Rita Leonor Franco Paulo Guimarães

Licenciatura em Biologia Molecular e Celular

ZNF675 evolutionary history and its novel role in neuronal regulatory networks

Dissertação para obtenção do Grau de Mestre em Genética Molecular e Biomedicina

Orientador: Frank M.J. Jacobs, Assistant Professor, Swammerdam Institute for Life Sciences, University of Amsterdam
Co-orientador: Gerrald Lodewijk, PhD Student, Swammerdam Institute for Life Sciences, University of Amsterdam

Júri:

Presidente: Prof. Doutora Margarida Casal Ribeiro Castro Caldas Braga, Faculdade de Ciências e Tecnologia da Universidade Nova de Lisboa
Arguente: Doutora Maria João de Jesus Nunes, Faculdade de Farmácia da Universidade de Lisboa.
Vogal: Prof. Doutor Pedro Miguel Ribeiro Viana Baptista, Faculdade de Ciências e Tecnologia da Universidade Nova de Lisboa

FACULDADE DE CIÊNCIAS E TECNOLOGIA
UNIVERSIDADE NOVA DE LISBOA

Setembro 2018

Rita Leonor Franco Paulo Guimarães

Licenciatura em Biologia Molecular e Celular

ZNF675 evolutionary history and its novel role in neuronal regulatory networks

Dissertação para obtenção do Grau de Mestre em Genética Molecular e Biomedicina

Orientador: Frank M.J. Jacobs, Assistant Professor, Swammerdam Institute for Life Sciences, University of Amsterdam
Co-orientador: Gerrald Lodewijk, PhD Student, Swammerdam Institute for Life Sciences, University of Amsterdam

Júri:

Presidente: Prof. Doutora Margarida Casal Ribeiro Castro Caldas Braga, Faculdade de Ciências e Tecnologia da Universidade Nova de Lisboa
Arguente: Doutora Maria João de Jesus Nunes, Faculdade de Farmácia da Universidade de Lisboa.
Vogal: Prof. Doutor Pedro Miguel Ribeiro Viana Baptista, Faculdade de Ciências e Tecnologia da Universidade Nova de Lisboa

Setembro 2018

ZNF675 evolutionary history and its novel role in neuronal regulatory networks

Acknowledgments

I would like to express my very deep gratitude to Dr. Frank Jacobs, for the great opportunity of working in his laboratory. Thank you for the guidance, the discussions and advices, and for the trust placed in me.

My special thanks are to Gerrald Lodewijk, for being the best daily supervisor in the world ever in the universe. For teaching me how to do good science every day, for cheering me up and calming me down, and for all the conversations about everything and anything. This thesis would not have been possible without you.

I am particularly grateful for having the fellow Portuguese colleague Diana Pereira Fernandes in the laboratory, with whom I shared my achievements and frustrations. Thank you for the friendship, for all the talks over a beer and the honest advices.

I wish to acknowledge the FCT/UNL and the European Commission for the Erasmus+ SMT (student mobility for traineeships) program, for the chance to do the internship in the Netherlands.

An important thank you to my mother, father, brother and sister for loving me unconditionally. For supporting me, believing in me and for giving the necessary strength to be one year far away from you. Thank you to my whole family, especially in the Netherlands, who made me feel safe and at home.

Thank you to my friends that were close to me even when I was in another country. To my friends Anna Davidson, David Huson and Sarah Tommouhi, who walked alongside me in this project. I sincerely thank you for the friendship, all the fun activities we did together and for never letting me down! I feel very heart-warmed for having found you guys.

Finally, I would like to thank to Fred White, for the love, protection and patience. Thank you for supporting me and being by my side no matter what. For making me feel truly happy and for being my safe place.

Abstract

ZNF675 is a primate-specific KRAB-ZNF protein that arose in the last common ancestor of New World Monkeys and Old World Monkeys. It represses THE1 and MST families of retrotransposons and binds to gene promoters like MCPH1 and SESN3. Copy-number variation of ZNF675 is associated with neurological disorders, including microcephaly and intellectual disability. Unravelling the evolutionary history and the regulatory effect of ZNF675 will clarify its role in primates' evolution.

The analysis of ZNF675 sequence revealed a significant difference in zinc-finger domains between the marmoset and the human ZNF675. In a reporter assay, the marmoset ZNF675 did not bind to MSTA elements. However, the human ZNF675 had a strong repressive effect on them. This indicates structural changes in ZNF675 emerged to silence THE1 and MST retrotransposons, after New World Monkeys split from our last common ancestor. As a collateral effect, ZNF675 created a new regulatory network, by binding to gene promoters, 25% of which are involved in development and neuronal processes. At the same time, some ZNF675 binding sites in gene promoters also underwent changes, as found in the MCPH1 and SESN3 promoters. These changes affect the ZNF675 binding capacity and led to the evasion or stabilization of its regulatory function. This highlights the dynamics of genome evolution in primates.

To further investigate these findings, the CRISPR/Cas9 system was used for targeted deletion of ZNF675 and ZNF681, or for modifying promoter region sequences using homology-directed repair.

The evolutionary arms race between ZNF675 and THE1 and MST elements led to the creation of a novel ZNF675, with the ability to silence them. As a collateral consequence, ZNF675 binds to gene promoters, regulating their expression. This new regulation network might have had a strong impact on the evolution of the human brain.

Keywords: ZNF675, KRAB-ZNF protein, retrotransposons, gene regulation, evolution, brain.

Resumo

O ZNF675 pertence à família das proteínas KRAB-ZNF e surgiu há 35 milhões de anos, no último ancestral comum dos Macacos do Novo e do Velho Mundo. Esta proteína silencia especificamente retrotransposões das famílias THE1 e MST, mas também pode ligar-se a promotores de genes neuronais como o MCPH1 e o SESN3. Variações no número de cópia do ZNF675 estão associadas a doenças neurológicas como microcefalia e deficiência intelectual. Este projeto pretende desvendar a função do ZNF675 no genoma humano e a sua história evolutiva.

A análise de sequências do ZNF675 ao longo da evolução revelou uma diferença significativa entre o domínio de *zinc-fingers* do ZNF675 presente no primata sagui-de-tufos-brancos e no de humanos. Num luciferase *reporter assay*, o ZNF675 do sagui-de-tufos-brancos não se ligou ao retrotransposão MSTA. Contudo, o ZNF675 humano exerceu uma forte repressão no mesmo. Isto indica que as alterações estruturais no ZNF675 evoluíram no sentido de silenciar os retrotransposões THE1 e MST, depois da divergência dos Macacos do Novo Mundo com o nosso último ancestral comum. Consequentemente, o ZNF675 terá adquirido uma nova função regulatória ao ligar-se a promotores de genes, alguns dos quais foram sofrendo alterações ao longo da evolução dos primatas. Estas alterações levaram a uma evasão ou estabilização de função regulatória do ZNF675.

Para investigar esta ideia, uma deleção nos genes ZNF675 e ZNF681 foi gerada através da técnica CRISPR/Cas9, e conjuntamente com o mecanismo de *homology-directed repair*, para modificar sequências dos promotores.

Os resultados descritos sugerem que a luta evolutiva entre KRAB-ZNFs e os retrotransposões THE1 e MST terá levado à emergência do ZNF675, que passou a silenciar estes elementos. Colateralmente, o ZNF675 também começou a regular genes envolvidos no desenvolvimento neurológico devido à afinidade pelo seu promotor. Esta regulação colateral poderá ter tido um forte impacto na evolução do cérebro humano.

Palavras-chave: ZNF675, proteínas KRAB-ZNF, retrotransposões, regulação génica, evolução, cérebro.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

A	**APS**	Ammonium persulfate
C	**cMCPH1**	*Callithrix jacchus* MCPH1
	cSESN3	*Callithrix jacchus* MCPH1
	cZNF675	*Callithrix jacchus* ZNF675
	cDNA	Complementary DNA
	ChIP	Chromatin immunoprecipitation
	ChIP-seq	Chromatin immunoprecipitation sequencing
D	**DNA**	Deoxyribonucleic acid
	DMEM	Dulbecco's modified Eagle medium
E	***E.coli***	*Escherichia coli*
	ERV	Endogenous retrovirus
	eGFP	Enhanced green fluorescent protein
	EDTA	Ethylenediamine tetraacetic acid
G	**gDNA**	Genomic DNA
	GTEx	Genotype-Tissue Expression
	GMEM	Glasgow's MEM
	gRNA	Guide RNA
H	**HES1**	Hairy and enhancer of split homolog 1
	HiFBS	Heat inactivated fetal bovine serum
	HP1	Heterochromatin protein 1
	HBS	HEPES Buffered Saline
	H3K9me3	Histone 3 lysine 9-trimethylation
	hMCPH1	*Homo sapiens* MCPH1
	hSESN3	*Homo sapiens* SESN3
	hZNF675	*Homo sapiens* ZNF675
	HDR	Homology-directed repair
	HEK293T	Human embryonic kidney 293 T cells
	HERV	Human endogenous retroviruses
	hiPSC	Human induced pluripotent stem cells
I	**IP**	Immunoprecipitation
K	**KAP1**	KRAB-associated protein 1
	KZNF	KRAB-ZNF
	KRAB-ZNF	Krüppel-associated box zinc-fingers
L	**LCA**	Last common ancestor
	LIF	Leukemia inhibitory factor
	LINE	Long interspersed elements
	LTR	Long terminal repeats
	LB	Luria-Bertani

M	**mZNF675**	*Macaca fascicularis* ZNF675
	MCPH1	Microcephalin 1
	Mya	Millions of years ago
	MEM	Minimum essential medium
	mESC	Mouse embryonic stem cells
N	**NWM**	New World Monkeys
	NHEJ	Non-homologous end-joining
	NuRD	Nucleosome remodelling and deacetylase
O	**OWM**	Old World Monkeys
	ORF	Open reading frame
P	**P/S**	Penicillin/streptomycin
	PBS	Phosphate buffered saline
	PAGE	Polyacrylamide gel electrophoresis
	PCR	Polymerase chain reaction
	PEI	Polyethylenimine
Q	**qPCR**	quantitative PCR
R	**RT-PCR**	Reverse transcriptase PCR
	ROS	Reactive oxygen species
	RNA	Ribonucleic acid
S	**SESN3**	Sestrin 3
	SVA	SINE-VNTR-Alu element
	ssDNA	Single-stranded DNA
	SINE	Short interspersed element
	SETDB1	SET domain bifurcated 1
T	**TEMED**	Tetramethylethylenediamine
	TF	Transcription factors
	TSS	Transcription start site
	TE	Transposable element
	THE1	Transposon-like human element 1
	TRIM28	Tripartite motif-containing 28
V	**VNTR**	Variable number of tandem repeats
Z	**ZNF**	Zinc-finger

1 Introduction

1.1 Evolution of the human brain

The human brain has increased remarkably in size and complexity throughout primate evolution. The encephalization quotient, a way of measuring brain size relatively to body weight, differs significantly between human and other primates, which is not observed when comparing nonhuman primates (Williams, 2002). In fact, the human brain has undergone an overdevelopment, specifically the neocortex and cerebellum, that has been correlated to an increase in cell populations (Herculano-Houzel, Collins, Wong, & Kaas, 2007). A lot of effort has been devoted to find the underlying causes of the human brain singularity, i.e. the genetic and molecular events that led to a bigger-than-expected brain throughout primate evolution.

In order to search for the genetic basis of brain evolution in primates, differences between humans and chimpanzees have been investigated. In 1975, Mary-Claire King and Allan Wilson published a landmark article that shed light on the molecular differences and similarities of humans and chimpanzees, where they found that both primates share more than 99% of the average polypeptide coding sequence (King & Wilson, 1975). Even though years later it was described that the difference between these two species is in fact 4% (Varki & Altheide, 2005), such a small change could not explain their major anatomical and behavioural differences. However, together with the non-coding part of the genome, important regulatory regions which control gene expression, and structural genomic rearrangements might have had an impact on the emergence of the human species.

Brain evolution has been linked to natural selection and genome reshaping, of which Transposable Elements (TEs) are a major cause (Ayarpadikannan & Kim, 2014; Cordaux & Batzer, 2009; Deininger, Moran, Batzer, & Kazazian, 2003). TEs are derived from ancient viruses that infected the host millions of years ago (Mya). They were first described by the Nobel Prize Barbara McClintock, being referred to as “controlling elements” because they are able to regulate gene expression (McClintock, 1950). They have the unique ability of moving throughout the genome, or even between genomes (Cordaux & Batzer, 2009), which is also the reason why between half (Lander et al., 2001) and two-thirds (de Koning, Gu, Castoe, Batzer, & Pollock, 2011) of the human genome is composed of TEs.

Krüppel-associated box zinc-fingers (KRAB-ZNF) proteins function as transcriptional repressors of TEs, in a type-specific manner, to prevent transposition and, thereby, maintain genome integrity (Imbeault, Helleboid, & Trono, 2017; Jacobs et al., 2014; Najafabadi et al., 2015). These proteins constitute ~50% of human zinc-finger (ZNF) proteins, representing the largest and most rapidly evolving gene family (Huntley et al., 2006; Najafabadi et al., 2015). They have been described as a driving force of a new level of gene regulation in primates, in a species-specific manner (Hamilton, Huntley, Gordon, & Stubbs, 2005; Imbeault et al., 2017; Jacobs et al., 2014; Nowick, Hamilton, Zhang, & Stubbs, 2010; Shannon, Hamilton, Gordon, Branscomb, & Stubbs, 2003). Interestingly, KRAB-ZNF (KZNF) proteins were found to be differentially expressed in the human brain, in comparison to that of the chimpanzee.

It was proposed that these regulatory modifications were in part responsible for the humans' over-sized and more complex brain (Nowick, Gernat, Almaas, & Stubbs, 2009).

1.2 Transposable elements

TEs are grouped in two major classes according to their transposition mechanism: class 1 and class 2. Class 2 is composed of transposons, DNA elements that can excise and insert themselves into another region of the genome, which constitute ~3% of the human genome (Pace, Feschotte, & Ii, 2007). Class 1 comprises TEs which make use of an RNA intermediate to transpose. For this reason, they are called retrotransposons. A reverse transcriptase, encoded in the element itself, copies the RNA into complementary DNA (cDNA) which then gets integrated in another region of the genome. Within this class, TEs can even be divided in families based on the presence or absence of long terminal repeats (LTRs): LTR retrotransposons (Figure 1.01), also known as human endogenous retroviruses (HERVs), are similar to simple retroviruses, lacking only the envelope-coding gene, and constitute ~8% of human genome (Deininger et al., 2003). In contrast, non-LTR retrotransposons, which do not contain an LTR region, are further subdivided in long interspersed elements (LINEs) and short interspersed elements (SINEs), e.g. L1 and *Alu* elements, respectively. There is also a hominoid-specific element family, the SVA elements, which comprise a SINE region, followed by a variable number of tandem repeats (VNTR) and an *Alu* element. The non-LTR elements constitute ~34% of the human genome (Cordaux & Batzer, 2009).

