocytes and disease susceptibility. PLoS ONE 5:e10693.

Zhang D, Hu X, Qian L, Chen SH, Zhou H, Wilson B, Miller DS, Hong JS. 2011. Microglial MAC1 receptor and PI3K are essential in mediating beta-amyloid peptide-induced microglial activation and subsequent neurotoxicity. J Neuroinflammation 8:3.

Zhang Y, Brady M, Smith S. 2001. Segmentation of brain MR images through a hidden markov random field model and the expectation-maximization algorithm. IEEE Trans Med Imaging 20:45–57.

# APPENDIX

# III

# APPENDIX III

This appendix consists on an article published in *BMC Genomics* in 2013 in collaboration with the group of the ICREA Professor Tomás Marqués-Bonet from the Universitat Pompeu Frabra (Barcelona, Spain).

For this article (PRADO-MARTINEZ *et al.* 2013), we focused in the prediction of inversions from paired-end mapping data, and specifically, I could participate in the experimental validation of five predicted inversions by PCR amplification: GgInv420 (3.1 Kb), GgInv160 (3.8 Kb), GgInv425 (17.5 Kb), GgInv187 (21.8 Kb), and GgInv315 (34.8 Kb). DNA from 2 anonymous humans, 2 gorillas (Snowflake and it son Urko), and one chimpanzee were used for the validation. In all cases, single bands of the expected size for the inverted and non-inverted orientation were obtained in the two gorilla and two human samples, respectively, confirming the presence of an inversion between the two species and the accurate location of the breakpoint. According to the distribution of the PCR bands across species and the analysis of the breakpoint regions in the genomes of the four ape species, two of the inversions appear to be specific to humans (GgInv187 and GgInv425), two appear to be specific to gorillas (GgInv315 and GgInv420), and one was specific to the lineage common to humans and chimpanzees (GgInv160).

Prado-Martinez et al. BMC Genomics 2013, 14:363
http://www.biomedcentral.com/1471-2164/14/363

BMC Genomics

RESEARCH ARTICLE Open Access

# The genome sequencing of an albino Western lowland gorilla reveals inbreeding in the wild

Javier Prado-Martinez[1], Irene Hernando-Herraez[1], Belen Lorente-Galdos[1], Marc Dabad[1], Oscar Ramirez[1], Carlos Baeza-Delgado[2], Carlos Morcillo-Suarez[1,3], Can Alkan[4,5], Fereydoun Hormozdiari[4], Emanuele Raineri[6], Jordi Estellé[6,7], Marcos Fernandez-Callejo[1], Mònica Valles[1], Lars Ritscher[8], Torsten Schöneberg[8], Elisa de la Calle-Mustienes[9], Sònia Casillas[10], Raquel Rubio-Acero[10], Marta Melé[1,11], Johannes Engelken[1,12], Mario Caceres[10,13], Jose Luis Gomez-Skarmeta[9], Marta Gut[6], Jaume Bertranpetit[1], Ivo G Gut[6], Teresa Abello[14], Evan E Eichler[4,15], Ismael Mingarro[2], Carles Lalueza-Fox[1], Arcadi Navarro[1,3,13,16] and Tomas Marques-Bonet[1,13*]

## Abstract

**Background:** The only known albino gorilla, named *Snowflake*, was a male wild born individual from Equatorial Guinea who lived at the Barcelona Zoo for almost 40 years. He was diagnosed with non-syndromic oculocutaneous albinism, i.e. white hair, light eyes, pink skin, photophobia and reduced visual acuity. Despite previous efforts to explain the genetic cause, this is still unknown. Here, we study the genetic cause of his albinism and making use of whole genome sequencing data we find a higher inbreeding coefficient compared to other gorillas.

**Results:** We successfully identified the causal genetic variant for *Snowflake*'s albinism, a non-synonymous single nucleotide variant located in a transmembrane region of *SLC45A2*. This transporter is known to be involved in oculocutaneous albinism type 4 (OCA4) in humans. We provide experimental evidence that shows that this amino acid replacement alters the membrane spanning capability of this transmembrane region. Finally, we provide a comprehensive study of genome-wide patterns of autozygogosity revealing that *Snowflake*'s parents were related, being this the first report of inbreeding in a wild born Western lowland gorilla.

**Conclusions:** In this study we demonstrate how the use of whole genome sequencing can be extended to link genotype and phenotype in non-model organisms and it can be a powerful tool in conservation genetics (e.g., inbreeding and genetic diversity) with the expected decrease in sequencing cost.

**Keywords:** Gorilla, Albinism, Inbreeding, Genome, Conservation

## Background

The only known albino gorilla named *Snowflake* (Figure 1) was a male wild-born Western lowland gorilla (*Gorilla gorilla gorilla*) from Equatorial Guinea. He was brought to the Barcelona Zoo in 1966 at young age [1], where he gained popularity worldwide. *Snowflake* presented the typical properties of albinism as seen in humans: white hair, pink skin, blue eyes, reduced visual acuity and photophobia. Given his lack of pigmentation and thus reduced protection from UV light, the aged albino gorilla developed squamous-cell carcinoma that led to his euthanasia in 2003 [2].

*Snowflake* was diagnosed with non-syndromic albinism (Oculocutaneous Albinism, OCA). This is a group of Mendelian recessive disorders characterized by the generalized reduction of pigmentation in skin, hair, and eyes. Pigmentation is determined by melanin compounds, which are produced in melanocytes and are transported via melanosomes into keratinocytes of the epidermis and hair follicles. It has been widely studied in humans and four genes are found to be causative of this disorder: (i) OCA1A/B (MIM 203100,606952) are caused by mutations in the gene *TYR* (*Tyrosinase*) (ii) mutations in the *OCA2* gene (previously known as *P-gene*) can cause OCA2

* Correspondence: tomas.marques@upf.edu
[1]Institut de Biologia Evolutiva, (CSIC-Universitat Pompeu Fabra), PRBB, Barcelona 08003, Spain
[13]Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona 08010, Spain
Full list of author information is available at the end of the article

**Figure 1 *Snowflake*, the only known albino gorilla.** This Western lowland gorilla was wild-born in Equatorial Guinea and he presented the typical characteristics of oculocutaneous albinism.

phenotype (MIM 203200) (iii) mutations in *TYRP1* cause OCA3 (MIM 203290) and (iv) OCA4 (MIM 606574) is caused by mutations in *SLC45A2* (formerly known as *MATP* and *AIM1*) [3]. Tyrosinase and TYRP1 are critical in the melanin synthesis pathway whereas P protein (OCA2) and SLC45A2 are involved in melanocytes maintenance or formation.

A previous study tried to assess whether the causative mutation of *Snowflake*'s albinism was located in the TYR gene but no causative mutation was found [4]. Here, we make use of whole genome sequencing to provide a better characterization of all known genes related to albinism to try to ascertain the genetic component causing this phenotype and to study genome wide patterns that can help the field of conservation genetics. Most of the knowledge about ecology, population dynamics, demography and social behavior about gorillas has been collected from mountain gorillas (*Gorilla beringei beringei*) and until recently this has not expanded to Western lowland gorillas [5,6]. This effort has been extremely helpful to improve our knowledge and conservation of this endangered species. With the development of conservation genetics we have gained insights into population genetics [7], demographic history [8] and group relationships through the usage of both microsatellites and mitochondrial markers. The main difficulty of these studies is that non-invasive samples such as hair or feces cannot provide DNA of high quality.

Here, using high quality DNA and next-generation sequencing, we have studied for the first time the whole genome of a wild born Western lowland gorilla. It is important to stress that previous whole-genome sequencing projects of Western lowland gorillas, involved captive-born individuals, Kamilah [9] and Kwan [10], individuals that do not belong to a wild population as it has been recently studied with microsatellite markers [11]. Studying this unique albino gorilla, we find the first evidence of inbreeding in wild Western lowland gorillas.

## Results

We sequenced the genome of *Snowflake* at 18.7× effective coverage using the Illumina GAIIx platform (114 bp paired-end reads). We aligned the reads to the reference human genome (NCBI build 37) using GEM [12], and used samtools [13] to identify single nucleotide variants (SNVs) (Methods). We found 73,307 homozygous non-synonymous *Snowflake*'s mutations compared to the human reference genome. Out of those, 20 were found within candidate genes for albinism (OCA related genes), but a single mutation was private compared to two other sequenced gorillas (Additional file 1: Tables S1 and S2) [9,10]. This substitution is located in the last exon of the *SLC45A2* gene at the position hg19: chr5_33944794_C/G and it causes a substantial amino acid change, Glycine to Arginine, (pGly518Arg) in a predicted transmembrane region of the protein. We then resequenced this mutation using capillary sequencing and it was confirmed as homozygous in *Snowflake* and heterozygous in all five tested non-albino offspring, as expected in Mendelian recessive disorders. To rule out the possible participation of other candidate genes, we also looked for structural variants that may be disrupting other genes related to pigmentation. We applied computational methods based on paired-end and split read approaches to detect genomic deletions (Methods), followed by experimental validation using array-comparative genomic hybridization (aCGH). We identified 1,390 validated deletions totaling to 9.5 Mbps, a similar proportion of the genome compared to previous reports [5] (Additional file 1: Table S3). These deletions overlap completely with 36 RefSeq transcripts and partially (>10%) with 660 transcripts (Additional file 1: Table S4) but none of them has a direct association with albinism.

Several pieces of evidence support the hypothesis that the non-synonymous mutation found in *SCL45A2* might be responsible for *Snowflake*'s albinism. First, this specific Glycine residue is conserved throughout all available vertebrate taxa (Additional file 2: Figure S1), suggesting a conserved role of this amino acid. Second, we predicted whether this amino acid change may affect the protein

structure and function based on sequence conservation and protein properties using SIFT [14] and PolyPhen-2 [15]. It is predicted as a "damaging" mutation by SIFT, and "probably damaging" by PolyPhen-2. Third, this gene was reported to be the genetic cause of albinism in several other species (*e.g.*, mouse [16], medaka fish [17], horse [18] and chicken [19]). Last, previous reports showed that Glycine to Arginine mutations within other transmembrane regions of *SCL45A2* in humans result in severe albino phenotypes [20].

We followed up on this finding with an experimental study to determine how this amino acid substitution affects the transmembrane segment where this mutation is present. For this purpose we used a functional assay based on *Escherichia coli* inner membrane protein leader peptidase (Lep) that detects and permits accurate measurements of the apparent free energy ($\Delta G_{app}$) of translocon-mediated integration of transmembrane helices into the endoplasmic reticulum (ER) membranes [21-23]. This procedure allows the quantification of the proper integration of the transmembrane region with the normal sequence and with the mutation. When we assayed the construct with the wild type sequence, we observed that 90% of the proteins were properly recognized for membrane insertion. However, translation of the mutant (G518R) found in *Snowflake* resulted in a significant reduction (~25%, p-value = 0.036 Mann–Whitney U test) in the membrane integration capability (Figure 2), suggesting that the replacement of a glycine by an arginine residue lowers the affinity of the transmembrane region and possibly alter the topology of the *SLC45A2* gene product.

Finally, the last piece of evidence supporting the role of the mutation in the phenotype is based on genome-wide patterns of heterozygosity in the genome of *Snowflake* (Additional file 2: Figure S2). We found that *SLC45A2* gene is located in a large run of homozygosity (40 Mbps) orthologous to human chromosome 5 (Figure 3a), meaning that this allele was inside a block identical by descent, which is characteristic for Mendelian recessive disorders. The other three candidate genes are not found in any autozygous regions. Overall, we found 25 large runs of homozygosity (longer than 2 Mbps), and a general reduction of heterozygosity compared to the other known genomes sequenced of the same species (Figure 3b). Some of the runs of homozygosis are particularly large, such as a continuous 68 Mbps segment in chromosome 4 (Additional file 2: Figure S2). This reduction of variation might allow the emergence of certain phenotypes otherwise masked by dominance, and they could lead to inbreeding depression [24], as previously reported in chimpanzee and other primates [25].

These patterns of heterozygosity allow the estimation of the amount of autozygosity, i.e. long regions of the genome identical by descent as a result of inbreeding. The minimum threshold of these stretches has been estimated in humans ranging between 1-2.25 Mbp depending on the population [26]. We conservatively quantified the inbreeding coefficient in *Snowflake* based on autozygosity ($F_{ROH}$) of 0.118 (306 Mbps out of 2,587 Mbps) compared to 0.002 and <0.001 estimated from the previously sequenced gorillas, Kamilah and Kwan respectively, using the same criteria (Methods). Assuming no previous inbreeding in any of the parents, 0.125 would correspond to parents being grandparent–grandchild, half-siblings, or uncle–niece/aunt–nephew. We performed a set of simulations that replicated the patterns of autozygosity in different pedigrees, accounting for the differential amount of recombinations derived from the number of meiosis

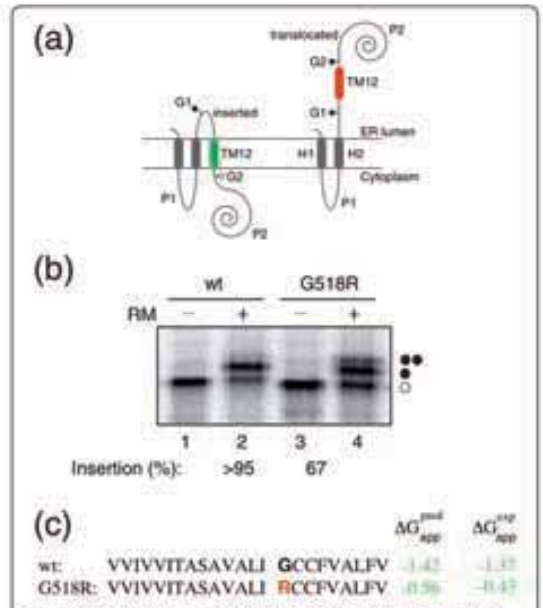



**Figure 2 Membrane integration of non-albino wild-type (wt) and mutant (*Snowflake*) G518R sequences.** (**a**) Schematic representation of the engineered leader peptidase (Lep) model protein. Wild-type Lep has two transmembrane (TM) helices (H1 and H2) and a large C-terminal luminal domain (P2). It inserts into rough microsomal (RM) membranes in an Nt/Ct ER luminal orientation. *SLC45A2* TM12 domains (TM12) wild-type and G518R mutant were inserted in the P2 domain flanked by two glycosylation acceptor sites (G1 and G2). If the inserted sequence integrates into the membrane, only the G1 site is glycosylated (left), whereas both G1 and G2 sites are glycosylated for the sequences that do not integrate into the membrane (right). (**b**) Plasmids encoding the constructs were transcribed and translated in vitro in the absence (–) and presence (+) of RM membranes. Non-glycosylated protein bands are indicated by a white dot; singly or doubly glycosylated proteins are indicated by one or two black dots, respectively. (**c**) *SLC45A2* TM12 sequence in each construct and $\Delta G_{app}$ values predicted using the ΔG Prediction Server (http://dgpred.cbr.su.se/) and deduced from the data in panel b are shown.

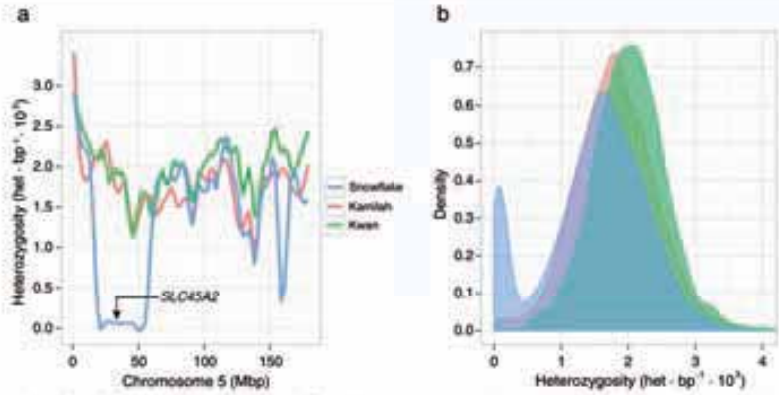



**Figure 3 Heterozygosity distributions of the studied gorillas.** (**a**) Heterozygosity values in 1 Mbp non-overlapping windows along the human chromosome 5. The *SLC45A2* gene is contained within a 40 Mbp run of homozygosity in *Snowflake* (blue). The other two gorillas do not show any run of homozygosity (Kamilah in red and Kwan in green). (**b**) Distribution of the heterozygosity values of 1Mbp non-overlapping windows in the aforementioned gorillas. While most of the genome shows a similar distribution, we observe an excess of windows with less heterozygosity as a result of the inbreeding observed in *Snowflake*.

and the sex of the ancestors that is known to influence recombination rates [27]. These simulations reconstruct the different recombination patterns in different possible pedigrees under a random paternal transmission, and we estimated which fragments may appear as autozygous in the offspring. Finally, we compared the distribution of sizes and number of autozygous segments in *Snowflake* with the different simulated outcomes to calculate a likelihood for each case. The uncle/niece or aunt/nephew combination is the most probable scenario, although there was not a unique best statistical pedigree (Additional file 2: Figure S3C-F and Additional file 1: Table S5).

## Discussion

We sequenced the whole genome of a phenotypically unique gorilla, identified and characterized the causative mutation for his albinism, and explored the origin of this trait. We found a private non-synonymous substitution in one of the candidate genes - the *SLC45A2* gene - associated with the OCA4 class of albinism. We provided several lines of evidence based on evolution, human disease and a functional assay supporting that this mutation in a transmembrane domain can modify the topology of the translated protein, therefore reinforcing its causative role in this rare case of albinism. Moreover, long runs of homozygosity in this wild born individual explain the emergence of this recessive trait through identity by descent, suggesting that inbreeding was an important factor towards the emergence of this phenotype.

We inferred that *Snowflake* was an offspring of closely related individuals supported by an inbreeding coefficient of 0.118. In general, inbreeding is avoided in the wild because the offsprings within gorilla societies disperse to other groups before maturity [28]. This is strictly true in patriarchal groups that are commonly composed of a silverback male and several females (97% of all the gorilla groups) [6] whereas in multimale groups, females can remain and have their first birth in the natal group. Multimale groups are usually composed of related males and therefore, newborn females are likely to be also related to them (commonly with relationships such as half brothers or half uncles). However, multimale groups have mainly been observed in mountain gorillas, while only two multimale groups have ever been reported in Western lowland gorillas, suggesting that they are extremely rare in these populations [6]. Therefore, it seems unlikely that a multimale group would explain the inbreeding found in this Western lowland gorilla.

Previous parentage studies in wild Western lowland gorillas have never found inbred mating, suggesting that is probably a rare behavior [29]. Despite this and considering that the observation of inbreeding in a single individual could be an extreme case, some social observations may point that inbreeding may still occur. First, gorillas seem to follow a patrilocal social structure, i.e. silverbacks are usually related to one or more nearby silverbacks [30]. Additionally, females transfer several times during their lifespan after the dispersal from their natal group [31], which may result in the arrival of a female to a new group where the silverback is related to her. Although father-daughter inbreeding is completely avoided, this hypothesis is feasible because other mating relationships, with half-brothers or even full brothers, are possible; suggesting that females do not detect consanguinity [6]. Other factors such as habitat loss, small population sizes and population fragmentation may influence the disposal of breeding

groups and therefore of unrelated silverbacks which may in turn favor inbreeding [32]. Other potential explanations are less likely; male takeovers are highly avoided and the death of a male silverback normally results in the disintegration of the group and female dispersal.

A previous study using microsatellite markers in captive gorilla populations showed that their genetic diversity is comparable to wild gorilla populations [11]. However, in our study, *Snowflake* shows different patterns of heterozygosity compared to the captive born gorillas,. The gorilla studbooks show that *Kamilah* (Studbook ID: 661) is a first generation captive-born gorilla, while *Kwan* (Studbook ID: 1107) is a second generation captive-born gorilla. When we compared the heterozygosity genome-wide, we observed that *Kwan* is the gorilla with higher heterozygosity, despite we cannot rule out that this was a result of some false positives due to the lower sequencing coverage. *Kamilah* and *Snowflake* have lower heterozygosity, with the albino gorilla showing the lowest values compared to the other captive-born individuals (Figure 3b) even accounting for the regions of inbreeding. This suggests that breeding programs could result in an increase of genetic variation but a bigger sample size would be needed to systematically explore this effect.

In this particular study, we show that high throughput sequencing can be used not only to unravel the genetic mechanisms of fundamental phenotypes (including disease) in non-model organisms, but also to provide insights into conservation genetics through the detection of inbreeding of endangered species such as gorilla. However, in order to systematically explore relationships and breeding patterns from wild specimens using whole genome sequencing data, high quality DNA is required and in most field studies, non-invasive samples such as feces or hair are used, and the amount of DNA extracted from such samples precludes the application of this methodology to conservation studies. Still, it can be applied to analyze the genomes of wild born individuals in zoos where blood samples are usually taken during routine veterinary check-ups. However, sequencing technologies quickly and constantly improve, and recent developments that includes library construction with very little amounts of DNA [33] or single-cell sequencing [34,35] may allow the implementation of this kind of analyses into conservation genetics in the near future.

## Conclusions

Here we make use of next-generation sequencing to study the complete genome of a wild born Western lowland gorilla genome. Using these data we have been able to identify the genetic cause of a rare phenotype --albinism– in this non-model species and we provide several lines of evidence that reinforces this hypothesis, ranging from evolution to human disease. Moreover, we have been able to characterize that this individual was descendant of close relatives by studying the patterns of autozygosity genome-wide, providing the first genetic evidence of inbreeding in this species. We discuss this finding from the perspective of gorilla societies and we link several pieces of information in order to provide plausible scenarios where this event could have happened. We envision that the analysis of whole genome data of endangered species will be a standard in future conservation and management studies and will make available relevant information that has been missed in previous studies.

## Methods

### Sequencing

We extracted DNA using phenol-chlorophorm from a frozen blood sample previously taken from *Snowflake* (*Gorilla gorilla gorilla*). We constructed Illumina libraries using the standard protocol with two different fragment sizes at 250 bp and 450 bp. We sequenced the genome at ~18× coverage with paired-end reads (114 nt). For comparison purposes, we analyzed the genomes of two other Western lowland Gorillas (Kwan [10]) and Kamilah [9]) (Additional file 1: Table S1). The research did not involve any experiment on human subjects or animals and for this reason no ethical approval was necessary, the blood used for the sequencing of *Snowflake* was extracted after the death of the gorilla.

### Single nucleotide variants

We mapped all reads to the human reference genome (GRCh37) using GEM [12] allowing a divergence of 4% in order to capture all putative changes between human and gorilla keeping uniquely placed reads. We identified single nucleotide differences with *Samtools* [13] (v.0.1.9), and filtered out potential false positives by mapping quality and read depth (based on the different sequencing depths for the samples).

### Copy number variation discovery and validation

We assessed genomic structural variants compared to the human genome using a combination of paired-end and split read methods to provide an initial catalog of potential deletions. In order to validate these regions using an independent approach, we further analyzed them with Array-comparative genomic hybridization. Finally, we reported the regions that revealed a variation and that were concordant using both methodologies (Additional file 2).

### Inbreeding

To estimate the degree of heterozygosity in the genome of *Snowflake*, we divided the genome into 1 Mbp non-overlapping windows and calculated the heterozygous positions per Kbp. To avoid divergent outliers in the estimates of each window, we removed all regions that overlapped more than 40% with duplications, and we corrected the

number of heterozygous positions by the remaining effective bases of the windows. To calculate the inbreeding coefficient, we conservatively considered regions with a loss of heterozygosity when at least two consecutive 1 Mbp windows showed a reduction of heterozygosity.

### Inbreeding simulations

Computer simulations were run in order to infer the family history that may be responsible for the pattern of homozygous fragments found in the genome of *Snowflake*. We considered all the possible pedigrees: half-siblings, aunt/nephew, uncle/niece, grandfather/granddaughter and grandmother/grandson. A total of 10 models were defined to account for the different pedigree combinations that can generate the above parental origins (Additional file 2: Figure S3).

For each pedigree model, 10,000 simulated "*Snowflakes*" were created considering no relationship among founding members. We used different rates of recombination for males and females (8.9×10-9 crossovers/nucleotide for males, and 1.4×10-8 for females) following empirical data [27].

The simulations were performed using a Java program written ad hoc for this particular purpose. For every founder individual in the pedigree two sets of chromosomes are generated containing different alleles (zero inbreeding is assumed among all founders). In descendant individuals, chromosomes are generated by crossing over parental chromosomes and randomly passing one out of the two present in each parent to the offspring. Descendant individuals will have a mix of founder alleles in their chromosomes. Due to the inbred structure of the pedigrees, the simulated *Snowflake* is expected to present regions where both chromosomes have the same allele originated from a single founder individual.

Number and length distribution of homozygous fragments resulting from each model were compared with the actual values in the genome of *Snowflake* (Additional file 2: Figures S4 and S5). For each model, homozygous fragments obtained were classified according to their length into 5 Mbp beans and a multinomial distribution was defined using the resulting counts. The probability of these distributions of generating the actual *Snowflake* counts was used as a measure of likelihood for each model (Additional file 1: Table S5). To make the obtained data compatible with experimental data, we removed segments smaller than 2 Mbp and merged the segments separated by gaps smaller than 500 Kbp.

### Mutant membrane integration

Wild type and *Snowflake* constructs in pGEM1 were transcribed and translated in the TNT® SP6 Quick Coupled System from Promega. DNA template (~75 ng), 1 μl of [35S]Met/Cys (5 μCi), and 1 μl of dog pancreas RMs were added to 5 μl of lysate at the start of the reaction, and samples were incubated for 90 min at 30°C. The translation reaction mixture was diluted in 5 volumes of phosphate buffer saline (pH 7.4). Subsequently, membranes were collected by layering the supernatant onto a 50 μl sucrose cushion and centrifuged at 100,000 × g for 20 min at 4°C in a Beckman tabletop ultracentrifuge with a TLA-45 rotor. Finally, pellets were analyzed by SDS-PAGE, and gels were visualized on a Fuji FLA3000 phosphorimager using the ImageGauge. (Additional file 2).

## Additional files

**Additional file 1: Contains the supplementary tables. Table S1**: Summary of the samples used in this study. **Table S2**: Non-synonymous mutations found in OCA genes compared to human genes. **Table S3**: Summary of deletions found in Snowflake using different methodologies. **Table S4**: List of transcripts affected by deletions. **Table S5**: Likelihood values in the paternity simulations.

**Additional file 2: Contains detailed explanation on some methods and supplementary figures.**

### Competing interests

The authors declare that they have no competing interests.

### Authors' contributions

JP-M and TM-B designed the study and drafted the manuscript. JP-M, IB-G, MD, CM-S, CA, FH, ER, JE, MF-C, SC, MM, RR and MC conducted bioinformatics analysis. IH-H, OR, CB-D, MV, LR, TS, EC-M, JE, JLGS, MG, IGG and IM performed experiments. TA, JB, IM, EEE, CL-F and AN helped to write the manuscript. All authors read and approved the final manuscript.

### Acknowledgements

Jordi Camps, Luis Alberto Perez Jurado and Lorna Brocksopp for technical help. The Spanish Government for grants BFU2010-14839 to JLG-S, Spanish Government and FEDER for grants BFU2009-13409-C02-02 and BFU2012-38236 to AN and JP-M, BFU2012-39482 to IM, and BFU2011-28549 to TM-B. The Andalusian Government for grants CSD2007-00008 and CVI-3488, supported by FEDER to JLG-S. The Barcelona Zoo (Ajuntament de Barcelona) for an award to JP-M. EEE is an investigator with the Howard Hughes Medical Institute.The European Community for an ERC Starting Grant (StG_20091118)) to TM-B.

### Author details

[1]Institut de Biologia Evolutiva, (CSIC-Universitat Pompeu Fabra), PRBB, Barcelona 08003, Spain. [2]Departament de Bioquímica i Biologia Molecular, Universitat de València, Burjassot E-46100, Spain. [3]Instituto Nacional de Bioinformatica, UPF, Barcelona, Spain. [4]Department of Genome Sciences, University of Washington, 3720 15th AVE NE, Seattle, WA 98195, USA. [5]Department of Computer Engineering, Bilkent University, Ankara, Turkey. [6]Centro Nacional de Análisis Genómico, PCB, Barcelona 08028, Spain. [7]Current address: INRA, UMR1313 GABI, Jouy-en-Josas, France. [8]Institute of Biochemistry, University of Leipzig, Leipzig 04103, Germany. [9]Centro Andaluz de Biología del Desarrollo, Consejo Superior de Investigaciones Científicas, Universidad Pablo de Olavide and Junta de Andalucía, Carretera de Utrera Km1, Sevilla 41013, Spain. [10]Institut de Biotecnologia i de Biomedicina, Universitat Autònoma de Barcelona, Bellaterra, 08193, Barcelona, Spain. [11]Current address: Centre for Genomic Regulation and UPF, Doctor Aiguader 88, Barcelona 08003, Catalonia, Spain. [12]Department of Evolutionary Genetics, Max-Planck Institute for Evolutionary Anthropology, Leipzig 04103, Germany. [13]Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona 08010, Spain. [14]Parc Zoològic de Barcelona, Barcelona 08003, Spain. [15]Howard Hugues Medical Institute, 3720 15th AVE NE, Seattle, WA 98195, USA. [16]Centre for Genomic Regulation and UPF, Doctor Aiguader 88, Barcelona 08003, Catalonia, Spain.