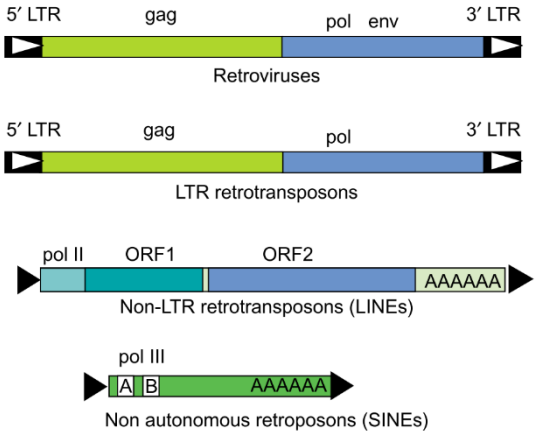


Figure 1.01 – **Schematic of retrotransposable elements.** Retrovirus is composed of three genes**:** gag (group-specific antigen), pol (reverse transcriptase) and env (envelope protein). LTR retrotransposons are similar, lacking only the env gene. Non-LTR retrotransposons (primarily LINEs in mammals) have an RNA polymerase II promoter, allowing the transcription of a full-length RNA that encodes two proteins: ORF1 and ORF2. Non-autonomous retroposons (primarily SINEs in mammals) have an RNA polymerase III promoter (A and B boxes) to transcribe a small RNA, similarly to the LINE retrotransposition machinery. *Adapted from* Deininger et al., 2003.

Retrotransposons are the major creators of genome instability by inserting themselves within or close to genes, or by inducing genomic ectopic rearrangements (Ayarpadikannan & Kim, 2014; Cordaux & Batzer, 2009; Lander et al., 2001). They can cause deleterious mutations when inserted in an exon, which might change the open reading frame (ORF) giving rise to non-functional proteins. TEs can also cause abnormal proteins, if they get inserted in an intron, eliminating the canonical splice site or even creating novel splice sites (Ayarpadikannan & Kim, 2014). On the other hand, TEs can generate ectopic rearrangement events, such as deletions, duplications and inversions, creating new genes, modifying pre-existing ones or just changing the GC content (Lander et al., 2001). Moreover, if TEs get inserted in close proximity with genes, in regulatory regions, they can change their expression levels, thus altering the network on which that gene acts on (Ayarpadikannan & Kim, 2014). Interestingly, genomic modifications induced by TEs are proposed to have accelerated evolution of gene regulatory networks over mammalian evolution (Lowe, Bejerano, & Haussler, 2007).

With the sequencing of the human genome it was possible to estimate TEs' activity status during primate evolution, by analysing the sequence divergence between species. It was shown that the activity of all transposons has declined 35-50 Mya, which means that they were active during early primate evolution (Lander et al., 2001). This includes some families of LTR (THE1A, THE1B, THE1C, MSTA and MSTB) and non-LTR retrotransposons (L1 and *Alu* elements), since they are primate-specific (Cordaux & Batzer, 2009; Pace et al., 2007). On the contrary, SVA elements have been active only during the hominid evolution. Some families of SVA, L1 and *Alu* elements are still active in the human genome (Mills, Bennett, Iskow, & Devine, 2007; H. Wang et al., 2005). Interestingly, they have been associated with a variety of genetic diseases, such as certain types of cancer, haemophilia and X-linked diseases (Deininger et al., 2003; Kazazian, 1998; Konkel & Batzer, 2010). However, TEs have also been described to play an important role in early development, being involved in essential processes such as in placental formation, meiotic recombination during gametogenesis and epigenetic inheritance (Gifford, Pfaff, & MacFarlan, 2013).

1.3 Krüppel-associated box zinc-finger proteins

KZNF proteins, the TEs' transcriptional repressors, are composed of C2H2-like ZNFs, which is the most prevalent DNA-binding motif of transcription factors (TFs) in eukaryotic organisms (Huntley et al., 2006; Thomas & Schneider, 2011). ZNF proteins often have between 1 and 30 ZNF motifs in tandem arrays, each one with two cysteines and two histidines interacting with a zinc ion, stabilizing the structure (Huntley et al., 2006; Krishna, Majumdar, & Grishin, 2003). These ZNF motifs consist of two β-sheets and one α-helix (Figure 1.02a), being the last one the DNA-binding domain by fitting in the major groove of the DNA (Lupo et al., 2013; Persikov & Singh, 2014). It is proposed that the ZNF protein binds to DNA with high specificity, as each ZNF motif contacts four consecutive base pairs (Figure 1.02b), overlapping on one, i.e. there is binding of three base pairs per ZNF motif (Persikov & Singh, 2014). There is currently no way of predicting the binding site of a ZNF protein with 100% accuracy. This is due to the possible

influence of amino acids close to those that bind the DNA, the impact of other ZNF motifs or even the possibility that not all the motifs bind to DNA (Najafabadi et al., 2015).

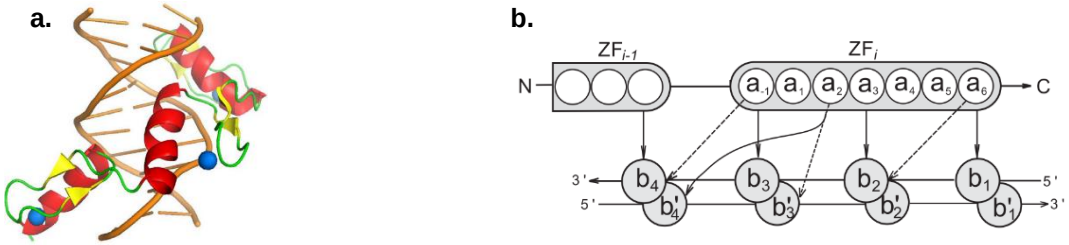


Figure 1.02 – **Representation of C2H2-like ZNF protein-DNA interaction.** (a) Structural model of a protein with three ZNF motifs interacting with DNA. Each ZNF motif has two β-sheets (yellow), one α-helix (red/orange) and a zinc ion (blue). *Adapted from* Yang, Wang, & Macfarlan, 2017. (b) Two consecutive ZNF motifs binding to DNA. Amino acids within the *i*-th ZNF are numbered according to their relative position from the start of the alpha helical domain. DNA bases b1, b2, b3 and b4 are numbered sequentially and the complementary bases are primed. The canonical contacts are shown with solid arrows, and the three additional contacts are shown with dashed arrows. *Adapted from* Persikov & Singh, 2014.

KZNF proteins are composed of a 75-amino acid KRAB domain in the N-terminal domain, in addition to the ZNF domain at the C-terminus. After DNA-binding by the ZNF domain (Figure 1.03), repression occurs through interaction of the KRAB domain with the KRAB-associated protein 1 (KAP1, or TRIM28) (Peng et al., 2000). The complex recruits SETDB1, a histone 3 lysine 9-specific methyltransferase, which in turn enhances heterochromatin protein 1 (HP1) binding (David C Schultz, Ayyanathan, Negorev, Maul, & Rauscher Iii, 2002). The nucleosome remodelling and deacetylase (NuRD) complex is also recruited by KZNF/KAP1 (D. C. Schultz, Friedman, & Rauscher, 2001), making up the repression complex, which induces heterochromatin formation and transcriptional silencing (Yang, Wang, & Macfarlan, 2017). The induced chromatin remodelling is spread in a long-range repression through H3K9me3 (histone 3 lysine 9-trimethylation) and HP1, model confirmed in a KZNF cluster, suggesting an auto-regulatory mechanism (Groner et al., 2010).

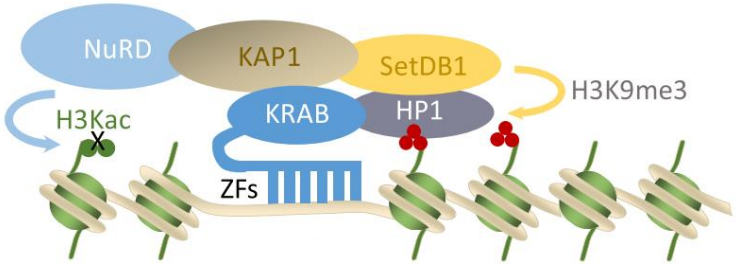


Figure 1.03 – **Representation of the KRAB-ZNF protein repression complex.** Firstly, ZNF domain binds to the DNA and the KRAB domain interacts with KAP1. The complex recruits other factors such as SETDB1, HP1 and the NuRD complex. Red circles on histone tails are H3K9m3 and green circles are deacetylation of multiple lysines. *Adapted from* Yang, Wang, & Macfarlan, 2017.

KZNF proteins are specific to tetrapod vertebrates and are often localized in familial gene clusters, as they are primarily originated by duplication events. For instance, out of 423 KZNF protein-coding loci, at least 136 are primate-specific KZNF and have diverged significantly after duplication (Huntley et al., 2006). Even though these genes coexist in the same loci, each one is under a distinct selective pressure, giving rise to novel proteins with new recognition sites, as changes occur mostly in the ZNF domain and the smallest nucleotide variation influence DNA binding (Huntley et al., 2006; Lupo et al., 2013). Segmental duplications of loci containing KZNF protein-coding genes are also the underlying cause for the observed diversity among their families, where 24 KZNF protein-coding genes are only present in great apes and hominids. Interestingly, gene duplications that are human and chimpanzee-specific have evolved in a higher rate than older genes, which may be evidence for the still ongoing evolution of regulatory factors (Nowick et al., 2010).

KZNF proteins are involved in essential physiological processes, such as development, differentiation, metabolism, cell division and cancer (Lupo et al., 2013), in a cell type-specific manner (Najafabadi et al., 2015). For instance, ZNF809 is a stem cell-specific retroviral repressor (Wolf & Goff, 2009); ZNF382 is a proapoptotic tumour suppressor (Cheng et al., 2010); ZNF268 participates in the erythroid cells' differentiation (Zeng et al., 2012); and PRDM9, a well-studied KZNF protein, determines meiotic recombination hotspots' location (Altemose et al., 2017). Moreover, KZNF proteins might have a role preventing segmental duplication events by suppressing nonallelic homologous recombination, stabilizing the genome (Najafabadi et al., 2015).

1.4 Evolutionary arms race between transposable elements and KZNF proteins

In 2009, Emerson and Thomas hypothesized that the evolution of KZNF proteins was being "driven by an arms race with their viral or transposon targets" (Emerson & Thomas, 2009). The first two main evidences that point in that direction were a strong correlation between the number of ZNF domains and LTR retrotransposons in the vertebrate genome; and whenever a new endogenous retrovirus (ERV) family arose in the genome, it was usually followed by a burst of ZNF genes. (Thomas & Schneider, 2011). This model was later supported by Jacobs and colleagues, who have shown that retrotransposons would evade repression due to mutations within the binding site of their primate-specific KZNF (Figure 1.04). In response, KZNF proteins evolved to keep the retrotransposons silenced, in an evolutionary arms race (Jacobs et al., 2014). Jacobs et al. showed this through luciferase reporter assays in mouse embryonic stem cells (mESC). An SVA element was cloned upstream of a minimal SV40 promoter, which controls the expression of the luciferase reporter. The constructs were co-expressed with different versions of the ZNF91: the human, the version probably present in the last common ancestor (LCA) of humans and gorillas, in the LCA of humans and orangutans, and the macaque version. The experiment revealed a difference in luciferase activity, being strongly repressed with the human ZNF91, and having a decreasing effect in the older species. This implies that the changes ZNF91 underwent improved its ability to bind and silence the SVA element.

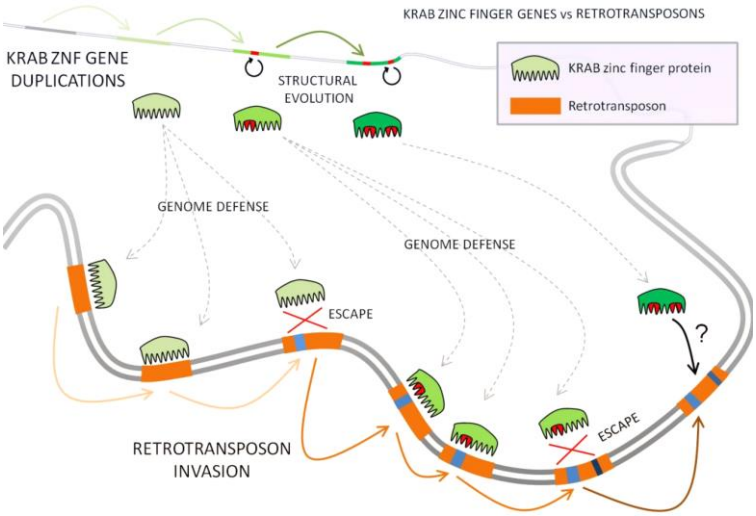


Figure 1.04 – **Evolutionary arms race between KZNF proteins and retrotransposons**. Retrotransposons have the ability to transpose to another location in the genome until a KZNF protein emerges and silences them. Later, retrotransposons escape repression due to mutations. KZNF protein mutates too, recovering the ability to silence the retrotransposons. *Adapted from* Frank Jacobs, frankjacobslab.com (retrieved in: 20/09/2018).

KZNF proteins have been protecting our genome integrity against transposable elements' emergence since 100 Mya, in the early mammalian radiation (Najafabadi et al., 2015). Throughout evolution the ZNF domain of KZNF proteins have undergone significant changes, whilst the KRAB domain stayed conserved, meaning that recently emerged TEs can be silenced by the same KAP1-mediated repression complex (Huntley et al., 2006; Thomas & Schneider, 2011). These DNA-binding proteins make up a suitable and smart way of fighting TEs movement, preventing whole-genome instability.

The arms race can explain in part the rapid expansion of KZNF protein families, but not entirely since the majority of TEs have lost the transposition potential and the correspondent KZNF protein still has its full binding capacity (Imbeault et al., 2017). As a collateral consequence of the evolutionary battle, these family of repressors have acquired novel regulatory functions. They are involved in regulation of gene expression, acting either directly by binding promoter regions (Lupo et al., 2013), or indirectly by binding TEs, which they benefit from and use as regulatory platforms nearby genes (Imbeault et al., 2017; Najafabadi et al., 2015).