Received: 14 January 2013 Accepted: 23 May 2013
Published: 31 May 2013

## References

1. Sabater Pi J: **An albino lowland gorilla from rio Muni, West Africa, and notes on its adaptation to captivity.** *Folia Primatol* 1967, **7**:155–160.
2. Márquez M, Serafín A, Fernández-Bellon H, Senut S, Ferrer-Admetlla A, Bertranpetit J, Ferrer I, Pumarola M: **Neuropathologic findings in an aged albino gorilla.** *Vet Pathol* 2008, **45**:531–537. doi:10.1354/vp.45-4-531.
3. Grønskov K, Ek J, Brondum-Nielsen K: **Oculocutaneous albinism.** *Orphanet J Rare Dis* 2007, **2**:43. doi:10.1186/1750-1172-2-43.
4. Martínez-Arias R, Comas D, Andrés A, Abelló MT, Domingo-Roura X, Bertranpetit J: **The tyrosinase gene in gorillas and the albinism of "Snowflake".** *Pigment Cell Res* 2000, **13**:467–470.
5. Robbins MM, Bermejo M, Cipolletta C, Magliocca F, Parnell RJ, Stokes E: **Social structure and life-history patterns in western gorillas (Gorilla gorilla gorilla).** *Am J Primatol* 2004, **64**:145–159. doi:10.1002/ajp.20069.
6. Harcourt AH, Stewart KJS: *Gorilla Society: conflict, compromise and cooperation between sexes.* Chicago: The University of Chicago Press; 2007.
7. Vigilant L, Bradley BJ: **Genetic variation in gorillas.** *Am J Primatol* 2004, **64**:161–172. doi:10.1002/ajp.20070.
8. Thalmann O, Fischer A, Lankester F, Paabo S, Vigilant L: **The complex evolutionary history of gorillas: insights from genomic data.** *Mol Biol Evol* 2007, **24**:146–158. doi:10.1093/molbev/msl160.
9. Scally A, Duthell JY, Hiller LW, Jordan GE, Goodhead I, Herrero J, Hobolth A, Lappalainen T, Mailund T, Marques-bonet T, Mccarthy S, Montgomery SH, Schwalie PC, Tang YA, Ward MC, Xue Y, Yngvadottir B, Alkan C, Andersen LN, Ayub Q, Ball EV, Beal K, Bradley BJ, Chen Y, Clee CM, Fitzgerald S, Graves TA, Gu Y, Heath P, Heger A, *et al*: **Insights into hominid evolution from the gorilla genome sequence.** *Nature* 2012, **483**:169–175. doi:10.1038/nature10842.
10. Ventura M, Catacchio CR, Alkan C, Marques-Bonet T, Sajjadian S, Graves TA, Hormozdiari F, Navarro A, Malig M, Baker C, Lee C, Turner EH, Chen L, Kidd JM, Archidiacono N, Shendure J, Wilson RK, Eichler EE: **Gorilla genome structural variation reveals evolutionary parallelisms with chimpanzee.** *Genome Res* 2011, **21**:1640–1649. doi:10.1101/gr.124461.111.
11. Nsubuga AM, Holzman J, Chemnick LG, Ryder OA: **The cryptic genetic structure of the North American captive gorilla population.** *Conserv Gen* 2009, **11**:161–172. doi:10.1007/s10592-009-0015-x.
12. Marco-Sola S, Sammeth M, Guigó R, Ribeca P: **The GEM mapper: fast, accurate and versatile alignment by filtration.** *Nat Methods* 2012. doi:10.1038/nmeth.2221.
13. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R: **The sequence alignment/Map format and SAMtools.** *Bioinformatics* 2009, **25**:2078–2079. doi:10.1093/bioinformatics/btp352.
14. Ng PC, Henikoff S: **SIFT: predicting amino acid changes that affect protein function.** *Nucleic Acids Res* 2003, **31**:3812–3814. doi:10.1093/nar/gkg509.
15. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR: **A method and server for predicting damaging missense mutations.** *Nat Methods* 2010, **7**:248–249. doi:10.1038/nmeth0410-248.
16. Newton JM, Cohen-Barak O, Hagiwara N, Gardner JM, Davisson MT, King RA, Brilliant MH: **Mutations in the human orthologue of the mouse underwhite gene (uw) underlie a new form of oculocutaneous albinism, OCA4.** *Am J Hum Genet* 2001, **69**:981–988. doi:10.1086/324340.
17. Fukamachi S, Shimada A, Shima A: **Mutations in the gene encoding B, a novel transporter protein, reduce melanin content in medaka.** *Nat Genet* 2001, **28**:381–385. doi:10.1038/ng584.
18. Marias D, Tacourt S, Guérin G: **A mutation in the *MATP* gene causes the cream coat colour in the horse.** *Genet Sel Evol* 2003, **35**:119–133. doi:10.1186/1297-9686-35-1-119.
19. Gunnarsson U, Hellström AR, Tixier-Boichard M, Minvielle F, Bed'hom B, Ito S, Jensen P, Rattink A, Vereijken A, Andersson L: **Mutations in SLC45A2 cause plumage color variation in chicken and Japanese quail.** *Genetics* 2007, **175**:867–877. doi:10.1534/genetics.106.063107.
20. Inagaki K, Suzuki T, Ito S, Suzuki N, Adachi K, Okuyama T, Nakata Y, Shimizu H, Matsuura H, Ozono T, Iwamatsu H, Kono M, Tomita Y: **Oculocutaneous albinism type 4: six novel mutations in the membrane-associated transporter protein gene and their phenotypes.** *Pigment Cell Res* 2006, **19**:451–453. doi:10.1111/j.1600-0749.2006.00332.x.
21. Hessa T, Kim H, Bihlmaier K, Lundin C, Boekel J, Andersson H, Nilsson I, White SH, Von Heijne G: **Recognition of transmembrane helices by the endoplasmic reticulum translocon.** *Nature* 2005, **433**:377–381.
22. Martínez-Gil L, Saurí A, Vilar M, Pallás V, Mingarro I: **Membrane insertion and topology of the p7B movement protein of Melon Necrotic Spot Virus (MNSV).** *Virology* 2007, **367**:348–357. doi:10.1016/j.virol.2007.06.006.
23. Martínez-Gil L, Pérez-Gil J, Mingarro I, Martínez-Gil L, Perez-Gil J: **The surfactant peptide KL 4 sequence is inserted with a transmembrane orientation into the endoplasmic reticulum membrane.** *Biophys J* 2008, **95**:36–38. doi:10.1529/biophysj.108.138602.
24. Charlesworth D, Willis JH: **The genetics of inbreeding depression.** *Nat Rev Genet* 2009, **10**:783–796. doi:10.1038/nrg2664.
25. Carpenter MJE, Widdig A, Alberts SC: **Inbreeding depression in non-human primates: a historical review of methods used and empirical data.** *Am J Primatol* 2007, **69**:1370–1386. doi:10.1002/ajp.20445.
26. Pemberton TJ, Absher D, Feldman MW, Myers RM, Rosenberg NA, Li JZ: **Genomic patterns of homozygosity in worldwide human populations.** *Am J Hum Genet* 2012, **91**:275–292. doi:10.1016/j.ajhg.2012.06.014.
27. Chowdhury R, Bois PRJ, Feingold E, Sherman SL, Vivian G: **Genetic analysis of variation in human meiotic recombination.** *PLoS Genet* 2009, **5**. doi:10.1371/journal.pgen.1000648.
28. Harcourt AH, Stewart KS, Fossey D: **Male emigration and female transfer in wild mountain gorilla.** *Nature* 1976, **263**:226–227.
29. Douadi MI, Gatti S, Levrero F, Duhamel G, Bermejo M, Vallet D, Menard N, Petit EJ: **Sex-biased dispersal in western lowland gorillas (Gorilla gorilla gorilla).** *Mol Ecol* 2007, **16**:2247–2259. doi:10.1111/j.1365-294X.2007.03286.x.
30. Bradley BJ, Doran-Sheehy DM, Lukas D, Boesch C, Vigilant L: **Dispersed male networks in western gorillas.** *Curr Biol* 2004, **14**:510–513. doi:10.1016/j.cub.2004.02.062.
31. Stokes EJ, Parnell RJ, Olejniczak C: **Female dispersal and reproductive success in wild western lowland gorillas (Gorilla gorilla gorilla).** *Behav Ecol Sociobiol* 2003, **54**:329–339. doi:10.1007/s00265-003-0630-3.
32. Bergl RA, Vigilant L: **Genetic analysis reveals population structure and recent migration within the highly fragmented range of the Cross River gorilla (Gorilla gorilla diehli).** *Mol Ecol* 2007, **16**:501–516. doi:10.1111/j.1365-294X.2006.03159.x.
33. Adey A, Morrison HG, Asan, Xun X, Kitzman JO, Turner EH, Stackhouse B, MacKenzie AP, Caruccio NC, Zhang X, Shendure J: **Rapid, low-input, low-bias construction of shotgun fragment libraries by high-density in vitro transposition.** *Genome Biol* 2010, **11**:R119. doi:10.1186/gb-2010-11-12-r119.
34. Navin N, Kendall J, Troge J, Andrews P, Rodgers L, McIndoo J, Cook K, Stepansky A, Levy D, Esposito D, Muthuswamy L, Krasnitz A, McCombie WR, Hicks J, Wigler M: **Tumour evolution inferred by single-cell sequencing.** *Nature* 2011, **472**:90–94. doi:10.1038/nature09807.
35. Peters BA, Kermani BG, Sparks AB, Alferov O, Hong P, Alexeev A, Jiang Y, Dahl F, Tang YT, Haas J, Robasky K, Zaranek AW, Lee J-H, Ball MP, Peterson JE, Perazich H, Yeung G, Liu J, Chen L, Kennemer MI, Pothuraju K, Konvicka K, Tsoupko-Sitnikov M, Pant KP, Ebert JC, Nilsen GB, Baccash J, Halpern AL, Church GM, Drmanac R: **Accurate whole-genome sequencing and haplotyping from 10 to 20 human cells.** *Nature* 2012, **487**:190–195. doi:10.1038/nature11236.

**doi:10.1186/1471-2164-14-363**
**Cite this article as:** Prado-Martinez *et al.*: **The genome sequencing of an albino Western lowland gorilla reveals inbreeding in the wild.** *BMC Genomics* 2013, **14**:363.

# BIBLIOGRAPHY


ADAMS, J. and LAWLER, J. (1993). "Extracellular matrix: the thrombospondin family." Curr Biol 3(3): 188-190.

ADAMS, J. and LAWLER, J. (2011). "The thrombospondins." Cold Spring Harbor perspectives in biology 3(10).

ADAMS, J.C. (2004). "Functions of the conserved thrombospondin carboxy-terminal cassette in cell-extracellular matrix interactions and signaling." Int J Biochem Cell Biol 36(6): 1102-1114.

ADAMS, J.C. and LAWLER, J. (2004). "The thrombospondins." Int J Biochem Cell Biol 36(6): 961-968.

ADAMS, J.C., MONK, R., TAYLOR, A.L., OZBEK, S., FASCETTI, N., BAUMGARTNER, S. and ENGEL, J. (2003). "Characterisation of Drosophila thrombospondin defines an early origin of pentameric thrombospondins." J Mol Biol 328(2): 479-494.

ADAMS, M.D., KELLEY, J.M., GOCAYNE, J.D., DUBNICK, M., POLYMEROPOULOS, M.H., XIAO, H., . . . ET AL. (1991). "Complementary DNA sequencing: expressed sequence tags and human genome project." Science 252(5013): 1651-1656.

AGGARWAL, A., HUNTER, W.J., 3RD, AGGARWAL, H., SILVA, E.D., DAVEY, M.S., MURPHY, R.F. and AGRAWAL, D.K. (2010). "Expression of leukemia/lymphoma-related factor (LRF/POKEMON) in human breast carcinoma and other cancers." Exp Mol Pathol 89(2): 140-148.

AIELLO, L. and WHEELER, P. (1995). "The expensive tissue hypothesis." Curr Anthropol 36: 199-221.

ALTHAMMER, S., GONZALEZ-VALLINAS, J., BALLARE, C., BEATO, M. and EYRAS, E. (2011). "Pyicos: a versatile toolkit for the analysis of high-throughput sequencing data." Bioinformatics 27(24): 3333-3340.

ANDERSSON, R., VALEN, E., FORREST, A., CARNINCI, P., DAUB, C., SUZUKI, H., . . . SANDELIN, A. (2012). A MAP OF TRANSCRIBED ENHANCERS ACROSS HUMAN PRIMARY CELLS AND TISSUES. The biology of genomes. Cold Spring Harbor Laboratory, Cold Sping Harbor.

ANDRES, A.M., HUBISZ, M.J., INDAP, A., TORGERSON, D.G., DEGENHARDT, J.D., BOYKO, A.R., . . . NIELSEN, R. (2009). "Targets of balancing selection in the human genome." Mol Biol Evol 26(12): 2755-2764.

ARBER, S. and CARONI, P. (1995). "Thrombospondin-4, an extracellular matrix protein expressed in the developing and adult nervous system promotes neurite outgrowth." J Cell Biol 131(4): 1083-1094.

ARMSTRONG, E., SCHLEICHER, A., OMRAN, H., CURTIS, M. and ZILLES, K. (1995). "The ontogeny of human gyrification." Cereb Cortex 5(1): 56-63.

ARORA, G., MEZENCEV, R. and MCDONALD, J.F. (2012). "Human cells display reduced apoptotic function relative to chimpanzee cells." PLoS One 7(9): e46182.

ARORA, G., POLAVARAPU, N. and MCDONALD, J.F. (2009). "Did natural selection for increased cognitive ability in humans lead to an elevated risk of cancer?" Med Hypotheses 73(3): 453-456.

AUDAS, T.E., LI, Y., LIANG, G. and LU, R. (2008). "A novel protein, Luman/CREB3 recruitment factor, inhibits Luman activation of the unfolded protein response." Mol Cell Biol 28(12): 3952-3966.

AUTISM GENOME PROJECT, C., SZATMARI, P., PATERSON, A.D., ZWAIGENBAUM, L., ROBERTS, W., BRIAN, J., . . . MEYER, K.J. (2007). "Mapping autism risk loci using genetic linkage and chromosomal rearrangements." Nat Genet 39(3): 319-328.

BABBITT, C.C., FEDRIGO, O., PFEFFERLE, A.D., BOYLE, A.P., HORVATH, J.E., FUREY, T.S. and WRAY, G.A. (2010). "Both noncoding and protein-coding RNAs contribute to gene expression evolution in the primate brain." Genome Biol Evol 2: 67-79.

BAENZIGER, N.L., BRODIE, G.N. and MAJERUS, P.W. (1971). "A thrombin-sensitive protein of human platelet membranes." Proc Natl Acad Sci U S A 68(1): 240-243.

BAIRD, G.S., ZACHARIAS, D.A. and TSIEN, R.Y. (2000). "Biochemistry, mutagenesis, and oligomerization of DsRed, a red fluorescent protein from coral." Proc Natl Acad Sci U S A 97(22): 11984-11989.

BAKER, M. (2012). "Structural variation: the genome's hidden architecture." Nat Methods 9(2): 133-137.

BANERJI, J., RUSCONI, S. and SCHAFFNER, W. (1981). "Expression of a beta-globin gene is enhanced by remote SV40 DNA sequences." Cell 27(2 Pt 1): 299-308.

BARRES, B.A. and RAFF, M.C. (1999). "Axonal control of oligodendrocyte development." J Cell Biol 147(6): 1123-1128.

BARSKI, A., CUDDAPAH, S., CUI, K., ROH, T.Y., SCHONES, D.E., WANG, Z., . . . ZHAO, K. (2007). "High-resolution profiling of histone methylations in the human genome." Cell 129(4): 823-837.

BAUER, A., LEIKAM, D., KRINNER, S., NOTKA, F., LUDWIG, C., L√§NGST, G. and WAGNER, R. (2010). "The impact of intragenic CpG content on gene expression." Nucleic acids research 38(12): 3891-3908.

BEDFORD, D.C., KASPER, L.H., FUKUYAMA, T. and BRINDLE, P.K. (2010). "Target gene context influences the transcriptional requirement for the KAT3 family of CBP and p300 histone acetyltransferases." Epigenetics 5(1): 9-15.

BEER, M.A. and TAVAZOIE, S. (2004). "Predicting gene expression from sequence." Cell 117(2): 185-198.

BEKPEN, C., TASTEKIN, I., SISWARA, P., AKDIS, C.A. and EICHLER, E.E. (2012). "Primate segmental duplication creates novel promoters for the LRRC37 gene family within the 17q21.31 inversion polymorphism region." Genome Res 22(6): 1050-1058.

BELL, D., BELL, A., ROBERTS, D., WEBER, R.S. and EL-NAGGAR, A.K. (2012). "Developmental transcription factor EN1--a novel biomarker in human salivary gland adenoid cystic carcinoma." Cancer 118(5): 1288-1292.

BENIASHVILI, D.S. (1989). "An overview of the world literature on spontaneous tumors in nonhuman primates." J Med Primatol 18(6): 423-437.

BENNER, E.J., LUCIANO, D., JO, R., ABDI, K., PAEZ-GONZALEZ, P., SHENG, H., . . . KUO, C.T. (2013). "Protective astrogenesis from the SVZ niche after injury is controlled by Notch modulator Thbs4." Nature 497(7449): 369-373.

BENTLEY, A.A. and ADAMS, J.C. (2010). "The evolution of thrombospondins and their ligand-binding activities." Mol Biol Evol 27(9): 2187-2197.

BERCHTOLD, N.C., CRIBBS, D.H., COLEMAN, P.D., ROGERS, J., HEAD, E., KIM, R., . . . COTMAN, C.W. (2008). "Gene expression changes in the course of normal brain aging are sexually dimorphic." Proc Natl Acad Sci U S A 105(40): 15605-15610.

BEREZIKOV, E., THUEMMLER, F., VAN LAAKE, L.W., KONDOVA, I., BONTROP, R., CUPPEN, E. and PLASTERK, R.H. (2006). "Diversity of microRNAs in human and chimpanzee brain." Nat Genet.

BERNSTEIN, B.E., MEISSNER, A. and LANDER, E.S. (2007). "The mammalian epigenome." Cell 128(4): 669-681.

BERSAGLIERI, T., SABETI, P.C., PATTERSON, N., VANDERPLOEG, T., SCHAFFNER, S.F., DRAKE, J.A., . . . HIRSCHHORN, J.N. (2004). "Genetic signatures of strong recent positive selection at the lactase gene." Am J Hum Genet 74(6): 1111-1120.

BIANCHI, S., STIMPSON, C.D., BAUERNFEIND, A.L., SCHAPIRO, S.J., BAZE, W.B., MCARTHUR, M.J., . . . SHERWOOD, C.C. (2012). "Dendritic Morphology of Pyramidal Neurons in the Chimpanzee Neocortex: Regional Specializations and Comparison to Humans." Cereb Cortex.

BIRD, A. (2002). "DNA methylation patterns and epigenetic memory." Genes Dev 16(1): 6-21.

BIRNEY, E., STAMATOYANNOPOULOS, J.A., DUTTA, A., GUIGO, R., GINGERAS, T.R., MARGULIES, E.H., . . . DE JONG, P.J. (2007). "Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project." Nature 447(7146): 799-816.

BLACKWOOD, E.M. and KADONAGA, J.T. (1998). "Going the distance: a current view of enhancer action." Science 281(5373): 60-63.

BLANCHETTE, M. (2012). "Exploiting ancestral mammalian genomes for the prediction of human transcription factor binding sites." BMC Bioinformatics 13 Suppl 19: S2.

BLENCOWE, B.J. (2006). "Alternative splicing: new insights from global analyses." Cell 126(1): 37-47.

BLOW, M.J., MCCULLEY, D.J., LI, Z., ZHANG, T., AKIYAMA, J.A., HOLT, A., . . . PENNACCHIO, L.A. (2010). "ChIP-Seq identification of weakly conserved heart enhancers." Nat Genet 42(9): 806-810.

BOESCH, C. (2007). "What makes us human (Homo sapiens)? The challenge of cognitive cross-species comparison." J Comp Psychol 121(3): 227-240.

BORLIKOVA, G.G., TREJO, M., MABLY, A.J., MC DONALD, J.M., SALA FRIGERIO, C., REGAN, C.M., . . . WALSH, D.M. (2013). "Alzheimer brain-derived amyloid beta-protein impairs synaptic remodeling and memory consolidation." Neurobiol Aging 34(5): 1315-1327.

BORNSTEIN, P. (2009). "Thrombospondins function as regulators of angiogenesis." J Cell Commun Signal 3(3-4): 189-200.

BOYLE, A.P., DAVIS, S., SHULHA, H.P., MELTZER, P., MARGULIES, E.H., WENG, Z., . . . CRAWFORD, G.E. (2008). "High-resolution mapping and characterization of open chromatin across the genome." Cell 132(2): 311-322.

BRITTEN, R.J. (2010). "Transposable element insertions have strongly affected human evolution." Proc Natl Acad Sci U S A 107(46): 19945-19948.

BRODMANN, K. (1912). "Neue Ergebnisse über die vergleichende histologische localisation der grosshirnrinde mit besonderer beru cksichtigung des stirnhirns." Anat. Anz 41 (suppl.), : 157–216

BRUNETTI-PIERRI, N., BERG, J.S., SCAGLIA, F., BELMONT, J., BACINO, C.A., SAHOO, T., . . . PATEL, A. (2008). "Recurrent reciprocal 1q21.1 deletions and duplications associated with microcephaly or macrocephaly and developmental and behavioral abnormalities." Nat Genet 40(12): 1466-1471.