1.5 ZNF675 and its target genes

ZNF675 is a KZNF gene which is expressed in the brain according to the Genotype-Tissue Expression (GTEx) project (Lonsdale et al., 2013). According to DECIPHER (Firth et al., 2009), it is associated with several neuronal disorders, such as microcephaly, macrocephaly, intellectual disability, autism and cerebellar atrophy. Like many other KZNF proteins, ZNF675 is a primate-specific protein and was originated in the LCA of New World Monkeys (NWMs) and Old World Monkeys (OWMs) (Figure

1.05) (Imbeault et al., 2017). However, a duplication event occurred in the ancestor of humans and OWMs, giving rise to ZNF681, with a quite different DNA recognizing ZNF domain. Furthermore, ZNF675 itself underwent significant structural modifications, which changed its DNA-recognizing motifs (Lodewijk et al., unpublished).

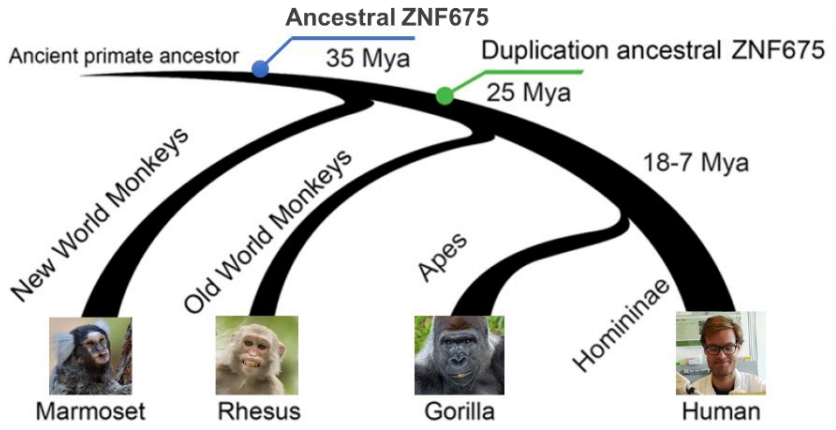


Figure 1.05 – **Evolutionary history of ZNF675.** ZNF675 first arose in the LCA of NWMs and OWMs. It underwent a duplication event in the LCA of OWM and apes, giving rise to ZNF681. *Adapted from* Lodewijk et al., unpublished.

ZNF675 represses LTR retrotransposons, specifically THE1B/C/D/A (transposon-like human element 1) and MSTA/B/B1 (Najafabadi et al., 2015), acting in a cell-type specific manner (Imbeault et al., 2017). Interestingly, it appears that a great expansion of THE1 and MST elements happened in the LCA of NWM and OWM, which was right before the structural changes of ZNF675 (Lodewijk et al., unpublished).

Chromatin immunoprecipitation sequencing (ChIP-seq) is a technique widely used to detect the interactions between DNA and proteins, and allows the identification of the DNA sequence involved, i.e. the binding site (Schmidt et al., 2009). ChIP-seq data for ZNF675 (GEO accession number: GSM1407624) (Najafabadi et al., 2015), suggests that besides binding TEs, ZNF675 binds to several genes related to neurodevelopment or involved in neurological disorders such as HES1 (hairy and enhancer of split homolog 1), MCPH1 (microcephalin 1) and SESN3 (sestrin 3).

HES1 is a transcription factor involved in the regulation of the central nervous system development, by maintaining neuronal progenitors' self-renewal and proliferative capacity (Nakamura et al., 2000). Preserving the neuroprogenitors' pool is crucial to reach the correct number of neuronal cells, which is accomplished by HES1 repression of the cyclin-dependent kinase inhibitor p27Kip1 (Murata et al., 2005).

MCPH1 is considered a pleiotropic factor, since it has a broad range of different functions. It is implicated in DNA damage repair, genomic stability, chromatin remodelling and cell division, by regulating the G2/M checkpoint (Liu, Zhou, & Wang, 2016) and the chromosomes biorientation (Arroyo et al., 2017). MCPH1 is also a tumour suppressor, found to be in low levels in several types of cancer,

including lung and breast cancer (Mantere et al., 2016; Zhou et al., 2016). MCPH1 plays an important role in regulating brain size during development, being the first gene to be associated with primary microcephaly. Microcephaly is a disorder characterized by a small brain, specifically the cerebral cortex, and a head circumference with more than 3 standard deviations below the age-related mean (Jackson et al., 2002). Genes involved in brain development are of particular interest since they might have played a role in humans' brain enlargement during evolution. In fact, several studies concerning MCPH1 gene have suggested that it was under positive selection in the lineage from OWMs ancestor to humans and great apes, especially after the split of lesser apes (Evans, Anderson, Vallender, Choi, & Lahn, 2004; Y. Q. Wang & Su, 2004).

SESN3 gene is a member of the highly conserved Sestrins family, characterized by having antioxidant properties. It is regulated by FoxO, subfamily of forkhead transcription factors, in response to high levels of reactive oxygen species (ROS) induced by Ras proteins. Therefore, SESN3 have a determinant role against oxidative stress and in preventing apoptosis (Hagenbuchner et al., 2012; Nogueira et al., 2008). In case of disease, the protective activity of SESN3 is especially important, which is why this protein has been described to act in certain types of diseases, such as cancer (colorectal and lung cancer), metabolism disorders, diabetes and cardiovascular disorders (M. Wang et al., 2018). However, regarding the brain, SESN3 has been associated with epilepsy. In the human epileptic hippocampus, SESN3 was found to act as a *trans*-acting genetic control of a proconvulsant gene-regulatory network, which positively regulates proconvulsive cytokines and Toll-like receptor signalling gene (Johnson et al., 2015).

Genes involved in major cellular processes need to be under a strict chronological and spatial regulatory mechanism in order to be expressed in a controlled manner. Consequently, changes in the regulatory machinery have a strong impact on gene expression, thus altering the downstream process that gene would act on. The LCA of humans and NWMs was probably the first primate suffering from the regulatory changes caused by the ZNF675 emerging, around 35 Mya.

1.6 Aim of this study

It is hypothesised that ZNF675 arose in the genome to repress THE1 and MST elements, but the promoters of several genes were caught in the arms race, such as HES1, MCPH1 and SESN3. Here, ZNF675 repression of MCPH1 and SESN3 is investigated in the context of evolution. By exploring how ZNF675 structural changes might have affected its DNA binding capacity and how small modifications in the promoters of these genes over the past millions of years have fine-tune the binding. ZNF675 structural changes over time are investigated with the Princeton predictive DNA recognition tool (Persikov & Singh, 2014), where DNA motifs recognized by ZNF675 are unravelled. The human ZNF675 and marmoset ZNF675 binding affinity to the MCPH1 and SESN3 promoter regions is analysed through Luciferase Assays in mESC, where there is no endogenous ZNF675. The MCPH1 and SESN3

promoters' changes during primate evolution are assessed with a multiple sequence alignment of sequences from NWMs or basal primates to humans. To further investigate these findings, the CRISPR/Cas9 system is used to induce a targeted deletion of the ZNF675 and ZNF681 genes. This technique is also used, together with the homology-directed repair (HDR) mechanism, to change the ZNF675 binding site in human cells by the marmoset. Finally, a chromatin immunoprecipitation (ChIP) of human and marmoset ZNF675 is performed.

2 Materials and methods

2.1 ZNF675 and ZNF681 structural analysis

The ZNF675 sequences from human (hZNF675), crab-eating macaque (mZNF675) and marmoset (cZNF675), and the ZNF681 from human (hZNF681) and crab-eating macaque (mZNF681), were previously obtained through the analysis of sequences from the Whole-Genome Shotgun contigs on GenBank, NCBI. Using the predictive DNA recognition tool (Persikov & Singh, 2014), ZNF motifs of ZNFs were characterized as functional, if the ZF score was higher than 25, or degenerated, and on which DNA motifs they bind to. With that information, the structural changes of all ZNFs were analysed.

2.2 Analysis of ZNF675 binding to THE1B-int elements

Using the RepeatMasker from the UCSC Genome Browser (Kent et al., 2002), all the THE1B-int sequences, which is the internal part of the elements, were obtained from the human reference genome GRCh37/hg19. Sequences were filtered by sequence similarity score above 8000 and length greater than 800bp. In order to get the THE1B-int elements that are bound and not bound by ZNF675, the selected sequences were intersected with Model-based Analysis of ChIP-seq (MACS) peak calling data from the ChIP-seq data for ZNF675 (GEO accession number: GSM1407624) (Najafabadi et al., 2015). ZNF675-bound was defined as 50% of the ZNF675 peak was overlapping with a THE1B-int element, whereas unbound was defined as no overlap at all. From the bound 641 and unbound 1505 THE1B-int elements, 15 sequences with the higher score of each group were aligned using the Multiple Sequence Alignment MUSCLE in UGENE (Okonechnikov et al., 2012).

2.3 ZNF675 ChIP-seq data for promoter binding analysis

Previously, raw ZNF675 ChIP-seq data fastq files (GSM2466628, GSM1407624) (Imbeault et al., 2017; Najafabadi et al., 2015) were trimmed with Trimmomatic using the following parameters: ILLUMINACLIP Truseqv3 single end for Hiseq, SLIDNIGWINDOW: 4:20, MINLEN: 30. Then, reads were mapped to hg19 or hg38 using Bowtie2, using very sensitive end-to-end settings. The BAM files were converted to bigwig using DeepTools bamCoverage (bin size = 1). Bigwig files were analysed using the UCSC genome browser. Using the table browser, bigwig files were filtered for regions with sequencing depth of >20 reads. Promoter regions were selected from 5kb upstream of genes' transcription start site (TSS) until 1kb downstream. Both promoter regions and the filtered bigwig files were intersected with any percentage of overlap. From the output list of intersections, regions with a lot of noise were excluded. Genes being influenced by the ZNF675 were annotated, as any class of repeats,

long non-coding RNA, antisense RNA and small nuclear RNA, if the ZNF675 binding site was in those elements. The list was further shortened by selecting neuronal related genes, involved in cell signalling or associated with neuronal diseases.

2.4 Retinoic Acid treatment of a human neuroblastoma cell line

Cells from the human neuroblastoma cell line SK-N-SH (Sigma-Aldrich #86012802) were seeded at a density of 5 x 10^4 cells per well in 2x 12-wells plates and grown in minimum essential medium (MEM, ThermoFisher Scientific #31095029) with 10% heat inactivated fetal bovine serum (HiFBS, ThermoFisher Scientific #10500-064) and 1x penicillin/streptomycin (P/S, ThermoFisher Scientific #15140122). After 24 hours (T0) cells were isolated in TRIzol (ThermoFisher Scientific #15596018) and kept in –80°C. Cells still in culture were treated with 10µM retinoic acid (RA, Sigma-Aldrich #R2625) or ethanol (Merck Millipore #1.00983.1000) as control. 24 hours later, cells were isolated in TRIzol and the RNA extracted, including from the T0 cells, following the Direct-zol™ RNA MicroPrep protocol (Zymo Research). A one-step RT-PCR (QIAGEN) was performed according to the manufacturer's protocol, on 20ng of RNA per reaction, with a program of 25 cycles. The following primers were used:

HES1_qFw1: GATGCTCTGAAGAAAGATAGC HES1_qRv1: TTCATGCACTCGCTGAAGC

The amplicons were analysed on a 1% agarose gel electrophoresis and the relative density of the DNA was analysed using ImageJ (Abràmoff, Magalhães, & Ram, 2004).

2.5 MCPH1 and SESN3 promoter region in primates

MCPH1 and SESN3 promoter sequences were obtain from the human reference genome GRCh37/hg19 on the UCSC Genome Browser. The sequences were then converted to other primate's genomes, being further described in the following table:

Table 2.01 – **Description of each primate species used on the analysis.** The columns from the first until the fifth correspond to: primate's group, primate's name, genome assembly used, coordinates of the sequence used of the MCPH1 and SESN3 genes' promoters.

		Genome Assembly	*MCPH1 Coordinates*	*SESN3 Coordinates*
Humans	Human	GRCh37/hg19	chr8:6,262,177-6,264,232	chr11:94,964,222-94,966,486
Apes	Chimpanzee	Pan_tro 3.0/panTro5	chr8:6,766,727-6,768,784	chr11:95,311,011-95,313,276
	Bonobo	MPI-EVA panpan1.1/panPan2	chr8:6,407,128-6,409,186	chr11:93,772,279-93,774,543

	Gorilla	gorGor3.1/gorGor3	chr8:6,207,578-6,209,653	chr11:92,578,851-92,581,113
	Orangutan	WUGSC 2.0.2/ponAbe2	chr8:6,674,658-6,676,746	chr11:91,241,710-91,243,978
	Gibbon	GGSC Nleu3.0/nomLeu3	chr4:145,850,941-145,853,014	chr15:40,467,613-40,469,875
Old World Monkeys	Green Monkey	Chlorocebus_sabeus 1.1/chlSab2	chr8:6,173,985-6,175,851 + chr8:6,190,182-6,190,384	chr1:86,486,434-86,488,709
	Crab-eating Macaque	Macaca_fascicularis_5.0/macFas5	chr8:6,619,567-6,621,649	chr14:91,115,965-91,118,237
	Rhesus Macaque	BCM Mmul_8.0.1/rheMac8	chr8:6,570,411-6,572,491	chr14:87,860,620-87,862,893
	Baboon	Baylor Panu_2.0/papAnu2	chr8:6,062,211-6,064,321	chr14:84,999,969-85,002,238
	Golden Snub-nosed Monkey	Rrox_v1/rhiRox1	KN298329v1:105,736-107,790	KN300223v1:207,295-209,579
New World Monkeys	Marmoset	WUGSC 3.2/calJac3	chr13:5,720,181-5,722,193	chr11:42,482,819-42,485,066
	Squirrel Monkey	Broad/saiBol1	JH378324:74,869-76,826	JH378333:178,252-180,348
Basal Primates	Mouse Lemur	Mouse lemur/micMur2	–	KQ057561v1:1,916,768-1,919,029
	Bushbaby	Broad/otoGar3	–	GL873520:268,417-270,652

The sequences were aligned using the Multiple Sequence Alignment MUSCLE in UGENE and their conservation was analysed.

2.6 Luciferase Reporter Assay on human and marmoset MCPH1 and SESN3 promoters

Cloning of sequences into pGL3 vector – A promoter region of 200bp from human and marmoset MCPH1 and SESN3 were obtained on the UCSC Genome Browser, further characterized in the following table:

Table 2.02 – **Description of human and marmoset promoter regions of MCPH1 and SESN3.** First column corresponds to the primate's name, second is the genome assembly used, third and fourth are the coordinates fo the sequences used of the MCPH1 and SESN3 genes' promoters.

	Genome Assembly	*MCPH1 coordinates*	*SESN3 coordinates*
Human	GRCh37/hg19	chr8:6,263,654-6,263,853	chr11:94,964,744-94,964,943
Marmoset	WUGSC 3.2/calJac3	chr13:5,721,612-5,721,811	chr11:42,484,350-42,484,543

These sequences were repeated 3 times and the following restriction sites were added: KpnI at the 5' end, and XhoI or BglII for the 3' end. The constructs were synthesized by GeneArt from ThermoFisher Scientific and cloned into a pGL3-Promoter luciferase vector (Promega #E1761). Ligation and transformation were done following the Quick ligation kit recommended protocol. pGL3-hMCPH1, pGL3-cMCPH1, pGL3-hSESN3 and pGL3cSESN3 were Sanger sequenced.

Cell Transfection and Luciferase Assay – Mouse 46c embryonic stem cells (Ying, Stavridis, Griffiths, Li, & Smith, 2003) were seeded from a full 100mm dish in a dilution of ¼ cells per well, in 24-wells plate coated with 0.1% gelatin. The cells were cultured in Glasgow's MEM (GMEM, ThermoFisher Scientific #11710-035) complete (10% HiFBS, 1x P/S, 5.5nM β-mercaptoethanol (ThermoFisher Scientific #31350010) 1x Non-essential aminoacids (ThermoFisher Scientific #11140035) and 1x sodium pyruvate (ThermoFisher Scientific #11360039)) supplemented with 1x leukemia inhibitory factor (LIF, ESGRO/Merck Millipore #ESG1106). On the following day the medium was changed to GMEM complete without P/S. Cells were transfected using Lipofectamine™ 3000 reagent (ThermoFisher Scientific), according to the manufacturer's recommended protocol. The transfection mix per well was composed of 5ng Red Fluorescent Protein (RFP), 5ng pRenilla (Promega #E2231) and 50ng pGL3 luciferase vector: pGL3-EV, pGL3-MSTA (previously cloned), pGL3-hMCPH1, pGL3-mMCPH1, pGL3-hSESN3 or pGL3-mSESN3. Per luciferase vector, it was added 440ng pCAG-EV (with 90ng pBluescript), pCAG-hZNF675-HA or pCAG-mZNF675-HA. 6 hours after the transfection the medium was changed again to GMEM complete with 1x LIF. On the next day the cells were washed once with Phosphate Buffered Saline (PBS, Lonza BioWhittaker™ #BE17-516F/12) and lysed using 100µl Passive Lysis Buffer from Dual-Luciferase® Reporter Assay System (Promega). The cell suspension was incubated for 15min incubation on an orbital shaker (250rpm) at room temperature and for 15min at -80°C. The Dual Luciferase Reporter Assay was performed following the manufacturer's recommended protocol.

2.7 Cloning of ZNF675-EGFP and ZNF681-EGFP fusion constructs

The enhanced green fluorescent protein (eGFP) was amplified by a PCR on pCAG-eGFP expression vector (Addgene #11150), with Phusion High-Fidelity DNA Polymerase (ThermoFisher Scientific). Two sets of primers were used, with different restriction sites, as follows:

pEGFP_ZNF675_PstI_F1: AGCTCTGCAGAACTGGAACGTGTACCCATACGATGGCGGAGGTGGAGGCGGAGGTGGAGTGAGCAAGGGCGAGGAGCTGTTC

pEGFP_ZNF675_NdeI_F2: AGCTCATATGATGGCGGAGGTGGAGGCGGAGGTGGAGTGAGCAAGGGCGAGGAGCTGTTC

pEGFP_ZNF675_NotI_R1: AGCTGCGGCCGCTCACTTGTACAGCTCGTCCATGCCG

Hereafter, the underlined sequence is the restriction site.

PCR products were analysed on a 1% agarose gel electrophoresis. Amplicons were excised, and DNA was extracted using QIAEX II Gel Extraction kit (QIAGEN).

The human, crab-eating macaque and marmoset ZNF675 coding sequences were previously obtained through the analysis of sequences from the Whole-Genome Shotgun contigs on GenBank, NCBI. An HA-tag sequence was added to the end of each ZNF and the constructs were synthesized by GeneArt from ThermoFisher Scientific. The vectors and the eGFP fragment were digested with the restriction enzymes: PstI (ThermoFisher Scientific, #ER0611) for pMA-hZNF675-HA; NdeI (ThermoFisher Scientific #ER0581) for pMA-mZNF675-HA, pMA-cZNF675-HA and pMA-hZNF681-HA; and NotI (ThermoFisher Scientific #ER0591) for every construct. The digested fragments were analysed with 1% agarose gel electrophoresis, the DNA excised and extracted using QIAEX II Gel Extraction kit. The vectors were treated with Alkaline Phosphatase (New England Biolabs #M0371S) and purified with DNA Clean & Concentrator-5 Kit (Zymo Research). The linearized vectors were ligated with the eGFP fragment in a 1:10 molar ratio using the Quick ligation kit (New England Biolabs). Competent *E.coli* DH5α was transformed with the construct following the manufacturer's recommended protocol. Bacteria was grown overnight in Luria-Bertani (LB)– agar plates (Merck Millipore #1.10285.0500, VWR Chemicals #84609.05) at 37°C. On the next day colonies were picked and grown in LB medium with ampicillin (Roche #10835242001) overnight. The plasmids were then isolated from the bacteria using the Plasmid Miniprep Kit (QIAGEN). The constructs were cloned into the pCAG expression vector (Addgene #51142) using the restriction enzymes EcoRI (ThermoFisher Scientific #ER0271) and NotI, for the 5' and 3' end, respectively. Finally, the pCAG-hZNF675-GFP, pCAG-mZNF675-GFP, pCAG-cZNF675-GFP and pCAG-hZNF681-GFP were Sanger sequenced.

2.8 MCPH1 promoter region on different primate species

Genomic DNA from Chimpanzee, Gorilla, Orangutan, Rhesus Macaque, Baboon, Gibbon and Marmoset were obtained from the Biomedical Primate Research Centre, in Rijswijk. Primers were designed in the MCPH1 promoter of every primate, which are the following:

MCPH1_Prom_F2_AllPrimates: AAGTACGTGAGTCTTCAATG

MCPH1_Prom_F3_Marmoset-AAGTACTTGGGTCTTCAATG

MCPH1_Prom_F4_Marmoset: TAGGCACCGTTGTCCTCGTTAC

MCPH1_Prom_R2_OWM: TTCCGGGATTTGTCTGTAGC

MCPH1_Prom_R3_Orangutan: TTTCCGGGATTTCTTTGTAGG

MCPH1_Prom_R4_Gibbon: TTTCCGGGATTTCTCTAGGTG

MCPH1_Prom_R5_Marmoset: GCGGGAGGCTGGCGGGAGCT

MCPH1_Prom_R6_Marmoset: CAGGAAGTTCCTCACCTTTC

MCPH1_Prom_R7_Marmoset: GTAACGAGGACAACGGTGCCTA

Genomic regions of 1200bp were amplified with Phusion High-Fidelity DNA Polymerase on 20ng genomic DNA (gDNA) per reaction. PCR products were analysed on a 1% agarose gel electrophoresis. Amplicons were extracted from gel using the QIAEX II Gel Extraction kit and purified with DNA Clean & Concentrator-5 Kit. A second PCR was performed, with the same DNA Polymerase, with primers with restriction sites, as shown below:

MCPH1_Prom-KpnI_F2_AllPrimates: ATGCGGTACCAAGTACGTGAGTCTTCAATG

MCPH1_Prom-NheI_R2_GreatApes (human, chimp, gorilla): ATGCGCTAGCTTTCCGGGATTTCTCTGTAGG

MCPH1_Prom-NheI _R2_Baboon: ATGCGCTAGCTTCCGGGATTTGTCTGTAGC

MCPH1_Prom-NheI _R8_Rhesus: ATGCGCTAGCTTCCGGGATTTATCTGTAGC

MCPH1_Prom-NheI _R3_Orangutan: ATGCGCTAGCTTTCCGGGATTTCTTTGTAGG

MCPH1_Prom-NheI _R4_Gibbon: ATGCGCTAGCTTTCCGGGATTTCTCTAGGTG

The amplicons were analysed on a 1% agarose gel electrophoresis and excised from the gel.

2.9 Replacement of human ZNF675 binding site in the MCPH1 promoter by the marmoset

Cloning of guide RNAs into pX330 vector – The guide RNAs (gRNAs) were designed upstream and downstream of the ZNF675 binding site in the MCPH1 promoter. To clone unidirectionally, on the 5' end of both sense (s) and antisense (as) oligos an overhang was added that can hybridize with DNA ends created when digesting pX330-U6-Chimeric_BB-CBh-hSpCas9 vector (Addgene #42230) with BbsI (ThermoFisher Scientific #ER1011). The sequences are the following:

pX330-MCPH1_bs_up_sense: CACCGCCGCACAAAGCCGTCCGCT

pX330-MCPH1_bs_up_antisense: AAACAGCGGACGGCTTTGTGCGGC

pX330-MCPH1_bs_down_sense: CACCGCGCGCCACGTTACTGTTCA

pX330-MCPH1_bs_down_antisense: AAACTGAACAGTAACGTGGCGCGC

10pmol of the gRNAs were mixed in 10x Annealing Buffer (10mM Tris pH8.0, 50mM Sodium chloride, 1mM ethylenediamine tetraacetic acid (EDTA) (Addgene)) and annealed using the thermocycler, where the temperature was increased up to 95°C for 10min, and slowly decreased until 25°C, in 30min total. The annealing efficiency was analysed with a 2% agarose gel electrophoresis. The gRNAs were then cloned into the pX330 with BbsI following the Quick ligation kit recommended protocol, with the single modification of adding only 0.2pmol of gRNA mix. Constructs were Sanger Sequenced.

Testing of gRNAs – Human embryonic kidney 293 T cells (HEK293T, uncertain origin) were seeded at passage 3+12 from a full 60mm dish in a dilution of ¼ cells per well, in a 12-wells plate and cultured in Dulbecco's Modified Eagle Medium (DMEM) (1x)+GlutaMax™ (ThermoFisher Scientific #61965026) with 10% HiFBS and 1x P/S. On the following day the medium was changed to DMEM+GlutaMax without 1x P/S. Using Lipofectamine™ 3000 reagent the cells were transfected with

250ng of pX330-gRNA or pX330-EV, 50ng of pCAG-GFP and 200ng of pBluescript, one well per gRNA, according to the manufacturer's recommended protocol. The medium was changed again to DMEM+GlutaMax with 10% HiFBS and 1x P/S, 6 hours after. 48 hours after transfection, the cells were lysed and the DNA isolated using Quick-DNA Miniprep Plus Kit (Zymo Research), following manufacturer's recommended protocol. A PCR was performed on 25ng of gDNA, using Phusion High-Fidelity DNA Polymerase, with the primers MCPH1_Prom_F2_primates and MCPH1_Prom_R2_greatapes.

10% of PCR product was analysed on a 1% agarose gel electrophoresis. Amplicon was purified using DNA clean & Concentrator-5 Kit and annealed, following the same protocol. An 8% Polyacrylamide gel electrophoresis (PAGE, 8% acrylamide (Bio-Rad #1610146), 1x TAE Buffer (Tris base (Sigma-Aldrich #10708976001), acetic acid (Merck Millipore #64-19-7), EDTA (Sigma-Aldrich #E5134)), 0.1% ammonium persulfate (APS, Sigma-Aldrich #A3678), 0.00001% Tetramethylethylenediamine (TEMED, Sigma-Aldrich #T9281)) was done with 100ng of gDNA. Samples migrated at 90V until the loading dye left the gel, then the gel was stained with 0.005% ethidium bromide in 1x TAE for 15min on orbital shaker at 50rpm. The gel was imaged using ChemiDoc™ MP Imaging System (Bio-Rad).

HDR mediated ZNF675 binding sequence replacement – The chosen gRNA was MCPH1_bs_down and a single-stranded DNA (ssDNA) oligo was designed with two homology-arms with 30bp each. In between the arms the sequence is from the marmoset genome WUGSC 3.2/calJac3. The oligo has on both ends three phosphorothioate bonds, as shown below:

MCPH1_ssODN_caljac:
G*A*G*CAGGCAGGGCGGTGCGTCTGGCCCTGAGCTGTAACGTGGCGCGCCGGCCCCAGGGGTGTCGGGCTGGAGTGGG
GGCGGAAGGCGCCGAAAGTCCGCACAAAGCCGTCCGC*T*G*G

* Indicate phosphothiorate bonds.

HEK293T were seeded at passage 3+22 from a full 100mm dish in a dilution of ¼ cells per well, in a 12-wells plate and cultured in DMEM+GlutaMax with 10% HiFBS and 1x P/S. Cells were transfected on the following day using Lipofectamine™ 3000 reagent with 250ng of pX330-MCPH1_bs_down or pX330-EV, 50ng of pCAG-GFP and 200ng of pBluescript. On 2 wells with the pX330- MCPH1_bs_down, the cells were co-transfected with 10nM and 30nM of MCPH1_ssODN_caljac. 48 hours after transfection, the cells were lysed and the DNA isolated using Quick-DNA Miniprep Plus Kit. A PCR was performed on 25ng of gDNA, using Phusion High-Fidelity DNA Polymerase, with following primers:

MCPH1_homJac_Fw1: TGAGGTCTGGAGGTACTCCTG MCPH1_homJac_Rv1: GGACTTTCGGCGCCTTCC

MCPH1_homJac_Rv2: CCACTCCAGCCCGACACC

Amplicons were analysed on a 1% agarose gel electrophoresis.