BUCHER, P. (1990). "Weight matrix descriptions of four eukaryotic RNA polymerase II promoter elements derived from 502 unrelated promoter sequences." J Mol Biol 212(4): 563-578.

BUÉE, L., HOF, P.R., ROBERTS, D.D., DELACOURTE, A., MORRISON, J.H. and FILLIT, H.M. (1992). "Immunohistochemical identification of thrombospondin in normal human brain and in Alzheimer's disease." Am J Pathol 141(4): 783-788.

BULLMORE, E. and SPORNS, O. (2012). "The economy of brain network organization." Nat Rev Neurosci 13(5): 336-349.

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." Proc Natl Acad Sci U S A 100(22): 13030-13035.

CÁCERES, M., SUWYN, C., MADDOX, M., THOMAS, J.W. and PREUSS, T.M. (2006). "Increased Cortical Expression of Two Synaptogenic Thrombospondins in Human Brain Evolution." Cereb Cortex 17: 2312-2321.

CÁCERES, M., SUWYN, C., MADDOX, M., THOMAS, J.W. and PREUSS, T.M. (2007). "Increased cortical expression of two synaptogenic thrombospondins in human brain evolution." Cereb Cortex 17(10): 2312-2321.

CAGLIANI, R., FUMAGALLI, M., RIVA, S., POZZOLI, U., COMI, G.P., MENOZZI, G., . . . SIRONI, M. (2008). "The signature of long-standing balancing selection at the human defensin beta-1 promoter." Genome Biol 9(9): R143.

CAGLIANI, R., GUERINI, F.R., RUBIO-ACERO, R., BAGLIO, F., FORNI, D., AGLIARDI, C., . . . SIRONI, M. (2013). "Long-Standing Balancing Selection in the THBS4 Gene: Influence on Sex-Specific Brain Expression and Gray Matter Volumes in Alzheimer Disease." Hum Mutat 34(5): 743-753.

CAHOY, J.D., EMERY, B., KAUSHAL, A., FOO, L.C., ZAMANIAN, J.L., CHRISTOPHERSON, K.S., . . . BARRES, B.A. (2008). "A transcriptome database for astrocytes, neurons, and oligodendrocytes: a new resource for understanding brain development and function." J Neurosci 28(1): 264-278.

CAI, H.N., ZHANG, Z., ADAMS, J.R. and SHEN, P. (2001). "Genomic context modulates insulator activity through promoter competition." Development 128(21): 4339-4347.

CALARCO, J.A., XING, Y., CÁCERES, M., CALARCO, J.P., XIAO, X., PAN, Q., . . . BLENCOWE, B.J. (2007). "Global analysis of alternative splicing differences between humans and chimpanzees." Genes Dev: doi:10.1101/gad.1606907.

CAMPBELL, R.E., TOUR, O., PALMER, A.E., STEINBACH, P.A., BAIRD, G.S., ZACHARIAS, D.A. and TSIEN, R.Y. (2002). "A monomeric red fluorescent protein." Proc Natl Acad Sci U S A 99(12): 7877-7882.

CANTALUPO, C. and HOPKINS, W.D. (2001). "Asymmetric Broca's area in great apes." Nature 414(6863): 505.

CARLSON, C.B., LAWLER, J. and MOSHER, D.F. (2008). "Structures of thrombospondins." Cell Mol Life Sci 65(5): 672-686.

CARNINCI, P., SANDELIN, A., LENHARD, B., KATAYAMA, S., SHIMOKAWA, K., PONJAVIC, J., . . . HAYASHIZAKI, Y. (2006). "Genome-wide analysis of mammalian promoter architecture and evolution." Nat Genet 38(6): 626-635.

CARROLL, S.B. (2003). "Genetics and the making of *Homo sapiens*." Nature 422(6934): 849-857.

CARROLL, S.B. (2005). "Evolution at two levels: on genes and form." PLoS Biol 3(7): e245.

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

CARTER, D., CHAKALOVA, L., OSBORNE, C.S., DAI, Y.F. and FRASER, P. (2002). "Long-range chromatin regulatory interactions in vivo." Nat Genet 32(4): 623-626.

CHAN, Y.F., MARKS, M.E., JONES, F.C., VILLARREAL, G., JR., SHAPIRO, M.D., BRADY, S.D., . . . KINGSLEY, D.M. (2010). "Adaptive evolution of pelvic reduction in sticklebacks by recurrent deletion of a Pitx1 enhancer." Science 327(5963): 302-305.

CHARLESWORTH, D. (2006). "Balancing selection and its effects on sequences in nearby genome regions." PLoS Genet 2(4): e64.

CHENDRIMADA, T.P., FINN, K.J., JI, X., BAILLAT, D., GREGORY, R.I., LIEBHABER, S.A., . . . SHIEKHATTAR, R. (2007). "MicroRNA silencing through RISC recruitment of eIF6." Nature 447(7146): 823-828.

CHRISTOPHERSON, K.S., ULLIAN, E.M., STOKES, C.C., MULLOWNEY, C.E., HELL, J.W., AGAH, A., . . . BARRES, B.A. (2005). "Thrombospondins are astrocyte-secreted proteins that promote CNS synaptogenesis." Cell 120(3): 421-433.

CIOFFI, F., LANNI, A. and GOGLIA, F. (2010). "Thyroid hormones, mitochondrial bioenergetics and lipid handling." Curr Opin Endocrinol Diabetes Obes 17(5): 402-407.

CLARK, A.G., GLANOWSKI, S., NIELSEN, R., THOMAS, P.D., KEJARIWAL, A., TODD, M.A., . . . CARGILL, M. (2003). "Inferring nonneutral evolution from human-chimp-mouse orthologous gene trios." Science 302(5652): 1960-1963.

COLANTUONI, C., LIPSKA, B.K., YE, T., HYDE, T.M., TAO, R., LEEK, J.T., . . . KLEINMAN, J.E. (2011). "Temporal dynamics and genetic control of transcription in the human prefrontal cortex." Nature 478(7370): 519-523.

CONSORTIUM, C.S.A.A. (2005). "Initial sequence of the chimpanzee genome and comparison with the human genome." Nature 437(7055): 69-87.

CONSORTIUM, E.P., BIRNEY, E., STAMATOYANNOPOULOS, J.A., DUTTA, A., GUIGO, R., GINGERAS, T.R., . . . DE JONG, P.J. (2007). "Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project." Nature 447(7146): 799-816.

CONSORTIUM, I.H.G.S. (2004). "Finishing the euchromatic sequence of the human genome." Nature 431(7011): 931-945.

CONSORTIUM, T.E.P. (2004). "The ENCODE (ENCyclopedia Of DNA Elements) Project." Science 306(5696): 636-640.

COOPER, S.J., TRINKLEIN, N.D., ANTON, E.D., NGUYEN, L. and MYERS, R.M. (2006). "Comprehensive analysis of transcriptional promoter structure and function in 1% of the human genome." Genome Res 16(1): 1-10.

CORBALLIS, M.C. (2009). "The evolution and genetics of cerebral asymmetry." Philos Trans R Soc Lond B Biol Sci 364(1519): 867-879.

CORSETTI, J.P., RYAN, D., MOSS, A.J., MCCARTHY, J., GOLDENBERG, I., ZAREBA, W. and SPARKS, C.E. (2011). "Thrombospondin-4 polymorphism (A387P) predicts cardiovascular risk in postinfarction patients with high HDL cholesterol and C-reactive protein levels." Thromb Haemost 106(6): 1170-1178.

COSTELLO, J.F. and PLASS, C. (2001). "Methylation matters." J Med Genet 38(5): 285-303.

COWLEY, M. and OAKEY, R.J. (2013). "Transposable elements re-wire and fine-tune the transcriptome." PLoS Genet 9(1): e1003234.

CRESPI, B.J. and CROFTS, H.J. (2012). "Association testing of copy number variants in schizophrenia and autism spectrum disorders." J Neurodev Disord 4(1): 15.

CUI, J., RANDELL, E., RENOUF, J., SUN, G., GREEN, R., HAN, F.Y. and XIE, Y.G. (2006). "Thrombospondin-4 1186G>C (A387P) is a sex-dependent risk factor for myocardial infarction: a large replication study with increased sample size from the same population." Am Heart J 152(3): 543 e541-545.

CUI, J., RANDELL, E., RENOUF, J., SUN, G., HAN, F.Y., YOUNGHUSBAND, B. and XIE, Y.G. (2004). "Gender dependent association of thrombospondin-4 A387P polymorphism with myocardial infarction." Arterioscler Thromb Vasc Biol 24(11): e183-184.

DARWIN, C. (1859). On the origin of the species by means of natural selection, or, The preservation of favoured races in the struggle for life. London,, J. Murray.

DEGNER, J.F., PAI, A.A., PIQUE-REGI, R., VEYRIERAS, J.B., GAFFNEY, D.J., PICKRELL, J.K., . . . PRITCHARD, J.K. (2012). "DNase I sensitivity QTLs are a major determinant of human expression variation." Nature 482(7385): 390-394.

DERMITZAKIS, E.T., REYMOND, A. and ANTONARAKIS, S.E. (2005). "Conserved non-genic sequences - an unexpected feature of mammalian genomes." Nat Rev Genet 6(2): 151-157.

DEUTZMANN, R. (2004). "Structural characterization of proteins and peptides." Methods Mol Med 94: 269-297.

DISKIN, S.J., HOU, C., GLESSNER, J.T., ATTIYEH, E.F., LAUDENSLAGER, M., BOSSE, K., . . . MARIS, J.M. (2009). "Copy number variation at 1q21.1 associated with neuroblastoma." Nature 459(7249): 987-991.

DOWELL, R.D. (2010). "Transcription factor binding variation in the evolution of gene regulation." Trends Genet 26(11): 468-475.

DUMAS, L., KIM, Y.H., KARIMPOUR-FARD, A., COX, M., HOPKINS, J., POLLACK, J.R. and SIKELA, J.M. (2007). "Gene copy number variation spanning 60 million years of human and primate evolution." Genome Res 17(9): 1266-1277.

DUMAS, L. and SIKELA, J.M. (2009). "DUF1220 domains, cognitive disease, and human brain evolution." Cold Spring Harb Symp Quant Biol 74: 375-382.

DUMAS, L.J., O'BLENESS, M.S., DAVIS, J.M., DICKENS, C.M., ANDERSON, N., KEENEY, J.G., . . . SIKELA, J.M. (2012). "DUF1220-domain copy number implicated in human brain-size pathology and evolution." Am J Hum Genet 91(3): 444-454.

DUONG, T.Q. (2010). "Diffusion tensor and perfusion MRI of non-human primates." Methods 50(3): 125-135.

EHRICH, M., ZOLL, S., SUR, S. and VAN DEN BOOM, D. (2007). "A new method for accurate assessment of DNA quality after bisulfite treatment." Nucleic Acids Res 35(5): e29.

ELSTON, G.N. (2003). "Cortex, cognition and the cell: new insights into the pyramidal neuron and prefrontal function." Cereb Cortex 13(11): 1124-1138.

ELSTON, G.N., BENAVIDES-PICCIONE, R., ELSTON, A., ZIETSCH, B., DEFELIPE, J., MANGER, P., . . . KAAS, J.H. (2006). "Specializations of the granular prefrontal cortex of primates: implications for cognitive processing." Anat Rec A Discov Mol Cell Evol Biol 288(1): 26-35.

ENARD, W., GEHRE, S., HAMMERSCHMIDT, K., HOLTER, S.M., BLASS, T., SOMEL, M., . . . PAABO, S. (2009). "A humanized version of Foxp2 affects cortico-basal ganglia circuits in mice." Cell 137(5): 961-971.

ENARD, W., KHAITOVICH, P., KLOSE, J., ZOLLNER, S., HEISSIG, F., GIAVALISCO, P., . . . PAABO, S. (2002). "Intra- and interspecific variation in primate gene expression patterns." Science 296(5566): 340-343.

ENARD, W., PRZEWORSKI, M., FISHER, S.E., LAI, C.S., WIEBE, V., KITANO, T., . . . PAABO, S. (2002). "Molecular evolution of FOXP2, a gene involved in speech and language." Nature 418(6900): 869-872.

ENATTAH, N.S., SAHI, T., SAVILAHTI, E., TERWILLIGER, J.D., PELTONEN, L. and JARVELA, I. (2002). "Identification of a variant associated with adult-type hypolactasia." Nat Genet 30(2): 233-237.

EPSTEIN, D.J. (2009). "Cis-regulatory mutations in human disease." Brief Funct Genomic Proteomic 8(4): 310-316.

ERNST, J. and KELLIS, M. (2010). "Discovery and characterization of chromatin states for systematic annotation of the human genome." Nature biotechnology 28(8): 817-825.

ERNST, J., KHERADPOUR, P., MIKKELSEN, T., SHORESH, N., WARD, L., EPSTEIN, C., . . . BERNSTEIN, B. (2011). "Mapping and analysis of chromatin state dynamics in nine human cell types." Nature 473(7345): 43-49.

EROGLU, C. (2009). "The role of astrocyte-secreted matricellular proteins in central nervous system development and function." J Cell Commun Signal 3(3-4): 167-176.

EROGLU, C., ALLEN, N.J., SUSMAN, M.W., O'ROURKE, N.A., PARK, C.Y., OZKAN, E., . . . BARRES, B.A. (2009). "Gabapentin receptor alpha2delta-1 is a neuronal thrombospondin receptor responsible for excitatory CNS synaptogenesis." Cell 139(2): 380-392.

ESTECIO, M.R. and ISSA, J.P. (2011). "Dissecting DNA hypermethylation in cancer." FEBS Lett 585(13): 2078-2086.

FALK, D. (2012). "Hominin paleoneurology: where are we now?" Prog Brain Res 195: 255-272.

FANG, X., HAN, H., STAMATOYANNOPOULOS, G. and LI, Q. (2004). "Developmentally specific role of the CCAAT box in regulation of human gamma-globin gene expression." J Biol Chem 279(7): 5444-5449.

FARAJOLLAHI, S. and MAAS, S. (2010). "Molecular diversity through RNA editing: a balancing act." Trends Genet 26(5): 221-230.

FAZIUS, E., SHELEST, V. and SHELEST, E. (2011). "SiTaR: a novel tool for transcription factor binding site prediction." Bioinformatics 27(20): 2806-2811.

FEHRMANN, R.S., JANSEN, R.C., VELDINK, J.H., WESTRA, H.J., ARENDS, D., BONDER, M.J., . . . FRANKE, L. (2011). "Trans-eQTLs reveal that independent genetic variants associated with a complex phenotype converge on intermediate genes, with a major role for the HLA." PLoS Genet 7(8): e1002197.

FELDMAN, D.E. (2009). "Synaptic mechanisms for plasticity in neocortex." Annu Rev Neurosci 32: 33-55.

FERLAY, J., SHIN, H.R., BRAY, F., FORMAN, D., MATHERS, C. and PARKIN, D.M. (2010). "Estimates of worldwide burden of cancer in 2008: GLOBOCAN 2008." Int J Cancer 127(12): 2893-2917.

FEUK, L., MACDONALD, J.R., TANG, T., CARSON, A.R., LI, M., RAO, G., . . . SCHERER, S.W. (2005). "Discovery of human inversion polymorphisms by comparative analysis of human and chimpanzee DNA sequence assemblies." PLoS Genet 1(4): e56.

FINLAY, B.L., DARLINGTON, R.B. and NICASTRO, N. (2001). "Developmental structure in brain evolution." Behav Brain Sci 24(2): 263-278; discussion 278-308.

FIRTH, H.V., RICHARDS, S.M., BEVAN, A.P., CLAYTON, S., CORPAS, M., RAJAN, D., . . . CARTER, N.P. (2009). "DECIPHER: Database of Chromosomal Imbalance and Phenotype in Humans Using Ensembl Resources." Am J Hum Genet 84(4): 524-533.

FLACK, J.C. and DE WAAL, F.B.M. (2000). "'Any animal whatever'. Darwinian building blocks of morality in monkeys and apes." Journal of Consciousness Studies Volumen 7, Numbers 1-2, 2000 , pp. 1-29(29).

FORTNA, A., KIM, Y., MACLAREN, E., MARSHALL, K., HAHN, G., MELTESEN, L., . . . SIKELA, J.M. (2004). "Lineage-specific gene duplication and loss in human and great ape evolution." PLoS Biol 2(7): E207.

FOUNDAS, A.L., EURE, K.F., LUEVANO, L.F. and WEINBERGER, D.R. (1998). "MRI asymmetries of Broca's area: the pars triangularis and pars opercularis." Brain Lang 64(3): 282-296.

FRANGOGIANNIS, N.G., REN, G., DEWALD, O., ZYMEK, P., HAUDEK, S., KOERTING, A., . . . ENTMAN, M.L. (2005). "Critical role of endogenous thrombospondin-1 in preventing expansion of healing myocardial infarcts." Circulation 111(22): 2935-2942.

FRANKEL, N., EREZYILMAZ, D.F., MCGREGOR, A.P., WANG, S., PAYRE, F. and STERN, D.L. (2011). "Morphological evolution caused by many subtle-effect substitutions in regulatory DNA." Nature 474(7353): 598-603.

FROLOVA, E.G., SOPKO, N., BLECH, L., POPOVIC, Z.B., LI, J., VASANJI, A., . . . STENINA, O.I. (2012). "Thrombospondin-4 regulates fibrosis and remodeling of the myocardium in response to pressure overload." FASEB J 26(6): 2363-2373.

FUSTER-MATANZO, A., LLORENS-MARTIN, M., HERNANDEZ, F. and AVILA, J. (2013). "Role of neuroinflammation in adult neurogenesis and Alzheimer disease: therapeutic approaches." Mediators Inflamm 2013: 260925.

GANNON, P.J., HOLLOWAY, R.L., BROADFIELD, D.C. and BRAUN, A.R. (1998). "Asymmetry of chimpanzee planum temporale: humanlike pattern of Wernicke's brain language area homolog." Science 279(5348): 220-222.

GASZNER, M. and FELSENFELD, G. (2006). "Insulators: exploiting transcriptional and epigenetic mechanisms." Nat Rev Genet 7(9): 703-713.

GAZAVE, E., DARRE, F., MORCILLO-SUAREZ, C., PETIT-MARTY, N., CARRENO, A., MARIGORTA, U.M., . . . NAVARRO, A. (2011). "Copy number variation analysis in the great apes reveals species-specific patterns of structural variation." Genome Res 21(10): 1626-1639.

GEIMAN, T.M. and MUEGGE, K. (2010). "DNA methylation in early development." Mol Reprod Dev 77(2): 105-113.

GEORGES, A.B., BENAYOUN, B.A., CABURET, S. and VEITIA, R.A. (2010). "Generic binding sites, generic DNA-binding domains: where does specific promoter recognition come from?" FASEB J 24(2): 346-356.

GERLO, S., DAVIS, J.R., MAGER, D.L. and KOOIJMAN, R. (2006). "Prolactin in man: a tale of two promoters." Bioessays 28(10): 1051-1055.

GESCHWIND, N. and LEVITSKY, W. (1968). "Human brain: left-right asymmetries in temporal speech region." Science 161(3837): 186-187.

GEYER, P.K., GREEN, M.M. and CORCES, V.G. (1990). "Tissue-specific transcriptional enhancers may act in trans on the gene located in the homologous chromosome: the molecular basis of transvection in Drosophila." EMBO J 9(7): 2247-2256.

GHIRLANDO, R., GILES, K., GOWHER, H., XIAO, T., XU, Z., YAO, H. and FELSENFELD, G. (2012). "Chromatin domains, insulators, and the regulation of gene expression." Biochim Biophys Acta 1819(7): 644-651.

GILAD, Y., OSHLACK, A., SMYTH, G.K., SPEED, T.P. and WHITE, K.P. (2006). "Expression profiling in primates reveals a rapid evolution of human transcription factors." Nature 440(7081): 242-245.

GIRALDEZ, A.J., MISHIMA, Y., RIHEL, J., GROCOCK, R.J., VAN DONGEN, S., INOUE, K., . . . SCHIER, A.F. (2006). "Zebrafish MiR-430 promotes deadenylation and clearance of maternal mRNAs." Science 312(5770): 75-79.

GOMEZ-ROBLES, A., HOPKINS, W.D. and SHERWOOD, C.C. (2013). "Increased morphological asymmetry, evolvability and plasticity in human brain evolution." Proc Biol Sci 280(1761): 20130575.

GOULD, S.J. (1977). Ontogeny and phylogeny. Cambridge, Mass., Belknap Press of Harvard University Press.

GOULD, S.J. and SUBRAMANI, S. (1988). "Firefly luciferase as a tool in molecular and cell biology." Anal Biochem 175(1): 5-13.

GRALLE, M. and PAABO, S. (2011). "A comprehensive functional analysis of ancestral human signal peptides." Mol Biol Evol 28(1): 25-28.

GRECO, S.A., CHIA, J., INGLIS, K.J., COZZI, S.J., RAMSNES, I., BUTTENSHAW, R.L., . . . WHITEHALL, V.L. (2010). "Thrombospondin-4 is a putative tumour-suppressor gene in colorectal cancer that exhibits age-related methylation." BMC Cancer 10: 494.

GU, J. and GU, X. (2003). "Induced gene expression in human brain after the split from chimpanzee." Trends Genet 19(2): 63-65.

HAMBLIN, M.T. and DI RIENZO, A. (2000). "Detection of the signature of natural selection in humans: evidence from the Duffy blood group locus." Am J Hum Genet 66(5): 1669-1679.

HAMPSHIRE, A.J., RUSLING, D.A., BROUGHTON-HEAD, V.J. and FOX, K.R. (2007). "Footprinting: a method for determining the sequence selectivity, affinity and kinetics of DNA-binding ligands." Methods 42(2): 128-140.

HARLOW, H.F. and HARLOW, M. (1962). "Social deprivation in monkeys." Sci Am 207: 136-146.

HE, X., LING, X. and SINHA, S. (2009). "Alignment and prediction of cis-regulatory modules based on a probabilistic model of evolution." PLoS Comput Biol 5(3): e1000299.

HEINTZMAN, N., HON, G., HAWKINS, R., KHERADPOUR, P., STARK, A., HARP, L., . . . REN, B. (2009). "Histone modifications at human enhancers reflect global cell-type-specific gene expression." Nature 459(7243): 108-112.

HEINTZMAN, N., STUART, R., HON, G., FU, Y., CHING, C., HAWKINS, R., . . . REN, B. (2007). "Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome." Nature genetics 39(3): 311-318.

HEINTZMAN, N.D., STUART, R.K., HON, G., FU, Y., CHING, C.W., HAWKINS, R.D., . . . REN, B. (2007). "Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome." Nat Genet 39(3): 311-318.

HERCULANO-HOUZEL, S. (2012). "The remarkable, yet not extraordinary, human brain as a scaled-up primate brain and its associated cost." Proc Natl Acad Sci U S A 109 Suppl 1: 10661-10668.

HERMAN, J.G. and BAYLIN, S.B. (2003). "Gene silencing in cancer in association with promoter hypermethylation." N Engl J Med 349(21): 2042-2054.

HERNANDEZ, N. (1993). "TBP, a universal eukaryotic transcription factor?" Genes Dev 7(7B): 1291-1308.

HEUER, E., ROSEN, R.F., CINTRON, A. and WALKER, L.C. (2012). "Nonhuman primate models of Alzheimer-like cerebral proteopathy." Curr Pharm Des 18(8): 1159-1169.

HILLER, M., SCHAAR, B.T. and BEJERANO, G. (2012). "Hundreds of conserved non-coding genomic regions are independently lost in mammals." Nucleic Acids Res 40(22): 11463-11476.

HOEKSTRA, H.E. and COYNE, J.A. (2007). "The locus of evolution: evo devo and the genetics of adaptation." Evolution 61(5): 995-1016.

HOFFMAN, G.E. and LE, W.W. (2004). "Just cool it! Cryoprotectant anti-freeze in immunocytochemistry and in situ hybridization." Peptides 25(3): 425-431.

HOLLOWAY, R.L. (1983). "Human paleontological evidence relevant to language behavior." Hum Neurobiol 2(3): 105-114.

HOPKINS, W.D., MARINO, L., RILLING, J.K. and MACGREGOR, L.A. (1998). "Planum temporale asymmetries in great apes as revealed by magnetic resonance imaging (MRI)." Neuroreport 9(12): 2913-2918.

HOPKINS, W.D. and PILCHER, D.L. (2001). "Neuroanatomical localization of the motor hand area with magnetic resonance imaging: the left hemisphere is larger in great apes." Behav Neurosci 115(5): 1159-1164.

HORNER, V., CARTER, J.D., SUCHAK, M. and DE WAAL, F.B. (2011). "Spontaneous prosocial choice by chimpanzees." Proc Natl Acad Sci U S A 108(33): 13847-13851.

HOUSTON, I., PETER, C.J., MITCHELL, A., STRAUBHAAR, J., ROGAEV, E. and AKBARIAN, S. (2013). "Epigenetics in the human brain." Neuropsychopharmacology 38(1): 183-197.

HU, H., GUO, S., XI, J., YAN, Z., FU, N., ZHANG, X., . . . KHAITOVICH, P. (2011). "MicroRNA expression and regulation in human, chimpanzee, and macaque brains." PLoS genetics 7(10).

HUMPHREYS, D.T., WESTMAN, B.J., MARTIN, D.I. and PREISS, T. (2005). "MicroRNAs control translation initiation by inhibiting eukaryotic initiation factor 4E/cap and poly(A) tail function." Proc Natl Acad Sci U S A 102(47): 16961-16966.

IBRAHIM, A.E., THORNE, N.P., BAIRD, K., BARBOSA-MORAIS, N.L., TAVARE, S., COLLINS, V.P., . . . BRENTON, J.D. (2006). "MMASS: an optimized array-based method for assessing CpG island methylation." Nucleic Acids Res 34(20): e136.

INBAR-FEIGENBERG, M., CHOUFANI, S., BUTCHER, D.T., ROIFMAN, M. and WEKSBERG, R. (2013). "Basic concepts of epigenetics." Fertil Steril.

INTERNATIONAL HAPMAP, C. (2005). "A haplotype map of the human genome." Nature 437(7063): 1299-1320.

INTERNATIONAL SCHIZOPHRENIA, C. (2008). "Rare chromosomal deletions and duplications increase risk of schizophrenia." Nature 455(7210): 237-241.

JACOB, F. and MONOD, J. (1961). "Genetic regulatory mechanisms in the synthesis of proteins." J Mol Biol 3: 318-356.

JEONG, S., REBEIZ, M., ANDOLFATTO, P., WERNER, T., TRUE, J. and CARROLL, S.B. (2008). "The evolution of gene regulation underlies a morphological difference between two Drosophila sister species." Cell 132(5): 783-793.

JIN, V.X., SINGER, G.A., AGOSTO-PEREZ, F.J., LIYANARACHCHI, S. and DAVULURI, R.V. (2006). "Genome-wide analysis of core promoter elements from conserved human and mouse orthologous pairs." BMC Bioinformatics 7: 114.

JONES, P.A. and TAKAI, D. (2001). "The role of DNA methylation in mammalian epigenetics." Science 293(5532): 1068-1070.

JUGDUTT, B.I. (2003). "Ventricular remodeling after infarction and the extracellular collagen matrix: when is enough enough?" Circulation 108(11): 1395-1403.

KADONAGA, J.T. (2002). "The DPE, a core promoter element for transcription by RNA polymerase II." Exp Mol Med 34(4): 259-264.

KANDEL, E.R. and SQUIRE, L.R. (2000). "Neuroscience: breaking down scientific barriers to the study of brain and mind." Science 290(5494): 1113-1120.

KARAM, R., WENGROD, J., GARDNER, L.B. and WILKINSON, M.F. (2013). "Regulation of nonsense-mediated mRNA decay: implications for physiology and disease." Biochim Biophys Acta 1829(6-7): 624-633.

KAWAJI, H., KASUKAWA, T., FUKUDA, S., KATAYAMA, S., KAI, C., KAWAI, J., . . . HAYASHIZAKI, Y. (2006). "CAGE Basic/Analysis Databases: the CAGE resource for comprehensive promoter analysis." Nucleic Acids Res 34(Database issue): D632-636.

KELLER, S.S., ROBERTS, N. and HOPKINS, W. (2009). "A comparative magnetic resonance imaging study of the anatomy, variability, and asymmetry of Broca's area in the human and chimpanzee brain." J Neurosci 29(46): 14607-14616.

KHAITOVICH, P., ENARD, W., LACHMANN, M. and PAABO, S. (2006). "Evolution of primate gene expression." Nat Rev Genet 7(9): 693-702.

KHAITOVICH, P., HELLMANN, I., ENARD, W., NOWICK, K., LEINWEBER, M., FRANZ, H., . . . PAABO, S. (2005). "Parallel patterns of evolution in the genomes and transcriptomes of humans and chimpanzees." Science 309(5742): 1850-1854.

KHAITOVICH, P., MUETZEL, B., SHE, X., LACHMANN, M., HELLMANN, I., DIETZSCH, J., . . . PAABO, S. (2004). "Regional patterns of gene expression in human and chimpanzee brains." Genome Res 14(8): 1462-1473.

KHAMBATA-FORD, S., LIU, Y., GLEASON, C., DICKSON, M., ALTMAN, R.B., BATZOGLOU, S. and MYERS, R.M. (2003). "Identification of promoter regions in the human genome by using a retroviral plasmid library-based functional reporter gene assay." Genome Res 13(7): 1765-1774.

KHOURY, G. and GRUSS, P. (1983). "Enhancer elements." Cell 33(2): 313-314.

KIM, E., MAGEN, A. and AST, G. (2007). "Different levels of alternative splicing among eukaryotes." Nucleic Acids Res 35(1): 125-131.

KIMURA, K., WAKAMATSU, A., SUZUKI, Y., OTA, T., NISHIKAWA, T., YAMASHITA, R., . . . SUGANO, S. (2006). "Diversification of transcriptional modulation: large-scale identification and characterization of putative alternative promoters of human genes." Genome Res 16(1): 55-65.

KING, M.C. and WILSON, A.C. (1975). "Evolution at two levels in humans and chimpanzees." Science 188(4184): 107-116.

KLEINJAN, D.A. and VAN HEYNINGEN, V. (2005). "Long-range control of gene expression: emerging mechanisms and disruption in disease." Am J Hum Genet 76(1): 8-32.

KOCH, W., HOPPMANN, P., DE WAHA, A., SCHOMIG, A. and KASTRATI, A. (2008). "Polymorphisms in thrombospondin genes and myocardial infarction: a case-control study and a meta-analysis of available evidence." Hum Mol Genet 17(8): 1120-1126.

KODZIUS, R., KOJIMA, M., NISHIYORI, H., NAKAMURA, M., FUKUDA, S., TAGAMI, M., . . . CARNINCI, P. (2006). "CAGE: cap analysis of gene expression." Nat Methods 3(3): 211-222.

KONG, Y.W., CANNELL, I.G., DE MOOR, C.H., HILL, K., GARSIDE, P.G., HAMILTON, T.L., . . . BUSHELL, M. (2008). "The mechanism of micro-RNA-mediated translation repression is determined by the promoter of the target gene." Proc Natl Acad Sci U S A 105(26): 8866-8871.

KONOPKA, G., BOMAR, J.M., WINDEN, K., COPPOLA, G., JONSSON, Z.O., GAO, F., . . . GESCHWIND, D.H. (2009). "Human-specific transcriptional regulation of CNS development genes by FOXP2." Nature 462(7270): 213-217.

KONOPKA, G., FRIEDRICH, T., DAVIS-TURAK, J., WINDEN, K., OLDHAM, M.C., GAO, F., . . . GESCHWIND, D.H. (2012). "Human-specific transcriptional networks in the brain." Neuron 75(4): 601-617.

KORESSAAR, T. and REMM, M. (2007). "Enhancements and modifications of primer design program Primer3." Bioinformatics 23(10): 1289-1291.

KULAKOVSKIY, I.V., MEDVEDEVA, Y.A., SCHAEFER, U., KASIANOV, A.S., VORONTSOV, I.E., BAJIC, V.B. and MAKEEV, V.J. (2013). "HOCOMOCO: a comprehensive collection of human transcription factor binding sites models." Nucleic Acids Res 41(Database issue): D195-202.

LAGRANGE, T., KAPANIDIS, A.N., TANG, H., REINBERG, D. and EBRIGHT, R.H. (1998). "New core promoter element in RNA polymerase II-dependent transcription: sequence-specific DNA binding by transcription factor IIB." Genes Dev 12(1): 34-44.

LAI, C.S., FISHER, S.E., HURST, J.A., VARGHA-KHADEM, F. and MONACO, A.P. (2001). "A forkhead-domain gene is mutated in a severe speech and language disorder." Nature 413(6855): 519-523.

LANDER, E.S., LINTON, L.M., BIRREN, B., NUSBAUM, C., ZODY, M.C., BALDWIN, J., . . . CHEN, Y.J. (2001). "Initial sequencing and analysis of the human genome." Nature 409(6822): 860-921.

LANDRY, J.R., MAGER, D.L. and WILHELM, B.T. (2003). "Complex controls: the role of alternative promoters in mammalian genomes." Trends Genet 19(11): 640-648.

LAPPALAINEN, T. and DERMITZAKIS, E. (2010). "Evolutionary history of regulatory variation in human populations." Human molecular genetics 19(R2): 203.

LAWLER, J., DUQUETTE, M., URRY, L., MCHENRY, K. and SMITH, T.F. (1993). "The evolution of the thrombospondin gene family." J Mol Evol 36(6): 509-516.