2.10 ZNF675/ZNF681 KO using CRISPR/Cas9 system

Cloning of gRNAs into pX330 vector – The gRNAs were designed upstream (ex1up1 and ex1up2) and downstream (ex1down1, ex1down2, ex1down3 and ex1down4) of the first exon of ZNF681. The sequences are the following:

pX330-681_ex1up1_sense: CACCGGGCAAATACCCTGTTGGGT

pX330-681_ex1up1_antisense: AAACACCCAACAGGGTATTTGCCC

pX330-681_ex1up2_sense: CACCGACTATGGCTGCCAGAAACC

pX330-681_ex1up2_antisense: AAACGGTTTCTGGCAGCCATAGTC

pX330-681_ex1down1_sense: CACCGCGCCATCTTATGGTCCAAG

pX330-681_ex1down1_antisense: AAACCTTGGACCATAAGATGGCGC

pX330-681_ex1down2_sense: CACCGAAGTGGACTGAGGCCGCCT

pX330-681_ex1down2_antisense: AAACAGGCGGCCTCAGTCCACTTC

pX330-681_ex1down3_sense: CACCGCGTCAGGGGCGAATCCTGA

pX330-681_ex1down3_antisense: AAACTCAGGATTCGCCCCTGACGC

pX330-681_ex1down4_sense: CACCGGCGAATCCTGACGGGTGCC

pX330-681_ex1down4_antisense: AAACGGCACCCGTCAGGATTCGCC

The cloning strategy used was the same as previously described for gRNAs. The gRNAs for ZNF675 were previously designed upstream and downstream of exon 4 and cloned into a pX330 with BbsI. The sequences are the following:

pX330-675_ex4up2_sense: CACCGTACACTTAATTATGGCCTG

pX330-675_ex4up2_antisense: AAACCAGGCCATAATTAAGTGTAC

pX330-675_ex4down1_sense: CACCGATATAAGTGACACATGAGG

pX330-675_ex4down1_antisense: AAACCCTCATGTGTCACTTATATC

Testing of gRNAs for ZNF681 – HEK293T were seeded at passage 3+1 at a density of 3.5 x 10^5 cells per well, in a 12-wells plate. Cells were cultured in DMEM+GlutaMax with 10% HiFBS and 1x P/S. On the following day the medium was changed to DMEM+GlutaMax without 1x P/S. Using Lipofectamine™ 3000 reagent the cells were transfected with 250ng of pX330-gRNA or pX330-EV, 50ng of pCAG-GFP and 200ng of pBluescript, one well per gRNA, according to the manufacturer's recommended protocol. The medium was changed again to DMEM+GlutaMax with 10% HiFBS and 1x P/S, 6 hours after. 48 hours after transfection, the cells were lysed and the DNA isolated using Quick-DNA Miniprep Plus Kit, following manufacturer's recommended protocol. A PCR was performed on 25ng of gDNA from the transfected cells, using LongAmp® Taq DNA Polymerase (New England Biolabs), with the following primers:

681_ex1up_Fw1: ATCCAGGAAGATTTGAATGC 681_ex1down_Rv4: AAAAGCTCCTGGATTGTAGG

10% of PCR product was analysed on a 1% agarose gel electrophoresis. Amplicon was purified using DNA clean & Concentrator-5 Kit and annealed, as previously described. An 8% PAGE was done with 100ng of gDNA, following the same protocol, and the gel was imaged using ChemiDoc Imager.

ZNF681 KO and screening – HEK293T cells were seeded at passage 3+14 from a full 60mm dish in a dilution of ¼ cells per well in a 12-wells plate and cultured in DMEM+GlutaMax with 10% HiFBS and 1x P/S. Cells were transfected, using Lipofectamine™ 3000 reagent, with 2x 225ng of pX330-gRNA or 450ng of pX330-EV, and 50ng of pCAG-GFP. The gRNA pairs are as follows:

pX330-681_ex1up1 + pX330-681_ex1down1 pX330-681_ex1up1 + pX330-681_ex1down2

pX330-681_ex1up1 + pX330-681_ex1down3 pX330-681_ex1up1 + pX330-681_ex1down4

48hours after transfection, the cells were lysed and the DNA isolated using Quick-DNA Miniprep Plus Kit, following manufacturer's recommended protocol. A PCR was on 25ng of gDNA, using LongAmp® Taq DNA Polymerase with the primer pair 681_ex1up_Fw1/Rv4. Amplicons were analysed on a 1% agarose gel electrophoresis.

ZNF675/ZNF681 KO and screening – HEK293T cells were seeded at passage 9+16 from a full 60mm dish in a dilution of ¼ cells per well, in a 12-wells plate and cultured in DMEM+GlutaMax with 10% HiFBS and 1x P/S. Cells were transfected, using Lipofectamine™ 3000 reagent, with 2x 225ng of pX330-gRNA or 450ng of pX330-EV, and 50ng of pCAG-GFP. The gRNA pairs for ZNF675 and ZNF681 are as follows:

pX330-675_ex4down1 + pX330-681_ex1up1 pX330-675_ex4down1 + pX330-681_ex1down3

pX330-675_ex4down1 + pX330-681_ex1down4 pX330-675_ex4up2 + pX330-681_ex1up1

48 hours after transfection, the cells were lysed and the DNA isolated using Quick-DNA Miniprep Plus Kit, following manufacturer's recommended protocol. A PCR was performed on 25ng of gDNA, using LongAmp® Taq DNA Polymerase with the primers 681_ex1up_Fw1 and:

675_ex4up_Rv1: GCATAGTTGGTAAACATTGG 675_ex4down_Rv1: GGGATTACTAGAAATGTTTGTCC

Amplicon was analysed on a 1% agarose gel electrophoresis.

HDR-mediated KO – The gRNA pair 675_ex4down1/681_ex1down3 was chosen and a ssDNA oligo was designed with two homology-arms of 30bp each, with the following sequence:

675-681-dKO-ssDNA:
A*A*T*G*CAGAATATTACCCTGAACACCCACCTCTCGAGGGATTCGCCCCTGACGACCCTCCCGT*G*G*T*C

* Indicate phosphothiorate bonds.

HEK293T cells were seeded at passage 9+24 from a full 60mm dish in a dilution of ¼ cells per well, in a 12-wells plate and cultured in DMEM+GlutaMax with 10% HiFBS and 1x P/S. Cells were

transfected using Lipofectamine™ 3000 reagent, with 225ng of pX330-675_ex4down1 and pX330-681_ex1down3 or 450ng of pX330-EV, and 50ng of pCAG-GFP. On 4 wells, the cells were co-transfected with 2.5nM, 10nM, 30nM and 100nM of 675-681-dKO-ssDNA. 48 hours after transfection, the cells were lysed and the DNA isolated using Quick-DNA Miniprep Plus Kit, following manufacturer's recommended protocol. A PCR was on 25ng of gDNA, using LongAmp® Taq DNA Polymerase with the primers 681_ex1up_Fw1 and 675_ex4down_Rv1. Amplicons were analysed on a 1% agarose gel electrophoresis.

2.11 ChIP of human and marmoset ZNF675

Cell transfection – HEK293T cells were seeded at passage 3+24 from a full 100mm dish in a ¼ dilution of cells per dish, in 8x 100mm dish. Human bone osteosarcoma cells (U2OS, Sigma-Aldrich #92022711) were seeded at passage 4+5 in 12x 100mm dishes, at a density of 2.5 x 10^6 cells per dish. Both types of cells were cultured in DMEM+GlutaMax with 10% HiFBS and 1x P/S. On the following day the medium was changed to 9ml DMEM+GlutaMax without 1x P/S. The transfection protocol for one 100mm dish is the following:

Table 2.03 – **Description of protocols followed per transfection of one 100mm dish.** The first column corresponds to the cell type transfected, the second is the transfection method used with a detailed description in the third, fourth and fifth columns.

Cell type	*Transfection method*	*Quantity of transfection agent*	*DNA mix (12500ng)*	*Incubation time*
HEK293T	Calcium Phosphate	500µl of 2x HEPES Buffered Saline (HBS)	50µl of 2.5M CaCl2 + H2O to 500µl	60s
U2OS	Polyethylenimine (PEI)	25µl in 475µl Opti-MEM (ThermoFisher Scientific #31985047)	Opti-MEM to 500 µl	20 min

The 2x HBS (1.5mM Di-sodium hydrogen phosphate (Merck Millipore #1.06580.1000), 50mM HEPES-NaOH pH 7.5, 140mM Sodium chloride) was previously prepared.

For the U2OS cells transfection with PEI (Polysciences #23966-1), the DNA mixes per dish contained 12500ng pCAG-cZNF675-GFP, 11250ng pCAG-hZNF675-GFP and 1250ng pCAG-GFP + 8750ng pCAG-EV (pBluescript was added to make the 12500ng). For the HEK293T cells transfection, the DNA mixes per dish contained 12500ng of pCAG-cZNF675-GFP and pCAG-hZNF675-GFP. For both cell types, the medium was changed to DMEM+GlutaMax with 10% HiFBS and 1x P/S, 6 hours after transfection.

Magnetic beads pre-blocking and binding to antibody – Dynabeads™ M-280 Sheep anti-Rabbit IgG (Invitrogen/ThermoFisher Scientific #11203D) were used, 50µl per immunoprecipitation (IP). The beads were transferred to a 1.5ml tube and 1ml Block Solution (0.5% Bovine Serum Albumin (Sigma-Aldrich #A2153) in 1x PBS) was added. Beads were washed 2x with 500µ Block Solution, using the magnetic stand, and each 50µl of beads was resuspended in 125µl Block Solution. 0.50µl rabbit eGFP-antibody (ab290, Abcam) was added to each 50µl of beads and incubated minimum 4 hours at 4°C on a rotating platform.

Cell's crosslinking, lysis and sonication – 48 hours after transfection the medium was removed, 5ml cold 1x PBS was added and the cells were mechanically detached. The cell suspensions from each two dishes were transferred to a 15ml tube, 1ml Crosslinking Buffer (50mM HEPES-NaOH pH 7.5 (VWR Chemicals #441476L, Merck Millipore #1064691000), 100mM Sodium chloride (Sigma-Aldrich #S7653) 1mM EDTA, 0.5mM EGTA (Sigma-Aldrich #E3889) 11% Formaldehyde (Sigma-Aldrich #F8775)) was added and the samples incubated for 10min at room temperature on a rocking platform. 500µl 2.5M glycine (Sigma-Aldrich #G8898) was added and after 5min incubation at room temperature on a rocking platform. Samples were centrifuged 5min at 300xg at 4°C and washed 2x with 10ml cold 1x PBS. The pellet of crosslinked cells was resuspended in Lysis Buffer 1 (LB1 – 50mM HEPES-NaOH pH 7.5, 140mM Sodium chloride, 1mM EDTA, 10% Glycerol (Sigma-Aldrich #15523-1L-R-D), 0.5% Nonidet™ P 40 Substitute (Sigma-Aldrich #74385), 0.25% Triton X-100 (Merck Millipore #1086031000)) with Protease Inhibitor Cocktail (Roche #5892791001), transferred to a 1.5ml tube and incubated for 10min at 4°C on a rocking platform. The samples were centrifuged 5min at 300xg at 4°C, resuspended in Lysis Buffer 2 (LB2 – 10mM Tris-HCl pH 8 (HCl – Merck Millipore #1.00317.1000), 200mM Sodium chloride, 1mM EDTA, 0.5mM EGTA) with Protease Inhibitor Cocktail, and incubated for 5min at 4°C on rocking platform. The samples were centrifuged again for 5min at 2000xg at 4°C and resuspended in 400µl Lysis Buffer 3 (LB3 – 10mM Tris-HCl pH 8, 100mM Sodium chloride, 1mM EDTA, 0.5mM EGTA, 0.1% Na-Deoxycholate (Sigma-Aldrich #8220500100), 0.5% N-laurosylsarcosine (Sigma-Aldrich #L5125-50G)) with Protease Inhibitor Cocktail. The Bioruptor™ UCD-200 (Diagenode) was used and the settings were as follows: high intensity; 22 cycles of 30s on – 60s off; as the maximum temperature the water can reach is 8°C, some water was removed, and ice added every 4 cycles; the samples were spun down every 2 cycles. After the sonication, 150µl LB3 and 50µl 10% Triton X-100 were added to the samples, they were centrifuged for 10min at maximum speed at 4°C, the supernatant was transferred to a new tube, from where 50µl was added to other tube and stored at -20°C as the input control.

Chromatin Immunoprecipitation – After at least 4 hours incubation of Dynabeads with rabbit eGFP-antibody, the mix was washed 3x with 500µl LB3, using the magnetic stand. The antibody coated beads were resuspended with 1.5ml 1% Triton X-100 in LB3 and divided over 6x 1.5ml tubes already with 250µl 1% Triton X-100 in LB3. 500µl cell lysate was added to the antibody coated beads and incubated overnight on a rotator at 4°C. On the following day, the beads were washed 8x with 500µl RIPA Buffer (50mM HEPES-NaOH pH 7.5, 500mM Lithium chloride (Merck Millipore #1056790100),

1mM EDTA, 1% NP40, 0.7% Na-Deoxycholate) with Protease Inhibitor Cocktail, using the magnetic stand, washed 1x with 1ml TBS (50mM Tris-HCl pH 7.6, 150mM Sodium chloride), centrifuged for 3min at 950xg at 4°C and any residual TBS Buffer was removed. 200µl Elution Buffer (50mM Tris-HCl pH 8, 10mM EDTA, 1% Sodium dodecyl sulphate (SDS, Sigma-Aldrich #30970)) was added to the samples, and 150µl was added to the previously thawed input controls. Samples were incubated at 65°C overnight with rotation (400rpm).

DNA isolation and purification – On the next day, the samples were centrifuged for 1min at 14000rpm, the supernatant transferred to a new 1.5ml tube and 200µl Tris-EDTA (10mM Tris-HCl pH 8, 1mM EDTA) was added to every sample (including the input controls). 20µl of each sample was stored at -20°C. 8µl 1mg/ml RNase Cocktail Enzyme Mix (ThermoFisher Scientific #AM2286) was added and incubated for 30min at 37°C. 8µl 10mg/ml Proteinase K (Roche #03115887001) was added and incubated for 2 hours at 55°C. To extract the DNA, a phenol:chloroform (Biosolve Chemicals #16962344, Rathburn Chemicals #67-66-3) method was performed twice, where 400µl phenol:chloroform was added, the tubes agitated for 30s, incubated for 5min on ice, centrifuged for 15min at 14000rpm at 4°C and the aqueous phase was transferred to a new 1.5ml tube. This method was repeated but with chloroform only. The volume was set to exactly 400µl with Tris-EDTA and 16µl 5M Sodium chloride and 1.5µl 20mg/ml Glycogen were added to the samples. 900µl 100% ethanol was also added and after incubating for 45min at -80°C, the samples were centrifuged at 14000rpm at 4°C for 45min as well. The supernatant was removed, the DNA pellet washed with 500µl 70% ethanol and centrifuged at 14000rpm at 4°C for 10min. All traces of ethanol were removed from the samples and left to air-dry briefly. The DNA was resuspended in 20µl water and purified using DNA Clean & Concentrator-5 Kit.