LAWLER, J., DUQUETTE, M., WHITTAKER, C.A., ADAMS, J.C., MCHENRY, K. and DESIMONE, D.W. (1993). "Identification and characterization of thrombospondin-4, a new member of the thrombospondin gene family." J Cell Biol 120(4): 1059-1067.

LEE, B.M., BUCK-KOEHNTOP, B.A., MARTINEZ-YAMOUT, M.A., DYSON, H.J. and WRIGHT, P.E. (2007). "Embryonic neural inducing factor churchill is not a DNA-binding zinc finger protein: solution structure reveals a solvent-exposed beta-sheet and zinc binuclear cluster." J Mol Biol 371(5): 1274-1289.

LEE, T.I. and YOUNG, R.A. (2000). "Transcription of eukaryotic protein-coding genes." Annu Rev Genet 34: 77-137.

LENHARD, B., SANDELIN, A. and CARNINCI, P. (2012). "Metazoan promoters: emerging characteristics and insights into transcriptional regulation." Nature reviews. Genetics 13(4): 233-245.

LI, W.H., GU, Z., WANG, H. and NEKRUTENKO, A. (2001). "Evolutionary analyses of the human genome." Nature 409(6822): 847-849.

LIANG, H. and LI, W.H. (2009). "Lowly expressed human microRNA genes evolve rapidly." Mol Biol Evol 26(6): 1195-1198.

LIM, C.Y., SANTOSO, B., BOULAY, T., DONG, E., OHLER, U. and KADONAGA, J.T. (2004). "The MTE, a new core promoter element for transcription by RNA polymerase II." Genes Dev 18(13): 1606-1617.

LINDBLAD-TOH, K., GARBER, M., ZUK, O., LIN, M., PARKER, B., WASHIETL, S., . . . KELLIS, M. (2011). "A high-resolution map of human evolutionary constraint using 29 mammals." Nature 478(7370): 476-482.

LINDBLAD-TOH, K., GARBER, M., ZUK, O., LIN, M.F., PARKER, B.J., WASHIETL, S., . . . KELLIS, M. (2011). "A high-resolution map of human evolutionary constraint using 29 mammals." Nature 478(7370): 476-482.

LIU, C., TENG, Z.Q., SANTISTEVAN, N.J., SZULWACH, K.E., GUO, W., JIN, P. and ZHAO, X. (2010). "Epigenetic regulation of miR-184 by MBD1 governs neural stem cell proliferation and differentiation." Cell Stem Cell 6(5): 433-444.

LIU, S., LIN, L., JIANG, P., WANG, D. and XING, Y. (2011). "A comparison of RNA-Seq and high-density exon array for detecting differential gene expression between closely related species." Nucleic Acids Res 39(2): 578-588.

LIU, X., SOMEL, M., TANG, L., YAN, Z., JIANG, X., GUO, S., . . . KHAITOVICH, P. (2012). "Extension of cortical synaptic development distinguishes humans from chimpanzees and macaques." Genome Research 22(4): 611-622.

LOISEL, D.A., ROCKMAN, M.V., WRAY, G.A., ALTMANN, J. and ALBERTS, S.C. (2006). "Ancient polymorphism and functional variation in the primate MHC-DQA1 5' cis-regulatory region." Proc Natl Acad Sci U S A 103(44): 16331-16336.

LOMVARDAS, S., BARNEA, G., PISAPIA, D.J., MENDELSOHN, M., KIRKLAND, J. and AXEL, R. (2006). "Interchromosomal interactions and olfactory receptor choice." Cell 126(2): 403-413.

LOOTS, G. and OVCHARENKO, I. (2007). "ECRbase: database of evolutionary conserved regions, promoters, and transcription factor binding sites in vertebrate genomes." Bioinformatics 23(1): 122-124.

LOOTS, G.G., OVCHARENKO, I., PACHTER, L., DUBCHAK, I. and RUBIN, E.M. (2002). "rVista for comparative sequence-based discovery of functional transcription factor binding sites." Genome Res 12(5): 832-839.

LOSOS, J.B., MASON, K.A., SINGER, S.R., RAVEN, P.H. and JOHNSON, G.B. (2008). Biology. Boston, McGraw-Hill Higher Education.

LOWE, S.W. and LIN, A.W. (2000). "Apoptosis in cancer." Carcinogenesis 21(3): 485-495.

LURIA, A.R. (1970). "The functional organization of the brain." Sci Am 222(3): 66-72 passim.

LYNCH, J.M., MAILLET, M., VANHOUTTE, D., SCHLOEMER, A., SARGENT, M.A., BLAIR, N.S., . . . MOLKENTIN, J.D. (2012). "A thrombospondin-dependent pathway for a protective ER stress response." Cell 149(6): 1257-1268.

MARIONI, J.C., MASON, C.E., MANE, S.M., STEPHENS, M. and GILAD, Y. (2008). "RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays." Genome Res 18(9): 1509-1517.

MARPLE, B.F., STANKIEWICZ, J.A., BAROODY, F.M., CHOW, J.M., CONLEY, D.B., COREY, J.P., . . . AMERICAN ACADEMY OF OTOLARYNGIC ALLERGY WORKING GROUP ON CHRONIC, R. (2009). "Diagnosis and management of chronic rhinosinusitis in adults." Postgrad Med 121(6): 121-139.

MARQUES, A.C., VINCKENBOSCH, N., BRAWAND, D. and KAESSMANN, H. (2008). "Functional diversification of duplicate genes through subcellular adaptation of encoded proteins." Genome Biol 9(3): R54.

MARQUES, S.M. and ESTEVES DA SILVA, J.C. (2009). "Firefly bioluminescence: a mechanistic approach of luciferase catalyzed reactions." IUBMB Life 61(1): 6-17.

MARTOGLIO, B. (2003). "Intramembrane proteolysis and post-targeting functions of signal peptides." Biochem Soc Trans 31(Pt 6): 1243-1247.

MARVANOVA, M., MENAGER, J., BEZARD, E., BONTROP, R.E., PRADIER, L. and WONG, G. (2003). "Microarray analysis of nonhuman primates: validation of experimental models in neurological disorders." Faseb J 17(8): 929-931.

MAUNAKEA, A.K., CHEPELEV, I. and ZHAO, K. (2010). "Epigenome mapping in normal and disease States." Circ Res 107(3): 327-339.

MAUNAKEA, A.K., NAGARAJAN, R.P., BILENKY, M., BALLINGER, T.J., D'SOUZA, C., FOUSE, S.D., . . . COSTELLO, J.F. (2010). "Conserved role of intragenic DNA methylation in regulating alternative promoters." Nature 466(7303): 253-257.

MCCLURE, H.M. (1973). "Tumors in nonhuman primates: observations during a six-year period in the Yerkes primate center colony." Am J Phys Anthropol 38(2): 425-429.

MCLEAN, C., RENO, P., POLLEN, A., BASSAN, A., CAPELLINI, T., GUENTHER, C., . . . KINGSLEY, D. (2011). "Human-specific loss of regulatory DNA and the evolution of human-specific traits." Nature 471(7337): 216-219.

MCMANUS, C.J., COOLON, J.D., DUFF, M.O., EIPPER-MAINS, J., GRAVELEY, B.R. and WITTKOPP, P.J. (2010). "Regulatory divergence in Drosophila revealed by mRNA-seq." Genome Res 20(6): 816-825.

MEFFORD, H.C., SHARP, A.J., BAKER, C., ITSARA, A., JIANG, Z., BUYSSE, K., . . . EICHLER, E.E. (2008). "Recurrent rearrangements of chromosome 1q21.1 and variable pediatric phenotypes." N Engl J Med 359(16): 1685-1699.

MENZEL, E.W., JR., DAVENPORT, K., JR. and ROGERS, C.M. (1963). "Effects of environmental restriction upon the chimpanzee's responsiveness in novel situations." J Comp Physiol Psychol 56: 329-334.

MURPHY-ULLRICH, J.E. and IOZZO, R.V. (2012). "Thrombospondins in physiology and disease: new tricks for old dogs." Matrix Biol 31(3): 152-154.

MUSTONEN, E., ARO, J., PUHAKKA, J., ILVES, M., SOINI, Y., LESKINEN, H., . . . RYSA, J. (2008). "Thrombospondin-4 expression is rapidly upregulated by cardiac overload." Biochem Biophys Res Commun 373(2): 186-191.

MUSTONEN, E., RUSKOAHO, H. and RYSA, J. (2012). "Thrombospondin-4, tumour necrosis factor-like weak inducer of apoptosis (TWEAK) and its receptor Fn14: novel extracellular matrix modulating factors in cardiac remodelling." Ann Med 44(8): 793-804.

MUSTONEN, E., RUSKOAHO, H. and RYSA, J. (2013). "Thrombospondins, potential drug targets for cardiovascular diseases." Basic Clin Pharmacol Toxicol 112(1): 4-12.

MYERS, A.J., GIBBS, J.R., WEBSTER, J.A., ROHRER, K., ZHAO, A., MARLOWE, L., . . . HARDY, J. (2007). "A survey of genetic human cortical gene expression." Nat Genet 39(12): 1494-1499.

MYERS, R.M., TILLY, K. and MANIATIS, T. (1986). "Fine structure genetic analysis of a beta-globin promoter." Science 232(4750): 613-618.

NATARAJAN, A., YARDIMCI, G.G., SHEFFIELD, N.C., CRAWFORD, G.E. and OHLER, U. (2012). "Predicting cell-type-specific gene expression from regions of open chromatin." Genome Res 22(9): 1711-1722.

NOTTROTT, S., SIMARD, M.J. and RICHTER, J.D. (2006). "Human let-7a miRNA blocks protein production on actively translating polyribosomes." Nat Struct Mol Biol 13(12): 1108-1114.

O'BLENESS, M., SEARLES, V.B., VARKI, A., GAGNEUX, P. and SIKELA, J.M. (2012). "Evolution of genetic and genomic features unique to the human lineage." Nat Rev Genet 13(12): 853-866.

O'BLENESS, M.S., DICKENS, C.M., DUMAS, L.J., KEHRER-SAWATZKI, H., WYCKOFF, G.J. and SIKELA, J.M. (2012). "Evolutionary history and genome organization of DUF1220 protein domains." G3 (Bethesda) 2(9): 977-986.

OBERHEIM, N.A., TAKANO, T., HAN, X., HE, W., LIN, J.H., WANG, F., . . . NEDERGAARD, M. (2009). "Uniquely hominid features of adult human astrocytes." J Neurosci 29(10): 3276-3287.

OSHLACK, A., ROBINSON, M.D. and YOUNG, M.D. (2010). "From RNA-seq reads to differential expression results." Genome Biol 11(12): 220.

OVCHARENKO, I., LOOTS, G.G., GIARDINE, B.M., HOU, M., MA, J., HARDISON, R.C., . . . MILLER, W. (2005). "Mulan: multiple-sequence local alignment and visualization for studying function and evolution." Genome Res 15(1): 184-194.

OVCHARENKO, I., NOBREGA, M.A., LOOTS, G.G. and STUBBS, L. (2004). "ECR Browser: a tool for visualizing and accessing data from comparisons of multiple vertebrate genomes." Nucleic Acids Res 32(Web Server issue): W280-286.

PAI, A., BELL, J., MARIONI, J., PRITCHARD, J. and GILAD, Y. (2011). "A genome-wide study of DNA methylation patterns and gene expression levels in multiple human and chimpanzee tissues." PLoS genetics 7(2).

PALAGI, E., PAOLI, T. and TARLI, S.B. (2004). "Reconciliation and consolation in captive bonobos (Pan paniscus)." Am J Primatol 62(1): 15-30.

PAN, Q., SHAI, O., LEE, L.J., FREY, B.J. and BLENCOWE, B.J. (2008). "Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing." Nat Genet 40(12): 1413-1415.

PAN, Y., TSAI, C.J., MA, B. and NUSSINOV, R. (2010). "Mechanisms of transcription factor selectivity." Trends Genet 26(2): 75-83.

PARK, P.J. (2009). "ChIP-seq: advantages and challenges of a maturing technology." Nat Rev Genet 10(10): 669-680.

PENNACCHIO, L.A., BICKMORE, W., DEAN, A., NOBREGA, M.A. and BEJERANO, G. (2013). "Enhancers: five essential questions." Nat Rev Genet 14(4): 288-295.

PERRY, G.H., DOMINY, N.J., CLAW, K.G., LEE, A.S., FIEGLER, H., REDON, R., . . . STONE, A.C. (2007). "Diet and the evolution of human amylase gene copy number variation." Nat Genet 39(10): 1256-1260.

PERRY, G.H., YANG, F., MARQUES-BONET, T., MURPHY, C., FITZGERALD, T., LEE, A.S., . . . REDON, R. (2008). "Copy number variation and evolution in humans and chimpanzees." Genome Res 18(11): 1698-1710.

PETERSEN, C.P., BORDELEAU, M.E., PELLETIER, J. and SHARP, P.A. (2006). "Short RNAs repress translation after initiation in mammalian cells." Mol Cell 21(4): 533-542.

PETERSEN, T.N., BRUNAK, S., VON HEIJNE, G. and NIELSEN, H. (2011). "SignalP 4.0: discriminating signal peptides from transmembrane regions." Nat Methods 8(10): 785-786.

PETRYKOWSKA, H.M., VOCKLEY, C.M. and ELNITSKI, L. (2008). "Detection and characterization of silencers and enhancer-blockers in the greater CFTR locus." Genome Res 18(8): 1238-1246.

PICKRELL, J.K., MARIONI, J.C., PAI, A.A., DEGNER, J.F., ENGELHARDT, B.E., NKADORI, E., . . . PRITCHARD, J.K. (2010). "Understanding mechanisms underlying human gene expression variation with RNA sequencing." Nature 464(7289): 768-772.

PINTO, D., PAGNAMENTA, A.T., KLEI, L., ANNEY, R., MERICO, D., REGAN, R., . . . BETANCUR, C. (2010). "Functional impact of global rare copy number variation in autism spectrum disorders." Nature 466(7304): 368-372.

PLUSKOTA, E., STENINA, O.I., KRUKOVETS, I., SZPAK, D., TOPOL, E.J. and PLOW, E.F. (2005). "Mechanism and effect of thrombospondin-4 polymorphisms on neutrophil function." Blood 106(12): 3970-3978.

POLAVARAPU, N., ARORA, G., MITTAL, V.K. and MCDONALD, J.F. (2011). "Characterization and potential functional significance of human-chimpanzee large INDEL variation." Mob DNA 2: 13.

PONICSAN, S.L., KUGEL, J.F. and GOODRICH, J.A. (2010). "Genomic gems: SINE RNAs regulate mRNA production." Curr Opin Genet Dev 20(2): 149-155.

POPESCO, M.C., MACLAREN, E.J., HOPKINS, J., DUMAS, L., COX, M., MELTESEN, L., . . . SIKELA, J.M. (2006). "Human lineage-specific amplification, selection, and neuronal expression of DUF1220 domains." Science 313(5791): 1304-1307.

PRABHAKAR, S., VISEL, A., AKIYAMA, J.A., SHOUKRY, M., LEWIS, K.D., HOLT, A., . . . NOONAN, J.P. (2008). "Human-specific gain of function in a developmental enhancer." Science 321(5894): 1346-1350.

PRADO-MARTINEZ, J., HERNANDO-HERRAEZ, I., LORENTE-GALDOS, B., DABAD, M., RAMIREZ, O., BAEZA-DELGADO, C., . . . MARQUES-BONET, T. (2013). "The genome sequencing of an albino Western lowland gorilla reveals inbreeding in the wild." BMC Genomics 14: 363.

PREUSS, T.M. (2004). What's it like to be a human? The Cognitive Neurosciences III. M. S. Gazzaniga. Cambridge, MA, MIT Press: (in press).

PREUSS, T.M. (2011). "The human brain: rewired and running hot." Ann N Y Acad Sci 1225 Suppl 1: E182-191.

PREUSS, T.M., CÁCERES, M., OLDHAM, M.C. and GESCHWIND, D.H. (2004). "Human brain evolution: insights from microarrays." Nat Rev Genet 5(11): 850-860.

PREUSS, T.M. and COLEMAN, G.Q. (2002). "Human-specific organization of primary visual cortex: alternating compartments of dense Cat-301 and calbindin immunoreactivity in layer 4A." Cereb Cortex 12(7): 671-691.

PRUETZ, J.D. and BERTOLANI, P. (2007). "Savanna chimpanzees, Pan troglodytes verus, hunt with tools." Curr Biol 17(5): 412-417.

PRUFER, K., MUNCH, K., HELLMANN, I., AKAGI, K., MILLER, J.R., WALENZ, B., . . . PAABO, S. (2012). "The bonobo genome compared with the chimpanzee and human genomes." Nature 486(7404): 527-531.

PUENTE, X.S., VELASCO, G., GUTIERREZ-FERNANDEZ, A., BERTRANPETIT, J., KING, M.C. and LOPEZ-OTIN, C. (2006). "Comparative analysis of cancer genes in the human and chimpanzee genomes." BMC Genomics 7: 15.