ChIP-qPCR – A qPCR was performed on HEK293T DNA, using QuantiTect® SYBR® Green PCR (QIAGEN), according to the manufacturer's recommended protocol. Four sets of primers were used, three as positive control and one as negative control, which are the following:

THE1_F2: CTCCCATCACAGGCCCAGAG

THE1_R1: CCTTTCAGCCATGGCTGGAG

SESN3_prom_qFw1: TCCTTCAGGGCAGTGGTTAG

SESN3_prom_qRv1: AAGTTCTTTGAGATCCCTGGA

MCPH1_prom_qFw1: GCATAACGGTGCCGAAAGTC

MCPH1_prom_qRv1: CTGAGCTGAGGATCAGGAAG

Control1_chr5_Fw: GTGGCATACCTCTCAACC

Control1_chr5_Rv: ATAGCATAGGCCAAGAGC

Western Blot – The gels for the SDS-PAGE were made: Running gel (10% acrylamide, 0.35M Tris-HCl pH 8.8, 0.1% APS, 0.1% SDS, 0.00001% TEMED) and Stacking gel (5% acrylamide, 0.125M Tris-HCl pH 6.8, 0.1% APS, 0.1% SDS, 0.00001% TEMED). Samples were prepared in 5% dithiothreitol (Sigma-Aldrich #D0632) in 2x Laemmli Sample Buffer (Bio-Rad #1610737). 20µl per sample (from U2OS and HEK293T cells) was loaded on the gel and migrated in Running Buffer (1x Tris-Glycine (Bio-Rad #161-0771), 0.1% SDS) at 100V for 25min, then at 160V for 1h. Blotting on a nitrocellulose membrane was performed in Transfer Buffer (1x Tris-Glycine, 20% Methanol (Rathburn Chemicals #67-

56-1)) at 100V for 2 hours. The membrane was stained with Ponceau S (Sigma-Aldrich #P3504), then washed with 1x TBS-T (0.15M Sodium chloride, 0.02M Tris-HCl pH 7.4, 0.001% Tween-20 (Sigma-Aldrich #P9416)). The membrane was blocked in 5% Milk powder in 1x TBS-T for 1hour on orbital shaker and washed once with 1x TBS-T. Primary antibody rabbit anti-GFP (Abcam ab290) (1:2500) was incubated in 1x TBS-T overnight at 4°C on rotator. The membrane was rinsed once and washed 3x with 1x TBS-T for 15min each on orbital shaker. Secondary antibody goat anti-rabbit-HRP (ThermoFisher Scientific #65-6120) (1:25000) was incubated in 1x TBS-T, for 1hour on orbital shaker. Membranes were incubated with Supersignal westdura ECL substrate (ThermoFisher Scientific #34075), according to the manufacturer's recommended protocol, and imaged using ChemiDoc imager.

3 Results

3.1 Divergent evolution of ZNF675 and ZNF681

To investigate the evolutionary history of the ZNF675 and ZNF681, a predictive DNA recognition tool was used. The analysed ZNF675 protein sequences are from human, crab-eating macaque and marmoset, and the ZNF681 sequence from human and crab-eating macaque. Since the marmoset ZNF675 is the oldest, it is certainly the most similar to the first ZNF675. It is composed of 4 initial degenerate and 19 functional ZNF motifs (Figure 3.01), making it the largest. Regarding the differences between ZNF675 and ZNF681 of crab-eating macaque and human, the ZNF675 of the OWM's ancestor was recreated. This ZNF675 got a nonsense mutation, preventing the last 6 ZNF motifs from being translated. Then a duplication event happened, giving rise to ZNF681. Finally, a deletion of 2 ZNF motifs occurred in the ZNF675, creating a protein with only 11 functional ZNF motifs. This novel ZNF675 is significantly different from the marmoset ZNF675, since the DNA sequence it recognizes is also different and shorter (considering that every ZNF motif binds to DNA). ZNF675 is present from the crab-eating macaque onwards, with a high sequence conservation between species. Even though ZNF681 remained with the same number of ZNF motifs, 2 of which degenerated, the DNA motifs they recognize changed considerably among the two species.

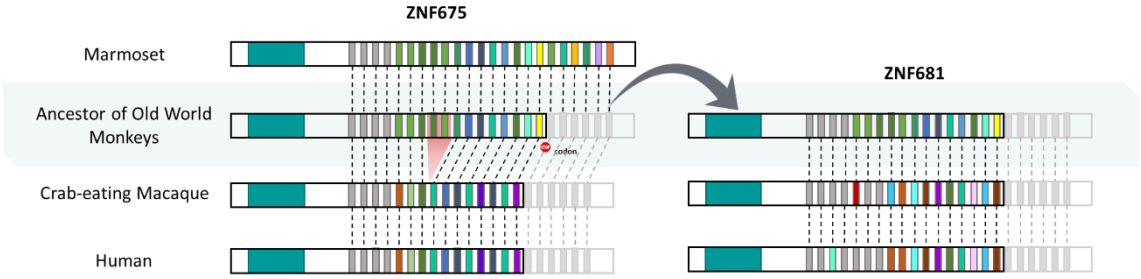
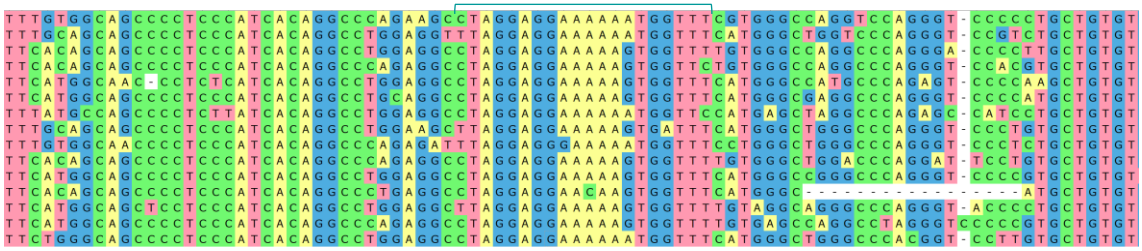


Figure 3.01 – **Structural changes of ZNF675 and ZNF681 throughout primate evolution.** ZNF motifs characterized as degenerate are represented in grey, whereas functional are in different colours, meaning different DNA motifs which they recognize. The shady regions are untranslated ZNF motifs. The KRAB domain is represented as the blue box.

3.2 Modifications to ZNF675 binding site in THE1B-int elements

To investigate the ZNF675 binding difference across the THE1B-int elements, 15 sequences were chosen, based on high sequence similarity score, and aligned with Multiple Sequence Alignment MUSCLE on UGENE. When comparing both THE1B-int bound and unbound elements, there is a clear decrease in sequence conservation in the centre region (Figure 3.02), which coincides with the ZNF675 binding site. Within the binding site, specifically in the stretch of adenines, there is frequently an indel or a substitution. Nevertheless, the same degree of conservation is maintained in the surrounding regions.

a.



b.

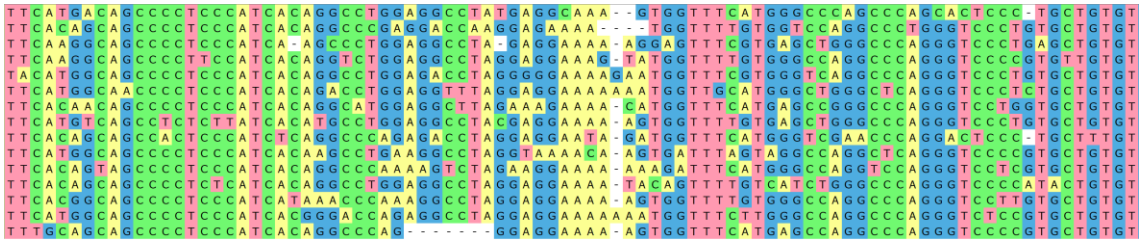


Figure 3.02 – **THE1B-int elements.** (a) THE1B-int elements bound by ZNF675, with its binding site marked in the centre region. (b) THE1B-int elements not bound by ZNF675. Every element represented here has a length of ~1600bp and a similarity score higher than 12400.

3.3 ZNF675 binds to promoters of gene related to development and neuronal processes

Through the analysis of combined ChIP-seq data (GSM2466628, GSM1407624), a list of genes whose promoters are bound by ZNF675 was created. Out of the 351 genes, 25% are involved in development and neuronal processes, for instance involved in cell signalling, metabolism or associated with neuronal diseases. In 54% of the genes, the ZNF675 binding site coincides with the binding of transcription factors and RNA Polymerase II. Moreover, in 48% of the cases the ZNF675 binds to a retrotransposon nearby the gene, and, in 7% ZNF675 binds to the transcribed region of a long non-coding RNA.

Representation of the selected genes, with the correspondent MACS peak from both ChIP-seq data sets and the gene function, is present in table 3.01. Even though most of the genes are present in both data sets, some are only present in one. The MACS peak is different when comparing both data sets, since their method is also different.

Table 3.01 – **Gene promoters bound by ZNF675.** First column corresponds to genes whose promoters are bound by ZNF675, the second is the chromosome where it locates. Third and fourth columns are the peak signal from the ChIP-seq data GSM2466628 and GSM1407624, respectively. The fifth column corresponds to the gene function.

Gene	Chromosome	Peak signal from GSM2466628	Peak Signal from GSM1407624	Function
OAT	10	211	51	Key enzyme glutamate/GABA synthesis
ZNF778	16	194	41	Autism and variable cognitive impairment

TCF25	16	174	–	Transcription factor important in embryonic development
WNT5A	3	168	–	Wnt signalling
CRY1	12	154	83	Circadian clock gene regulation
CHST10	2	147	–	Indirectly involved in synapse plasticity of the hippocampus
HES1	3	126	20	NOTCH signalling
TXLNA	1	117	47	Exocytosis of neuroendocrine cells
MEGF10	5	108	13	Schizophrenia, myopathy
TEX28	X	107	50	Encephalopathy and mental retardation
TCTN3	10	100	73	Cilium biogenesis, Hedgehog signalling
CHMP1A	16	86	12	Vesicle transport, brain development
MCPH1	8	43	11	Regulation brain size
SESN3	11	22	23	Master regulator of pro-epileptic genes
NEUROG3	10	–	21	Neurogenesis

3.4 HES1 expression is up-regulated after induced neuronal differentiation

HES1 expression levels are known to change upon RA treatment in neuroblastoma cells, which induces differentiation (Jögi, Persson, Grynfeld, Påhlman, & Axelson, 2002). SK-N-SH cells (Figure 3.03a) were treated with 10µM RA, the RNA was extracted, and a one-step RT-PCR was performed. The amplicons were analysed on gel (Figure 3.03b), relative density of which was assessed. Comparing to the expression of HES1 on T0, which is set to 1, the fold-change in expression levels is greater after 24 hours of treatment (6-fold) than after 2 hours (3-fold) (Figure 3.03c).

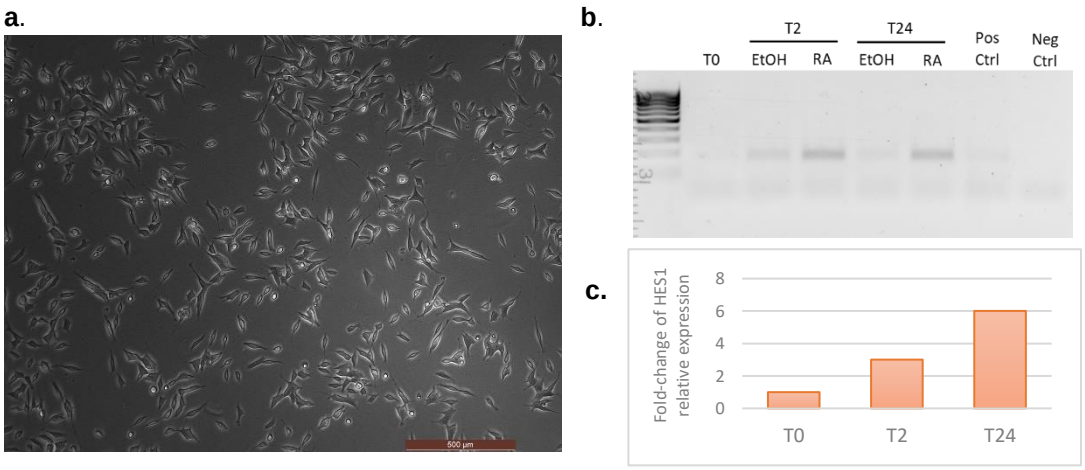


Figure 3.03 – **HES1 expression levels change after Retinoic Acid treatment.** (a) Heterogeneous SK-N-SH cells, with different subtypes present. Magnification of 10x. (b) 1% agarose gel electrophoresis of HES1 amplicons from a one-step RT-PCR, after cells treatment with RA and ethanol, as control. DNA ladder used is GeneRuler™ Low Range (#SM1193, Thermo Scientific). (c) Fold-change of HES1 relative expression after RA treatment, analysed by relative density of HES1 amplicon on 1% agarose gel electrophoresis. The graph is the result of a single experiment.

3.5 Significant changes in the binding site of ZNF675 in MCPH1 and SESN3 promoters in primates

To assess the ZNF675 binding site location on MCPH1, ChIP-seq data of ZNF675 (GSM1407624, GSM2466628) was analysed (Figure 3.04, *top*). In the MCPH1 gene, ZNF675 binds in the promoter region, 360bp from the TSS, within the long non-coding RNA LOC100287015. Sequence from the centre region of the ChIP peak, containing the predicted ZNF675 binding site, was analysed in several primate species, selected from the UCSC Genome Brower. Sequence conservation was assessed using the Multiple Sequence Alignment MUSCLE on UGENE. Clear modifications occurred in the binding site of ZNF675, specifically in the LCA of apes and humans (Figure 3.04, *bottom*). Both groups of primates have a 3-4 nucleotides deletion in the centre of the ZNF675 binding site, comparing to the NWMs. Regarding the OWM, a 1-2 nucleotides insertion and substitutions occurred in the same region. Though, considering the sequence similarity of apes and NWMs, mutations probably took place in the ancestor of OWM, being a family-specific change. There are some species-specific mutations across all primates, which are mostly substitutions.

Figure 3.04 – **MCPH1 promoter region.** Visualization of ZNF675 ChIP signal on the UCSC Genome Browser, localized within a long non-coding RNA LOC100287015 (*top*). Zoomed in region of MCPH1 promoter sequences from NWM until Humans, with 67bp in length (*bottom*).

The same analysis previously described for MCPH1 was reconducted for SESN3. The ZNF675 binding site location on SESN3 was analysed (Figure 3.05, *top*). In the SESN3 gene, the binding site of ZNF675 is in the promoter region of the standard gene, 590bp away from the TSS. Although if the gene starts with the alternative TSS, the binding site is localized within the first intron, 870bp away from the start of the first exon. In this case, the first exon of the standard gene is not transcribed. Sequence conservation was assessed using the Multiple Sequence Alignment MUSCLE on UGENE. When comparing the sequences, a 2 nucleotides insertion occurred in the LCA of OWM and apes, in the

ZNF675 binding site (Figure 3.05, *bottom*). From thereafter the sequence remained highly conserved, although there are species-specific substitutions, mostly outside of the binding site. A more drastic change happened in the ancestral version of the NWM, where a 4 nucleotides deletion occurred, being probably a family-specific mutation. There are several substitutions across all species analysed, the majority present in the basal primates' promoter region.

Figure 3.05 – **SESN3 promoter region.** Visualization of ZNF675 ChIP signal on the UCSC Genome Browser, localized in an intron of SESN3 gene with an alternative TSS (*top*). Zoomed in region of SESN3 promoter sequences from basal primates until Humans, with 67bp in length (*bottom*).

3.6 Differential effect of ZNF675 binding on MCPH1 and SESN3 promoters throughout evolution

To investigate the binding affinity of both ZNF675 from human and marmoset on the MCPH1 and SESN3 promoter regions, Luciferase Reporter Assays were performed. MSTA-Luciferase constructs were used as positive control in both assays (Figure 3.06, *top*). The hZNF675 repressed the reporter expression by 89%, although the cZNF675 did not have any effect (Figure 3.06a, *bottom*). On the human MCPH1 construct, though hZNF675 did not repress the reporter expression, the cZNF675 repressed it by 13%. In the marmoset MCPH1 construct, both human and marmoset ZNF675 influenced the reporter expression, repressing it by 11% and 8%, respectively. When comparing the repressive effect of hZNF675 on both human and marmoset MCPH1 constructs, there is a change in binding affinity, resulting in 13% less expression in the marmoset MCPH1. However, the repressive effect of cZNF675 on the human MCPH1 in comparison to the marmoset MCPH1, only differed at 6%.

On the human SESN3 construct (Figure 3.06b, *bottom*) the hZNF675 had a strong effect, repressing it by 48%. On the contrary, cZNF675 had no influence on the reporter expression. On the marmoset SESN3, neither of human and marmoset ZNF675 had an effect.

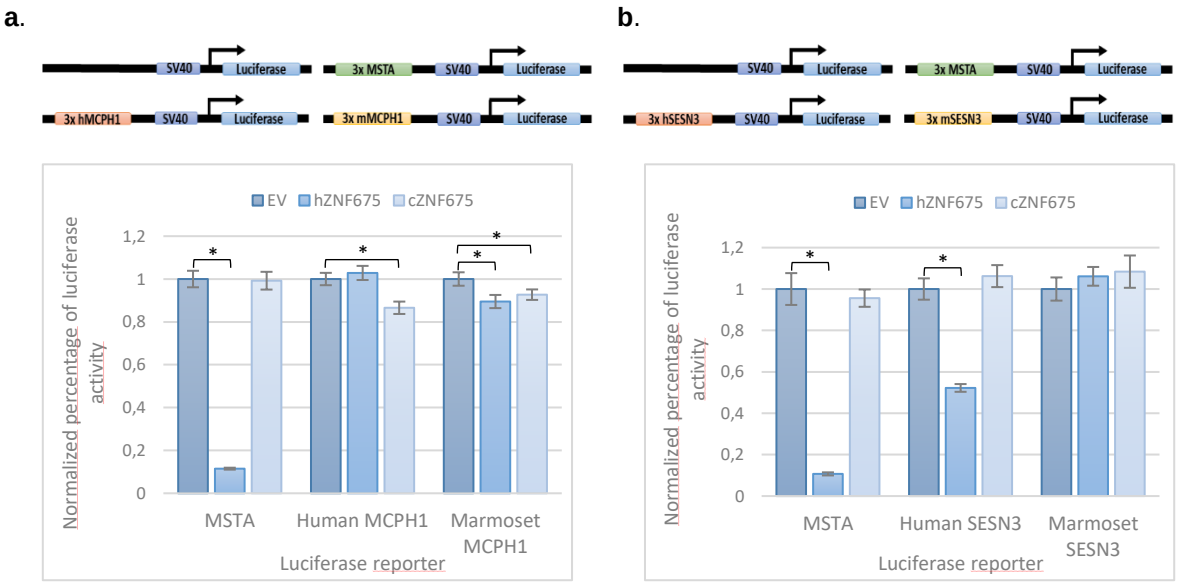


Figure 3.06 – **Luciferase Reporter Assays on MCPH1 and SESN3, in the presence of hZNF675 and cZNF675.** (a-b) Schematic of SV40-Luciferase constructs: EV, (3x) 100bp of MSTA, (3x) 200bp of hMCPH1, cMCPH1, hSESN3 and cSESN3 (*top*) and correspondent Luciferase Reporter Activity after transfection of the constructs in mESC (*bottom*). Each graph is the result of three independent experiments with n=6. Empty vector is set to 1 for comparison. Statistical analysis: Two-way repeated measures ANOVA; *P<0.001; error bars are s.e.m.

3.7 Isolation of MCPH1 promoter region from different primate species

To enable the analysis of the ZNF675 binding affinity to the MCPH1 promoter from different primates, a region of around 1200bp was amplified from gDNA of 7 species: human, chimpanzee, gorilla, orangutan, gibbon, rhesus macaque and baboon (Figure 3.07a). To be able to clone these sequences in a vector, the DNA was extracted and amplified with primers with restriction sites (Figure 3.07b).

The marmoset region, though, could not be amplified. A gradient PCR was performed with annealing temperatures ranging from 58°C to 63.9°C and the amplicons were analysed in a 1% agarose gel electrophoresis (Figure 3.07c). The two amplicons present in every amplification were extracted and Sanger sequenced. The result of which identified the DNA as being bacterial. A new approach was taken, and new primers designed: F4, R6 and R7. The F4 and R7 are the reverse and complement of each other, amplifying two sequences that overlap in 22bp. The primer pair F3/R7 would amplify a region of 1010bp, whilst the primer pairs F4/R6 and F4/R5 would amplify regions of 306bp and 126bp, respectively. In a second PCR, the overlapping region from the products of the amplification with the F3/R7 primer pair and F4/R5 or F4/R6 would hybridize and function as a primer for the amplification of the whole region. However, the amplification of the two sequences was unsuccessful, since there are several unspecific PCR products (Figure 3.07d).

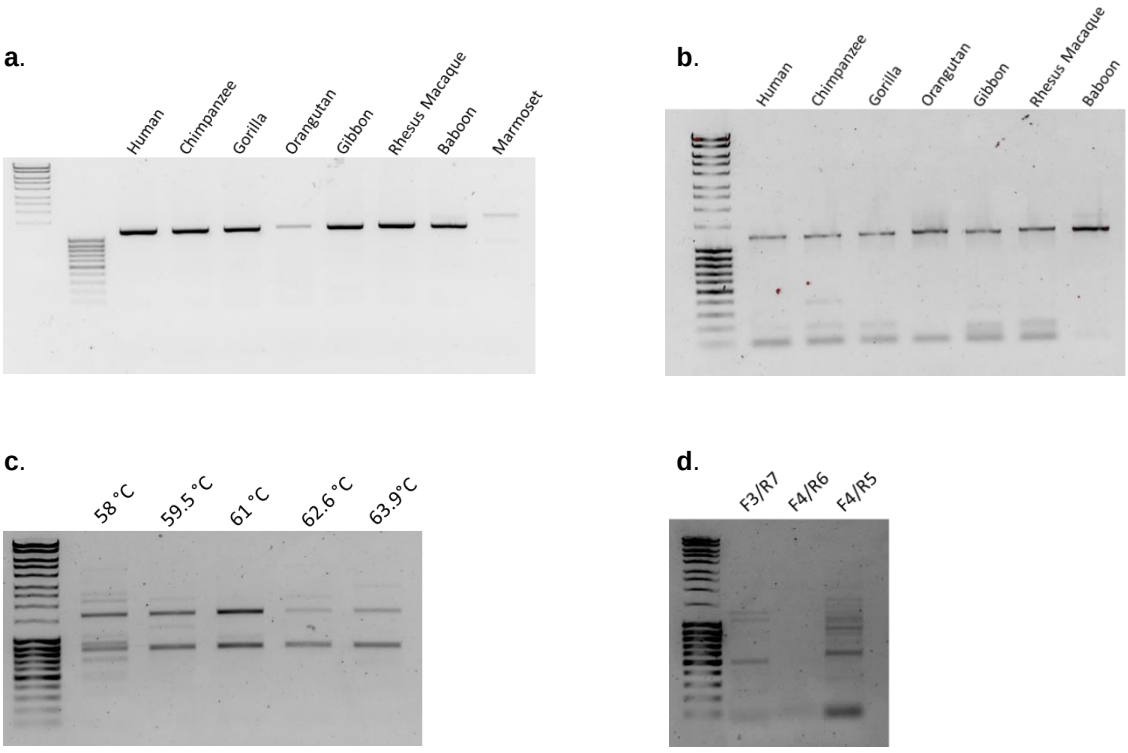


Figure 3.07 – **MCPH1 promoter region from several primate species analysed by 1% agarose gel electrophoresis.** (a) MCPH1 promoter amplicons. (b) MCPH1 promoter amplicons with restriction sites on both ends. (c) Unspecific products of amplification of marmoset MCPH1 promoter with F3/R5 primers, differing on the PCR annealing temperature (indicated in the image). (d) Unspecific products of amplification of marmoset MCPH1 promoter with different primer pairs: F3/R7, F4/R6 and F4/R5. In every 1% agarose gel electrophoresis the DNA ladders used were GeneRuler™ High Range (#SM1353, Thermo Scientific) and GeneRuler™ Low Range.

3.8 ZNF675 binding site of the marmoset MCPH1 promoter in a human context

The CRISPR/Cas9 system was used as a genome editing tool, by inducing the replacement of the ZNF675 binding site on the human MCPH1 promoter by the marmoset, through HDR. The gRNAs that target the ZNF675 binding site were designed (Figure 3.08a), cloned and transfected in HEK293T cells (Figure 3.08b). The transfection was verified by GFP-positive cells, visible in the cytoplasm. If the gRNA introduced a double stranded break on the DNA, it can form heteroduplexes after being randomly annealed. These DNA fragments will migrate slowly in the gel due to their structure. According to this, by the analysis of the PAGE (Figure 3.08c), both gRNAs are working correctly, though MCPH1_bs_down is the most promising.

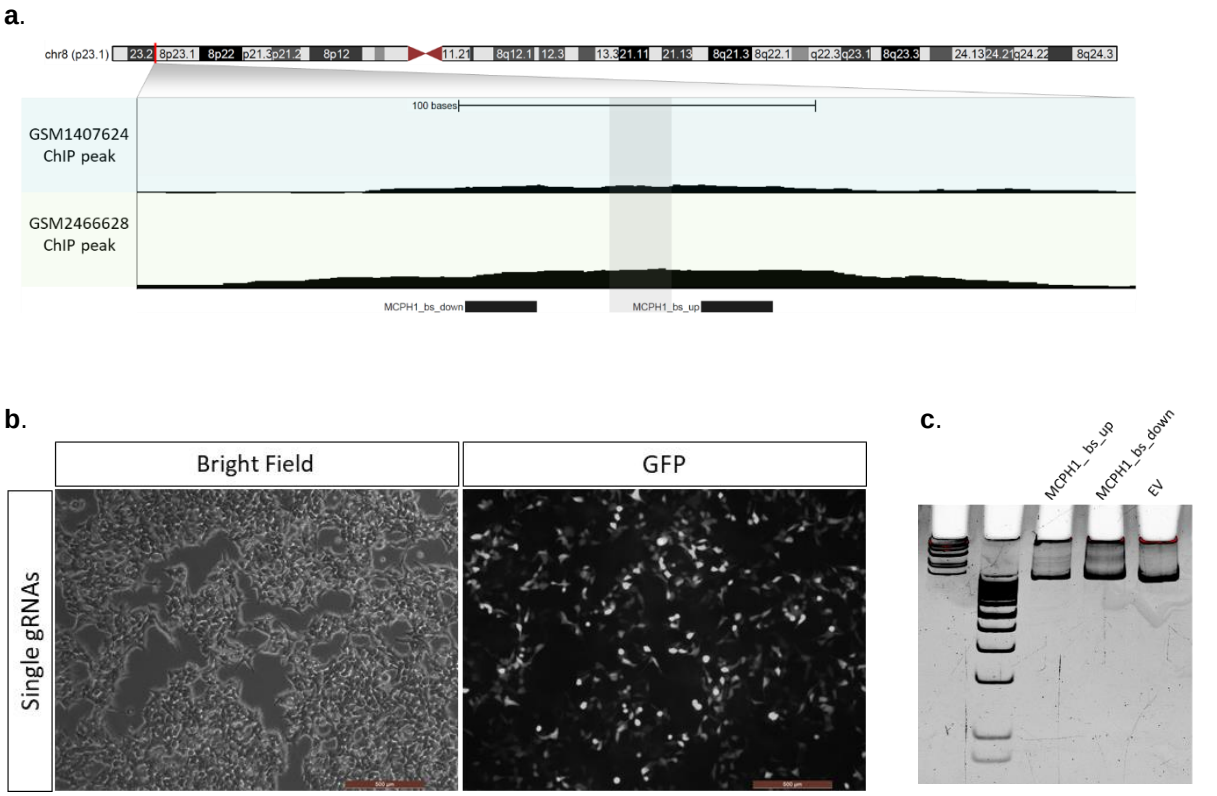
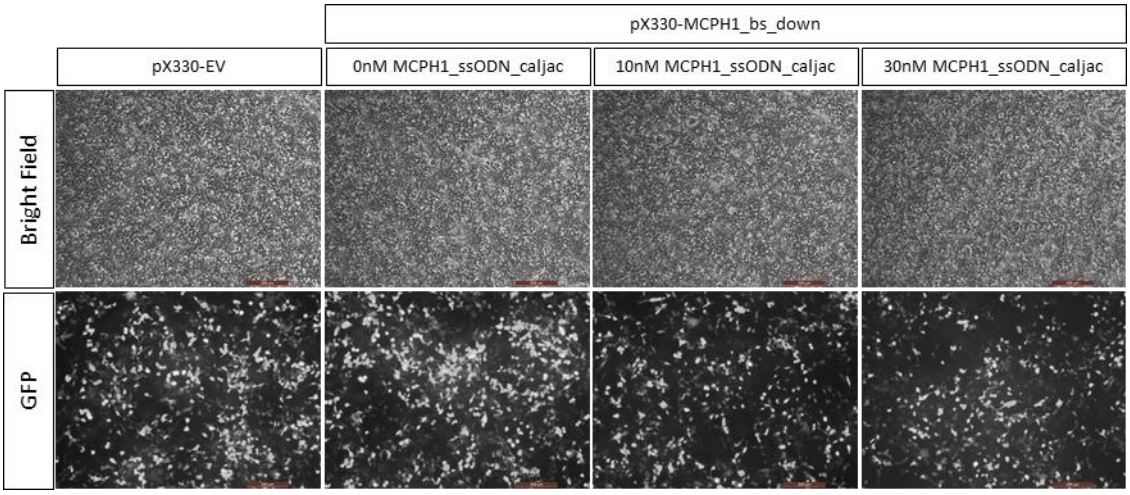


Figure 3.08 – **Testing of gRNAs that target the ZNF675 binding site on the MCPH1 promoter.** (a) Visualization of the gRNAs pX330-MCPH1_bs_up and pX330-MCPH1_bs_down on the UCSC Genome Browser. Grey bar is the predicted ZNF675 binding site. (b) Example of transfected HEK293T cells with gRNAs. Magnification of 10x. (c) 8% PAGE of MCPH1 promoter region after cells' transfection of pX330-MCPH1_bs_up, pX330-MCPH1_bs_down and pX330-EV. The DNA ladders used were GeneRuler™ High and Low Range.