QIU, J. (2006). "Epigenetics: unfinished symphony." Nature 441(7090): 143-145.

RAGHANTI, M.A., STIMPSON, C.D., MARCINKIEWICZ, J.L., ERWIN, J.M., HOF, P.R. and SHERWOOD, C.C. (2008). "Cortical dopaminergic innervation among humans, chimpanzees, and macaque monkeys: a comparative study." Neuroscience 155(1): 203-220.

REDDY, P.M., STAMATOYANNOPOULOS, G., PAPAYANNOPOULOU, T. and SHEN, C.K. (1994). "Genomic footprinting and sequencing of human beta-globin locus. Tissue specificity and cell line artifact." J Biol Chem 269(11): 8287-8295.

REID, A.T. and EVANS, A.C. (2013). "Structural networks in Alzheimer's disease." Eur Neuropsychopharmacol.

REIK, W. (2007). "Stability and flexibility of epigenetic gene regulation in mammalian development." Nature 447(7143): 425-432.

RILLING, J.K., BARKS, S.K., PARR, L.A., PREUSS, T.M., FABER, T.L., PAGNONI, G., . . . VOTAW, J.R. (2007). "A comparison of resting-state brain activity in humans and chimpanzees." Proc Natl Acad Sci U S A 104(43): 17146-17151.

RISHER, W.C. and EROGLU, C. (2012). "Thrombospondins as key regulators of synaptogenesis in the central nervous system." Matrix Biol 31(3): 170-177.

ROMERO, I.G., RUVINSKY, I. and GILAD, Y. (2012). "Comparative studies of gene expression and the evolution of gene regulation." Nat Rev Genet 13(7): 505-516.

ROTH, G. and DICKE, U. (2012). "Evolution of the brain and intelligence in primates." Prog Brain Res 195: 413-430.

SAMUELSON, L.C., PHILLIPS, R.S. and SWANBERG, L.J. (1996). "Amylase gene structures in primates: retroposon insertions and promoter evolution." Mol Biol Evol 13(6): 767-779.

SCALLY, A., DUTHEIL, J.Y., HILLIER, L.W., JORDAN, G.E., GOODHEAD, I., HERRERO, J., . . . DURBIN, R. (2012). "Insights into hominid evolution from the gorilla genome sequence." Nature 483(7388): 169-175.

SCHENKER, N.M., DESGOUTTES, A.M. and SEMENDEFERI, K. (2005). "Neural connectivity and cortical substrates of cognition in hominoids." J Hum Evol 49(5): 547-569.

SCHENKER, N.M., HOPKINS, W.D., SPOCTER, M.A., GARRISON, A.R., STIMPSON, C.D., ERWIN, J.M., . . . SHERWOOD, C.C. (2010). "Broca's area homologue in chimpanzees (Pan troglodytes): probabilistic mapping, asymmetry, and comparison to humans." Cereb Cortex 20(3): 730-742.

SCHMIDT, D., WILSON, M.D., BALLESTER, B., SCHWALIE, P.C., BROWN, G.D., MARSHALL, A., . . . ODOM, D.T. (2010). "Five-vertebrate ChIP-seq reveals the evolutionary dynamics of transcription factor binding." Science 328(5981): 1036-1040.

SCHRIER, A.M., HARLOW, H.F. and STOLLNITZ, F. (1965). Behavior of nonhuman primates; modern research trends. New York,, Academic Press.

SCHWARTZ, S., ELNITSKI, L., LI, M., WEIRAUCH, M., RIEMER, C., SMIT, A., . . . MILLER, W. (2003). "MultiPipMaker and supporting tools: Alignments and analysis of multiple genomic DNA sequences." Nucleic Acids Res 31(13): 3518-3524.

SCOTT, G.B.D. (1992). Comparative primate pathology. Oxford, Oxford University Press.

SEGAT, L. and CROVELLA, S. (2011). "MBL1 gene in nonhuman primates." Hum Immunol 72(11): 1084-1090.

SEIBOLD, H.R. and WOLF, R.H. (1973). "Neoplasms and proliferative lesions in 1065 nonhuman primate necropsies." Lab Anim Sci 23(4): 533-539.

SEMENDEFERI, K., LU, A., SCHENKER, N. and DAMASIO, H. (2002). "Humans and great apes share a large frontal cortex." Nat Neurosci 5(3): 272-276.

SEMENDEFERI, K., TEFFER, K., BUXHOEVEDEN, D.P., PARK, M.S., BLUDAU, S., AMUNTS, K., . . . BUCKWALTER, J. (2011). "Spatial organization of neurons in the frontal pole sets humans apart from great apes." Cereb Cortex 21(7): 1485-1497.

SEXTON, T., BANTIGNIES, F. and CAVALLI, G. (2009). "Genomic interactions: chromatin loops and gene meeting points in transcriptional regulation." Semin Cell Dev Biol 20(7): 849-855.

SHAPIRO, M.D., MARKS, M.E., PEICHEL, C.L., BLACKMAN, B.K., NERENG, K.S., JONSSON, B., . . . KINGSLEY, D.M. (2004). "Genetic and developmental basis of evolutionary pelvic reduction in threespine sticklebacks." Nature 428(6984): 717-723.

SHERWOOD, C.C., STIMPSON, C.D., RAGHANTI, M.A., WILDMAN, D.E., UDDIN, M., GROSSMAN, L.I., . . . HOF, P.R. (2006). "Evolution of increased glia-neuron ratios in the human frontal cortex." Proc Natl Acad Sci U S A 103(37): 13606-13611.

SHIBATA, Y., SHEFFIELD, N.C., FEDRIGO, O., BABBITT, C.C., WORTHAM, M., TEWARI, A.K., . . . CRAWFORD, G.E. (2012). "Extensive evolutionary changes in regulatory element activity during human origins are associated with altered gene expression and positive selection." PLoS Genet 8(6): e1002789.

SHIMOKAWA, K., OKAMURA-OHO, Y., KURITA, T., FRITH, M.C., KAWAI, J., CARNINCI, P. and HAYASHIZAKI, Y. (2007). "Large-scale clustering of CAGE tag expression data." BMC Bioinformatics 8: 161.

SHIRAKI, T., KONDO, S., KATAYAMA, S., WAKI, K., KASUKAWA, T., KAWAJI, H., . . . HAYASHIZAKI, Y. (2003). "Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage." Proceedings of the National Academy of Sciences of the United States of America 100(26): 15776-15781.

SHIRANGI, T.R., DUFOUR, H.D., WILLIAMS, T.M. and CARROLL, S.B. (2009). "Rapid evolution of sex pheromone-producing enzyme expression in Drosophila." PLoS Biol 7(8): e1000168.

SKOTHEIM, R.I. and NEES, M. (2007). "Alternative splicing in cancer: noise, functional, or systematic?" Int J Biochem Cell Biol 39(7-8): 1432-1449.

SMALE, S.T. and BALTIMORE, D. (1989). "The "initiator" as a transcription control element." Cell 57(1): 103-113.

SOMEL, M., FRANZ, H., YAN, Z., LORENC, A., GUO, S., GIGER, T., . . . KHAITOVICH, P. (2009). "Transcriptional neoteny in the human brain." Proc Natl Acad Sci U S A 106(14): 5743-5748.

SOMEL, M., GUO, S., FU, N., YAN, Z., HU, H.Y., XU, Y., . . . KHAITOVICH, P. (2010). "MicroRNA, mRNA, and protein expression link development and aging in human and macaque brain." Genome Res 20(9): 1207-1218.

SOMEL, M., LIU, X. and KHAITOVICH, P. (2013). "Human brain evolution: transcripts, metabolites and their regulators." Nat Rev Neurosci 14(2): 112-127.

SOMEL, M., LIU, X., TANG, L., YAN, Z., HU, H., GUO, S., . . . KHAITOVICH, P. (2011). "MicroRNA-driven developmental remodeling in the brain distinguishes humans from other primates." PLoS biology 9(12).

SONG, L. and CRAWFORD, G.E. (2010). "DNase-seq: a high-resolution technique for mapping active gene regulatory elements across the genome from mammalian cells." Cold Spring Harb Protoc 2010(2): pdb prot5384.

STANKIEWICZ, P. and LUPSKI, J.R. (2010). "Structural variation in the human genome and its role in disease." Annu Rev Med 61: 437-455.

STENINA, O.I., DESAI, S.Y., KRUKOVETS, I., KIGHT, K., JANIGRO, D., TOPOL, E.J. and PLOW, E.F. (2003). "Thrombospondin-4 and its variants: expression and differential effects on endothelial cells." Circulation 108(12): 1514-1519.

STENINA, O.I., TOPOL, E.J. and PLOW, E.F. (2007). "Thrombospondins, their polymorphisms, and cardiovascular disease." Arterioscler Thromb Vasc Biol 27(9): 1886-1894.

STENINA, O.I., USTINOV, V., KRUKOVETS, I., MARINIC, T., TOPOL, E.J. and PLOW, E.F. (2005). "Polymorphisms A387P in thrombospondin-4 and N700S in thrombospondin-1 perturb calcium binding sites." FASEB J 19(13): 1893-1895.

STERN, D.L. (2000). "Evolutionary developmental biology and the problem of variation." Evolution Int J Org Evolution 54(4): 1079-1091.

STERN, D.L. and ORGOGOZO, V. (2008). "The loci of evolution: how predictable is genetic evolution?" Evolution 62(9): 2155-2177.

STRANGER, B.E., FORREST, M.S., CLARK, A.G., MINICHIELLO, M.J., DEUTSCH, S., LYLE, R., . . . DERMITZAKIS, E.T. (2005). "Genome-wide associations of gene expression variation in humans." PLoS Genet 1(6): e78.

STRANGER, B.E., FORREST, M.S., DUNNING, M., INGLE, C.E., BEAZLEY, C., THORNE, N., . . . DERMITZAKIS, E.T. (2007). "Relative impact of nucleotide and copy number variation on gene expression phenotypes." Science 315(5813): 848-853.

STRANGER, B.E., NICA, A.C., FORREST, M.S., DIMAS, A., BIRD, C.P., BEAZLEY, C., . . . DERMITZAKIS, E.T. (2007). "Population genomics of human gene expression." Nat Genet 39(10): 1217-1224.

SUZUKI, Y., TSUNODA, T., SESE, J., TAIRA, H., MIZUSHIMA-SUGANO, J., HATA, H., . . . SUGANO, S. (2001). "Identification and characterization of the potential promoter regions of 1031 kinds of human genes." Genome Res 11(5): 677-684.

SZAMALEK, J.M., GOIDTS, V., CHUZHANOVA, N., HAMEISTER, H., COOPER, D.N. and KEHRER-SAWATZKI, H. (2005). "Molecular characterisation of the pericentric inversion that distinguishes human chromosome 5 from the homologous chimpanzee chromosome." Hum Genet 117(2-3): 168-176.

TAKAHASHI, H., KATO, S., MURATA, M. and CARNINCI, P. (2012). "CAGE (cap analysis of gene expression): a protocol for the detection of promoter and transcriptional networks." Methods Mol Biol 786: 181-200.

TAKAOKA, A.S., YAMADA, T., GOTOH, M., KANAI, Y., IMAI, K. and HIROHASHI, S. (1998). "Cloning and characterization of the human beta4-integrin gene promoter and enhancers." J Biol Chem 273(50): 33848-33855.

TASLIM, C., LIN, S., HUANG, K. and HUANG, T.H. (2012). "Integrative genome-wide chromatin signature analysis using finite mixture models." BMC Genomics 13 Suppl 6: S3.

TEFFER, K. and SEMENDEFERI, K. (2012). "Human prefrontal cortex: evolution, development, and pathology." Prog Brain Res 195: 191-218.

TIROSH, I., REIKHAV, S., LEVY, A.A. and BARKAI, N. (2009). "A yeast hybrid provides insight into the evolution of gene expression regulation." Science 324(5927): 659-662.

TISHKOFF, S.A., REED, F.A., RANCIARO, A., VOIGHT, B.F., BABBITT, C.C., SILVERMAN, J.S., . . . DELOUKAS, P. (2007). "Convergent adaptation of human lactase persistence in Africa and Europe." Nat Genet 39(1): 31-40.

TOPOL, E.J., MCCARTHY, J., GABRIEL, S., MOLITERNO, D.J., ROGERS, W.J., NEWBY, L.K., . . . DALEY, G.Q. (2001). "Single nucleotide polymorphisms in multiple novel thrombospondin genes may be associated with familial premature myocardial infarction." Circulation 104(22): 2641-2644.

TOURNAMILLE, C., COLIN, Y., CARTRON, J.P. and LE VAN KIM, C. (1995). "Disruption of a GATA motif in the Duffy gene promoter abolishes erythroid gene expression in Duffy-negative individuals." Nat Genet 10(2): 224-228.

TREVATHAN, W.R. (1996). "The evolution of bipedalism and assisted birth." Med Anthropol Q 10(2): 287-290.

TSACOPOULOS, M. and MAGISTRETTI, P.J. (1996). "Metabolic coupling between glia and neurons." J Neurosci 16(3): 877-885.

TSIEN, R.Y. (1998). "The green fluorescent protein." Annu Rev Biochem 67: 509-544.

UDDIN, M., OPAZO, J.C., WILDMAN, D.E., SHERWOOD, C.C., HOF, P.R., GOODMAN, M. and GROSSMAN, L.I. (2008). "Molecular evolution of the cytochrome c oxidase subunit 5A gene in primates." BMC Evol Biol 8: 8.

UDDIN, M., WILDMAN, D.E., LIU, G., XU, W., JOHNSON, R.M., HOF, P.R., . . . GOODMAN, M. (2004). "Sister grouping of chimpanzees and humans as revealed by genome-wide phylogenetic analysis of brain gene expression profiles." Proc Natl Acad Sci U S A 101(9): 2957-2962.

UNTERGASSER, A., CUTCUTACHE, I., KORESSAAR, T., YE, J., FAIRCLOTH, B.C., REMM, M. and ROZEN, S.G. (2012). "Primer3--new capabilities and interfaces." Nucleic Acids Res 40(15): e115.

VANE-WRIGHT, D. (2004). "Entomology: butterflies at that awkward age." Nature 428(6982): 477-480.

VARKI, A. and ALTHEIDE, T.K. (2005). "Comparing the human and chimpanzee genomes: searching for needles in a haystack." Genome Res 15(12): 1746-1758.

VARKI, N.M., STROBERT, E., DICK, E.J., JR., BENIRSCHKE, K. and VARKI, A. (2011). "Biomedical differences between human and nonhuman hominids: potential roles for uniquely human aspects of sialic acid biology." Annu Rev Pathol 6: 365-393.

VEDEL, V. and SCOTTI, I. (2011). "Promoting the promoter." Plant science : an international journal of experimental plant biology 180(2): 182-189.

VENTER, J.C., ADAMS, M.D., MYERS, E.W., LI, P.W., MURAL, R.J., SUTTON, G.G., . . . ZHU, X. (2001). "The sequence of the human genome." Science 291(5507): 1304-1351.

VISEL, A., BLOW, M., LI, Z., ZHANG, T., AKIYAMA, J., HOLT, A., . . . PENNACCHIO, L. (2009). "ChIP-seq accurately predicts tissue-specific activity of enhancers." Nature 457(7231): 854-858.

VISEL, A., BLOW, M.J., LI, Z., ZHANG, T., AKIYAMA, J.A., HOLT, A., . . . PENNACCHIO, L.A. (2009). "ChIP-seq accurately predicts tissue-specific activity of enhancers." Nature 457(7231): 854-858.

WALLACE, J.A. and FELSENFELD, G. (2007). "We gather together: insulators and genome organization." Curr Opin Genet Dev 17(5): 400-407.

WALLICH, R., BRENNER, C., BRAND, Y., ROUX, M., REISTER, M. and MEUER, S. (1998). "Gene structure, promoter characterization, and basis for alternative mRNA splicing of the human CD58 gene." J Immunol 160(6): 2862-2871.

WANG, E.T., SANDBERG, R., LUO, S., KHREBTUKOVA, I., ZHANG, L., MAYR, C., . . . BURGE, C.B. (2008). "Alternative isoform regulation in human tissue transcriptomes." Nature 456(7221): 470-476.

WANG, X., MITRA, N., CRUZ, P., DENG, L., PROGRAM, N.C.S., VARKI, N., . . . VARKI, A. (2012). "Evolution of siglec-11 and siglec-16 genes in hominins." Mol Biol Evol 29(8): 2073-2086.

WANG, Y. and NEUMANN, H. (2010). "Alleviation of neurotoxicity by microglial human Siglec-11." J Neurosci 30(9): 3482-3488.

WESSEL, J., TOPOL, E.J., JI, M., MEYER, J. and MCCARTHY, J.J. (2004). "Replication of the association between the thrombospondin-4 A387P polymorphism and myocardial infarction." Am Heart J 147(5): 905-909.

WEST, A.G. and FRASER, P. (2005). "Remote control of gene transcription." Hum Mol Genet 14 Spec No 1: R101-111.

WEST, A.G., GASZNER, M. and FELSENFELD, G. (2002). "Insulators: many functions, many mechanisms." Genes Dev 16(3): 271-288.

WHITEN, A. and ERDAL, D. (2012). "The human socio-cognitive niche and its evolutionary origins." Philos Trans R Soc Lond B Biol Sci 367(1599): 2119-2129.

WHITEN, A., GOODALL, J., MCGREW, W.C., NISHIDA, T., REYNOLDS, V., SUGIYAMA, Y., . . . BOESCH, C. (1999). "Cultures in chimpanzees." Nature 399(6737): 682-685.

WHITEN, A., MCGUIGAN, N., MARSHALL-PESCINI, S. and HOPPER, L.M. (2009). "Emulation, imitation, over-imitation and the scope of culture for child and chimpanzee." Philos Trans R Soc Lond B Biol Sci 364(1528): 2417-2428.

WINGENDER, E., DIETZE, P., KARAS, H. and KNUPPEL, R. (1996). "TRANSFAC: a database on transcription factors and their DNA binding sites." Nucleic Acids Res 24(1): 238-241.

WITTKOPP, P. and KALAY, G. (2012). "Cis-regulatory elements: molecular mechanisms and evolutionary processes underlying divergence." Nature reviews. Genetics 13(1): 59-69.

WITTKOPP, P.J., HAERUM, B.K. and CLARK, A.G. (2004). "Evolutionary changes in cis and trans gene regulation." Nature 430(6995): 85-88.

WITTKOPP, P.J., HAERUM, B.K. and CLARK, A.G. (2008). "Regulatory changes underlying expression differences within and between Drosophila species." Nat Genet 40(3): 346-350.

WRAY, G.A. (2007). "The evolutionary significance of cis-regulatory mutations." Nat Rev Genet 8(3): 206-216.

WRIGHT, S. (1950). "Genetical structure of populations." Nature 166(4215): 247-249.

WU, L., FAN, J. and BELASCO, J.G. (2006). "MicroRNAs direct rapid deadenylation of mRNA." Proc Natl Acad Sci U S A 103(11): 4034-4039.

WU, X., RAUCH, T.A., ZHONG, X., BENNETT, W.P., LATIF, F., KREX, D. and PFEIFER, G.P. (2010). "CpG island hypermethylation in human astrocytomas." Cancer Res 70(7): 2718-2727.

XU, A.G., HE, L., LI, Z., XU, Y., LI, M., FU, X., . . . KHAITOVICH, P. (2010). "Intergenic and repeat transcription in human, chimpanzee and macaque brains measured by RNA-Seq." PLoS Comput Biol 6: e1000843.

YUNIS, J.J. and PRAKASH, O. (1982). "The origin of man: a chromosomal pictorial legacy." Science 215(4539): 1525-1530.

ZENG, J., KONOPKA, G., HUNT, B.G., PREUSS, T.M., GESCHWIND, D. and YI, S.V. (2012). "Divergent whole-genome methylation maps of human and chimpanzee brains reveal epigenetic basis of human regulatory evolution." Am J Hum Genet 91(3): 455-465.

ZHANG, D., GUO, L., ZHU, D., LI, K., LI, L., CHEN, H., . . . LIU, T. (2012). "Diffusion tensor imaging reveals evolution of primate brain architectures." Brain Struct Funct.

ZHANG, J., WEBB, D.M. and PODLAHA, O. (2002). "Accelerated protein evolution and origins of human-specific features: Foxp2 as an example." Genetics 162(4): 1825-1835.

ZHANG, Y., LANDBACK, P., VIBRANOVSKI, M. and LONG, M. (2011). "Accelerated recruitment of new brain development genes into the human genome." PLoS biology 9(10).

ZHANG, Z. and GERSTEIN, M. (2003). "Of mice and men: phylogenetic footprinting aids the discovery of regulatory elements." J Biol 2(2): 11.

ZILLER, M.J., MULLER, F., LIAO, J., ZHANG, Y., GU, H., BOCK, C., . . . MEISSNER, A. (2011). "Genomic distribution and inter-sample variation of non-CpG methylation across human cell types." PLoS Genet 7(12): e1002389.

# ABBREVIATIONS

| | |
|---|---|
| aa | amino acid |
| AD | Alzheimer's disease |
| Aβ | β-amyloid |
| BAC | bacterial artificial chromosome |
| bp | base pairs |
| BRE | B recognition element |
| BSA | bovine serum albumin |
| cDNA | complementary DNA |
| ChIP | Chromatin immunoprecipitation |
| CpG | cytosine-phosphate-Guanina |
| CRE | *cis*-regulatory elements |
| DMEM | Dulbecco's modified eagle medium |
| DNA | deoxyribonucleic acid |
| DNase I | deoxyribonuclease I |
| DPE | downstream promoter element |
| DTI | diffusion-tensor imaging |
| ECR | evolutionary conserved regions |
| EDTA | ethylenediaminetetraacetic acid |
| eQTL | expression quantitative trait loci |
| ER | endoplasmic reticulum |
| EST | expressed sequence tag |
| FBS | fetal bovine serum |
| *HACNS1* | human-accelerated conserved non-coding sequence 1 |
| HSE | heat shock sequence elements |
| HRE | hormone response elements |
| Inr | Iniciator element |
| IgG | Immunoglobulin G |
| LB | Luria-Bertani |
| LTR | long terminal repeat |
| MEM | minimum essential medium |
| miRNA | microRNA |

| | |
|---|---|
| MRI | magnetic resonance imaging |
| mRNA | messenger RNA |
| MTE | motif ten element |
| NGS | next-generation sequencing |
| ORF | open reading frame |
| PBS | phosphate buffered saline |
| PCR | polymerase chain reaction |
| PET | positron emission tomography |
| RE | response elements |
| RNA | ribonucleic acid |
| RNAP-II | RNA polymerase II |
| RNase | ribonuclease |
| RT-PCR | reverse transcription polymerase chain reaction |
| SRE | serum response elements |
| SNP | single nucleotide polymorphism |
| TE | transposable element |
| TBP | TATA binding protein |
| TFBS | transcription factor binding site |
| TMRCA | time most recent common ancestor |
| TSS | transcription start site |
| UTR | untranslated region |

## UNITS

| | |
|---|---|
| g | gram |
| h | hour |
| M | molar |
| min | minute |
| ml | mililiter |
| mM | milimolar |
| ng | nanogram |
| sec | second |
| µg | microgram |

| | |
|---|---|
| µl | microliter |
| º C | degree Celsius |

## GENES

| | |
|---|---|
| *ACTB* | ß-actin |
| *COMP* | cartilage oliogomeric matrix protein |
| *COX5A* | COX5A cytochrome c oxidase subunit Va |
| *FOXP2* | forkhead box P2 |
| *GADD45G* | Growth arrest and DNA-damage-inducible protein GADD45 gamma |
| *GAPDH* | glyceraldehyde 3-phosphate dehydrogenase |
| *HACNS1* | human-accelerated conserved non-coding sequence 1 |
| *LCT* | Lactose gene |
| *MCM6* | DNA replication licensing factor MCM6 |
| *SIGLEC16P* | Sialic acid-binding Ig-like lectin 16 pseudogene |
| *SIGLEC11* | Sialic acid-binding Ig-like lectin 11 |
| *THBSs* | thrombospondins |
| *THBS1* | thrombospondin-1 |
| *THBS2* | thrombospondin-2 |
| *THBS3* | thrombospondin-3 |
| *THBS4* | thrombospondin-4 |
| *THBS5* | thrombospondin-5 |

# INDEX OF FIGURES

## INTRODUCTION



## MATERIALS AND METHODS



## RESULTS

## DISCUSION

# INDEX OF TABLES AND BOXES

## INTRODUCTION



## MATERIALS AND METHODS

# ACKNOWLEDGEMENTS

*"Every individual matters. Every individual has a role to play.*
*Every individual makes a difference."*

– JANE GOODALL –

All these years of work would not have been the same without the help and support of many people; I would like to convey my sincere thanks:

To my advisor Mario Cáceres, for giving me the opportunity to carry out this work. For all the things you have taught me, for your patience and all those "what do you bet?" that made me think several times how I was performing the experiments.

To the fellows and professors from S.E.K. university, for all the good moments that we have had together, because you were the beginning of this adventure. To Virginia Otones, my *little potential superpower*, for your smile and optimism, for being always there, for being yourself, for counting on me.

To Xavier Estivill, for allowing me to work from your laboratory the first two years. To all the people with who I shared those years at the CRG and made me have so good moments. To Lorena Pantano, for performing the ChIP-seq analysis, for answering all my "iDoubts" with OS X and of course for your friendship. To Cecilia Ballaré, for teaching me how to perform the ChIP-seq experiments.

To Sergio Villatoro, for your support, for your restless search for answers, because Mari, Nora and you have been my family these last years. You are priceless. THANKS.

To Jordi Esquinas, for the affection that you and your family gave me for so many years, for being interested about my lab experiments and try to understand them.

To Funtional and Comparative Genomics group, for accepting this lost sheep speaking about thrombospondins and not about inversions. Special thanks to Alexander Martinez and to David Vicente, for your good mood; to Marta Puig, for your advices and your stories about little Joan and Sara's adventures from time to time; to Cristina Aguado, for the Segovians' mutual understanding, for all the cinema evenings with the twins.

Thanks to all the people from the IBB and the UAB with who I shared these last years, to Fran Cortés, for your help in the culture room and all the time we spend talking; to Olivia Tort, for your help with western blots; to the Genomic, bioinformatics and evolution group, for all the good moments that we have spent together within and outside university.

To the former Ministerio de Ciencia e Innovación of Spain, for promoting the R&D, for the four years of FPI fellowship that have allowed me to carry out this thesis and for the short-stay travel grants for working at the United States.

To Todd Preuss, for allowing me to use the primate samples from your laboratory, and to Carolyn Suwyn and Mary Cree for giving me the information of the different individuals.

To James Thomas, for accepting me in your laboratory and for being so aware that all was good, for your interest in my project and for having always a smile for sharing. To Jamie Davis, for your help in the lab, for your friendship, for inviting me to the Mary Poppins's musical and for discovering after three months that I was using hand cream instead soap to wash my hands. To Karoun Bagamian and Gladys González, for counting on me so many times outside the lab.

I would like also to say thank you to all the people that I found at Villa International Atlanta and all of you that work hard for making it felling like home being so far of our places. Thanks to Zoltán, Sarah, Paola, Christian, Shiromi, Patrick, José, Saad, Mar, Marta... and many others, for all the good memories that I have, because friendship does not understand about frontiers or religion. To Arun Singh, for being such a good flatmate.

To my friends, for accepting me every time I return home, for updating me about how is everything going from time to time, for saying always the things as you think them.

To my family, for your support, for understand that I cannot be with you all the time that I would like to be, for missing me every time I am not there, for sending me pictures to have you closer. To my two guardian angels, for always protecting me.

To Stefan Kammermeier, for your help with the English writing this thesis, for your affection, for searching a way to make me happy everyday. To Luise and Rolf Kammermeier, for accepting me from the first day with a smile.

To my brother, David Rubio, for taking two sewing threads to explain me the DNA structure, for being such a good person, for your happiness and your positivism. I am proud of being your sister.

To my parents, Carlos Rubio and Mª Ángeles Acero, if I get until here is because of you, because only you know how hard was the way on many occasions. Thanks for being always there. This thesis is also yours.

To my cats, Alan and Suerte, for keeping me company while writing, for your contributions every time I left my laptop open ("*KLP'+:_OP0'ioasñle"*).

To all the people who I have not mentioned here, because both for good or for bad moments "Every individual matters. Every individual makes a difference" To all of you, thank you.

# AGRADECIMIENTOS

*"Cada individuo importa. Cada individuo tiene un papel que desempeñar.*
*Cada individuo marca la diferencia."*

– JANE GOODALL –

Porque estos años de trabajo no hubiesen sido lo mismo sin la ayuda y el apoyo de muchas personas, quisiera transmitir mi más sincero agradecimiento:

A mi director de tesis Mario Cáceres, por contar conmigo a la hora de llevar a cabo este trabajo. Por enseñarme tantísimas cosas, por la paciencia que has mostrado conmigo y por todos esos "¿qué te apuestas?" que me hicieron pensar varias veces como estaba llevando a cabo los experimentos.

A los compañeros y profesores de la universidad S.E.K. por todos los buenos momentos que hemos pasado juntos, porque vosotros fuisteis el principio de esta aventura. A Virginia Otones, mi pequeña gran superpotencia, por tu sonrisa y optimismo, por estar siempre ahí, por ser tu misma, por contar conmigo.

A Xavier Estivill, por permitirme trabajar desde tu laboratorio los dos primeros años. A toda la gente con la que compartí esos años por el CRG y me hicieron pasar tan buenos momentos. A Lorena Pantano, por los análisis del ChIP-seq, por responder todas mis "iDudas" con OS X, y sobre todo por tu amistad. A Cecilia Ballaré, por enseñarme como llevar a cabo los experimentos de ChIP-seq.

A Sergio Villatoro, por tu apoyo, por tu ansia de búsqueda incansable, porque Mari, Nora y tú habéis sido mi familia durante los últimos años. Vales muchísimo. GRACIAS.

A Jordi Esquinas, por el cariño con el que tú y tu familia me habéis tratado durante tantos años, por querer saber y entender los experimentos que hacía en el laboratorio.

Al grupo de Genómica comparativa y funcional, por aceptar a esta oveja descarriada hablando de thrombospondinas y no de inversiones. Gracias en especial a Alexander Martinez y a David Vicente, por vuestro buen rollo; a Marta Puig, por tus consejos y por contarnos las aventurillas de Joan y Sara de tanto en tanto; a Cristina Aguado, por tu complicidad segoviana, por las tardes de cine con los gemelos.

Gracias a la gente del IBB y de la UAB con la que he compartido los últimos años, a Fran Cortés, por tu ayuda en cultivos y tantísimas conversaciones; a Olivia Tort, por ayudarme con los western blots; al grupo de Genómica, bioinformática y evolución, por todos los buenos momentos que hemos pasado juntos dentro y fuera de la universidad.

Al que fue el Ministerio de Ciencia e Innovación de España, por impulsar las políticas de I+D+I, por los cuatro años de beca FPI que me han permitido llevar a cabo esta tesis y las becas para realizar estancias breves en Estados Unidos.

A Todd Preuss, por permitirme trabajar con las muestras de primates de tu laboratorio y a Carolyn Suwyn y Mary Cree por facilitarme la información de los distintos individuos.

A James Thomas, por aceptarme en tu laboratorio y estar tan pendiente de que todo esté bien, por interesarte por mi proyecto y siempre tener una sonrisa para los demás. A Jamie Davis, por tu ayuda en el laboratorio, por tu amistad, por invitarme al musical de Mary Poppins y descubrir después de tres meses que me estaba lavando las manos con crema y no con jabón. A Karoun Bagamian y Gladys González, por contar tantas veces conmigo fuera del laboratorio.

Quisiera agradecer también a toda la gente con la que me encontré en Villa International Atlanta y a todos los que trabajáis duro para que nos sintamos como en casa estando tan lejos. Gracias a Zoltán, Sarah, Paola, Christian, Shiromi, Patrick, José, Saad, Mar, Marta... y muchos más, por todos los buenos recuerdos que tengo, porque entendí que la amistad no entiende de fronteras ni de religión. A Arun Singh, por ser tan buen compañero de piso.

A mis amigas, por aceptarme cada vez que vuelvo a casa, por actualizarme de tanto en tanto como van las cosas, por contar conmigo, por decir siempre las cosas claras (y las narices rojas).

A mis familia, por vuestro apoyo, por entender que no puedo estar con vosotros todo lo que me gustaría, por echarme de menos cada vez que falto, por enviarme fotos para hacerme estar presente. A mis dos ángeles de la guarda, por protegerme siempre.

A Stefan Kammermeier, por ayudarme con el inglés a la hora de escribir esta tesis, por tu cariño, por buscar la forma de hacerme feliz todos lo días. A Luise y Rolf Kammermeier, por aceptarme desde el primer día con una sonrisa.

A mi hermano, David Rubio, por coger un par de hilos de costura y explicarme la estructura del ADN, por ser tan buena persona, por tu alegría y tu positivismo. Estoy orgullosa de ser tu hermana.

A mis padres, Carlos Rubio y Mª Ángeles Acero, porque si he llegado hasta aquí es gracias a vosotros, porque solo vosotros sabéis lo duro que fue el camino en muchas ocasiones. Gracias por estar siempre ahí. Esta tesis también es vuestra.

A mis gatos, Alan y Suerte, por hacerme tanta compañía escribiendo, por vuestras aportaciones cada vez que me he dejado el portátil abierto (“*KLP’+:_OP0‘ioasñle”*).

A todas las personas que no he nombrado aquí, porque tanto para los buenos como para los malos momentos “Cada individuo importa. Cada individuo marca la diferencia.” A todos vosotros, gracias.