The ssDNA oligo used, MCPH1_ssODN_caljac, comprises of two 30bp homology-arms which are complementary to the promoter sequence of human, whilst from the introduced double-stranded break until the end of the ZNF675 binding site the sequence is from the marmoset (Figure 3.09a). The ssDNA oligo was co-transfected with the gRNA MCPH1_bs_down in HEK293T cells (Figure 3.09b), which was verified with GFP-positive cells. Cells transfected with the EV and the gRNA have a higher number of GFP-positive cells, whereas in the co-transfection of 10nM and 30nM of ssDNA oligo, it appears to be slightly lower. To analyse if the genome editing was successful, the replaced region was amplified, with one primer within the marmoset region and another in the human MCPH1 promoter. The amplicon was analysed on gel (Figure 3.09c). Even though there are unspecific products from the PCR, there is one amplicon, with the expected size of 260bp, that is only present in the cells co-transfected with the gRNA and ssDNA oligo.

a.

b.



c.

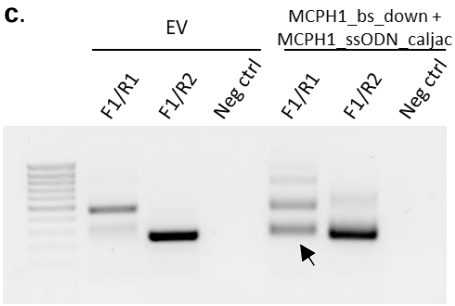


Figure 3.09 – **Replacement of ZNF675 binding site on the MCPH1 promoter by the marmoset.** (a) Visualization of the gRNA pX330-MCPH1_bs_down and the oligo MCPH1_ssODN_caljac on UCSC Genome Browser. Arrow indicates the induced double-stranded break. (b) HEK293T cells transfected with pX330-EV, pX330-MCPH1_bs_down and co-transfected with the gRNA and MCPH1_ssODN_caljac with different concentrations: 10nM and 30nM. Magnification of 10x. (c) 1% agarose gel electrophoresis of replaced region from HEK293T cells transfected with pX330-EV and co-transfected with pX330-MCPH1_bs_down and 30nM of MCPH1_ssODN_caljac. PCR was done with the primers F1/R1 and F1/R2. Arrow points to the band of 260bp. The DNA ladder used was GeneRuler™ Low Range.

3.9 ZNF675/ZNF681 KO using CRISPR/Cas9

In order to get a ZNF675/ZNF681$^{-/-}$ cell line, a working CRISPR/Cas9 system was created. gRNAs that target exon 1 of ZNF681 were designed (Figure 3.10a), cloned and tested in HEK293T cells (Figure 3.10b). The transfection was verified with GFP-positive cells. By the analysis of the PAGE (Figure 3.10c), all gRNAs work correctly, except for 681_ex1up2. Finally, HEK293T cells were transfected with the gRNA 681_ex1up1 paired with 681_ex1down1, 681_ex1down2, 681_ex1down3

and 681_ex1down4. gDNA was amplified and analysed on gel (Figure 3.10d). Every gRNA combination worked, inducing a deletion of ZNF681, which can be verified by the lowest bands present on the gel.

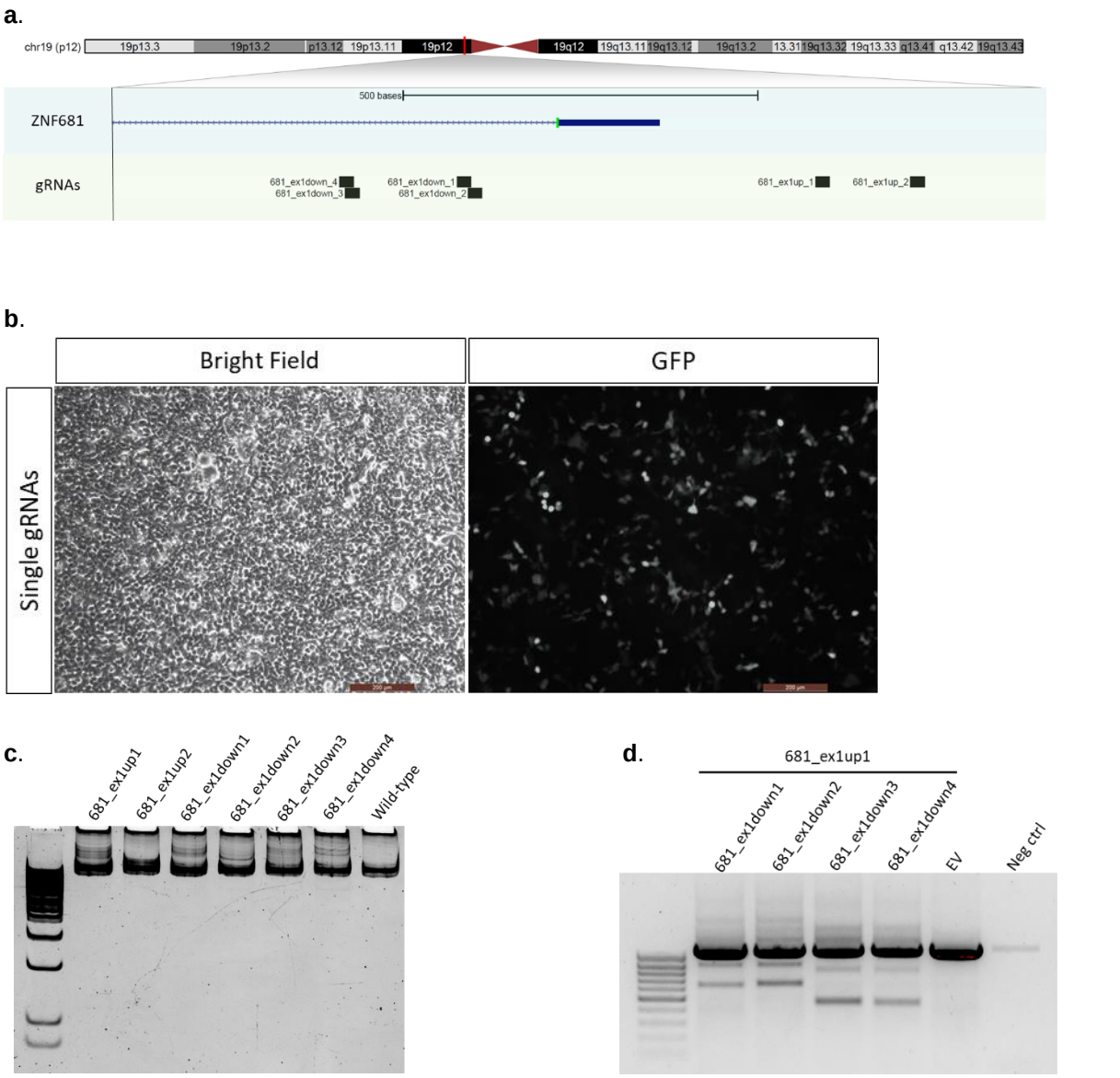


Figure 3.10 – **Testing of gRNAs that target exon 1 of ZNF681**. (a) gRNAs visualized on UCSC Genome Browser. (b) Example of transfected HEK293T cells with a gRNA with 10x of magnification. (c) 8% PAGE of ZNF681 region after HEK293T transfection with the gRNAs: pX330-681_ex1up1, pX330-681_ex1up2, pX330-681_ex1down1, pX330-681_ex1down2, pX330-681_ex1down3, pX330-681_ex1down4 and pX330-EV. (d) 1% agarose gel electrophoresis of ZNF681 region after transfection of the gRNA pX330-681_ex1up1 paired with pX330-681_ex1down1, pX330-681_ex1down2, pX330-681_ex1down3 and pX330-681_ex1down4, and transfection of pX330-EV. The DNA ladder used was GeneRuler™ Low Range.

To induce a deletion of both ZNF675 and ZNF681 genes, the gRNAs 681_ex1up1, 681_ex1down3 and 681_ex1down4 were paired with the previously tested gRNAs for ZNF675, 675_ex4down1 and 675_ex4up2 (Figure 3.11a). HEK293T were transfected with combinations of the gRNAs (Figure 3.11b), which was verified with GFP-positive cells. To assess the deletion induced by the CRISPR/Cas9 system, the ZNF675-ZNF681 region was amplified and analysed on gel (Figure

3.11c). All the ZNF681 gRNAs paired with the gRNA 675_ex4down1 were successful. The pair 675_ex4down1/681_ex1down3 was the most efficient, which can be verified by the intensity of the band in the fourth lane. The 675_ex4up2/681_ex1up1 pair did not work.

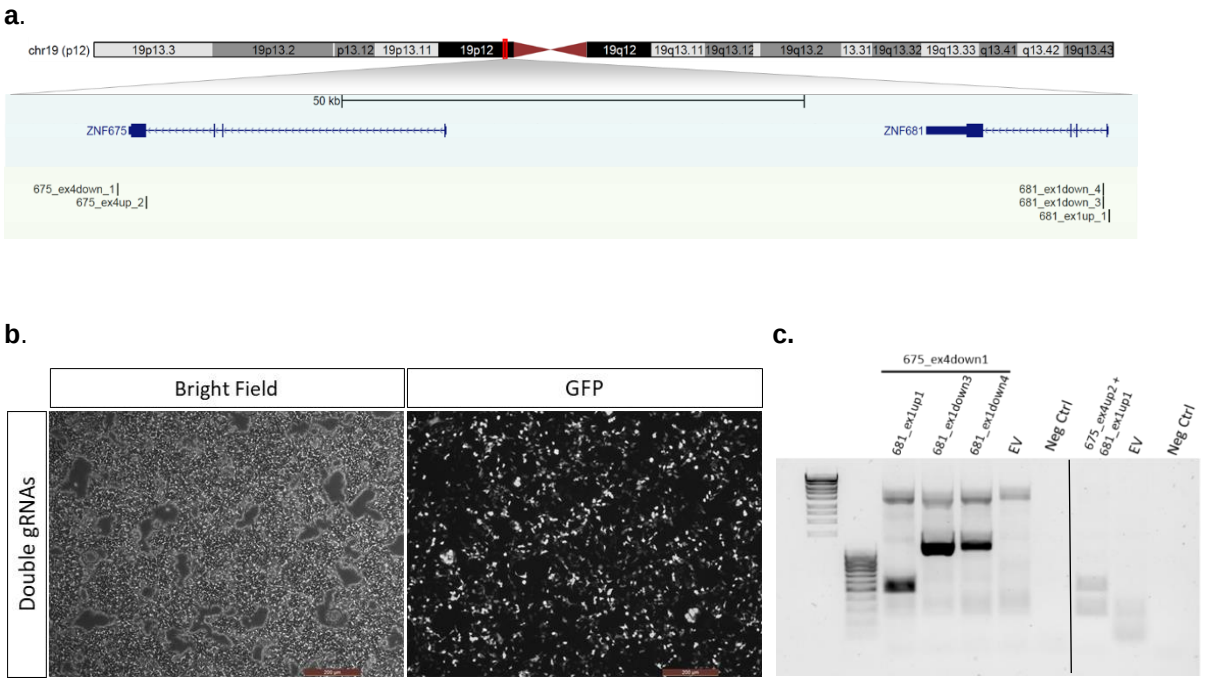


Figure 3.11 – **ZNF675/ZNF681 KO using CRISPR/Cas9 system.** (a) Visualization of gRNAs for ZNF675 and ZNF681 on UCSC Genome Browser. (b) Example of transfected HEK293T cells with a gRNA pair. Magnification of 10x. (c) 1% agarose gel electrophoresis of ZNF675-ZNF681 region after HEK293T transfection of pX330-675_ex4down1/pX330-681_ex1up1, pX330-675_ex4down1/pX330-681_ex1down3, pX330-675_ex4down1/pX330-681_ex1down4, pX330-675_ex4up2/pX330-681_ex1up1 and pX330-EV. The DNA ladders used were GeneRuler™ High and Low Range.

To improve the efficiency of the deletion, an approach using HDR was adopted. The ssDNA oligo used, 675-681-dKO-ssDNA, comprises of two 30bp homology-arms which are complementary to the sequence that remains after deletion, on both ends (Figure 3.12a). The ssDNA oligo was co-transfected with the gRNA pair 675_ex4down1/pX330-681_ex1down3 in HEK293T cells (Figure 3.12b). The transfection was verified with GFP-positive cells, which appears to be in higher number in the conditions with EV, gRNA pair and in the co-transfection of the gRNA pair and 2.5nM and 10nM of ssDNA oligo. On the contrary, in the conditions with 30nM and 100nM of ssDNA oligo, the number of GFP-positive cells is lower. To assess the deletion efficiency, the ZNF675-ZNF681 region was amplified and analysed on gel (Figure 3.12c). The efficiency of the deletion was greater in the transfection with the gRNA pair only, however, the co-transfection of gRNA pair and 10nM and 30nM of ssDNA oligo appears to be reasonable. Conditions with 2.5nM and 100nM of ssDNA oligo had the lowest efficiency. However, the deletion efficiency does not seem to improve with the addition of ssDNA oligo.

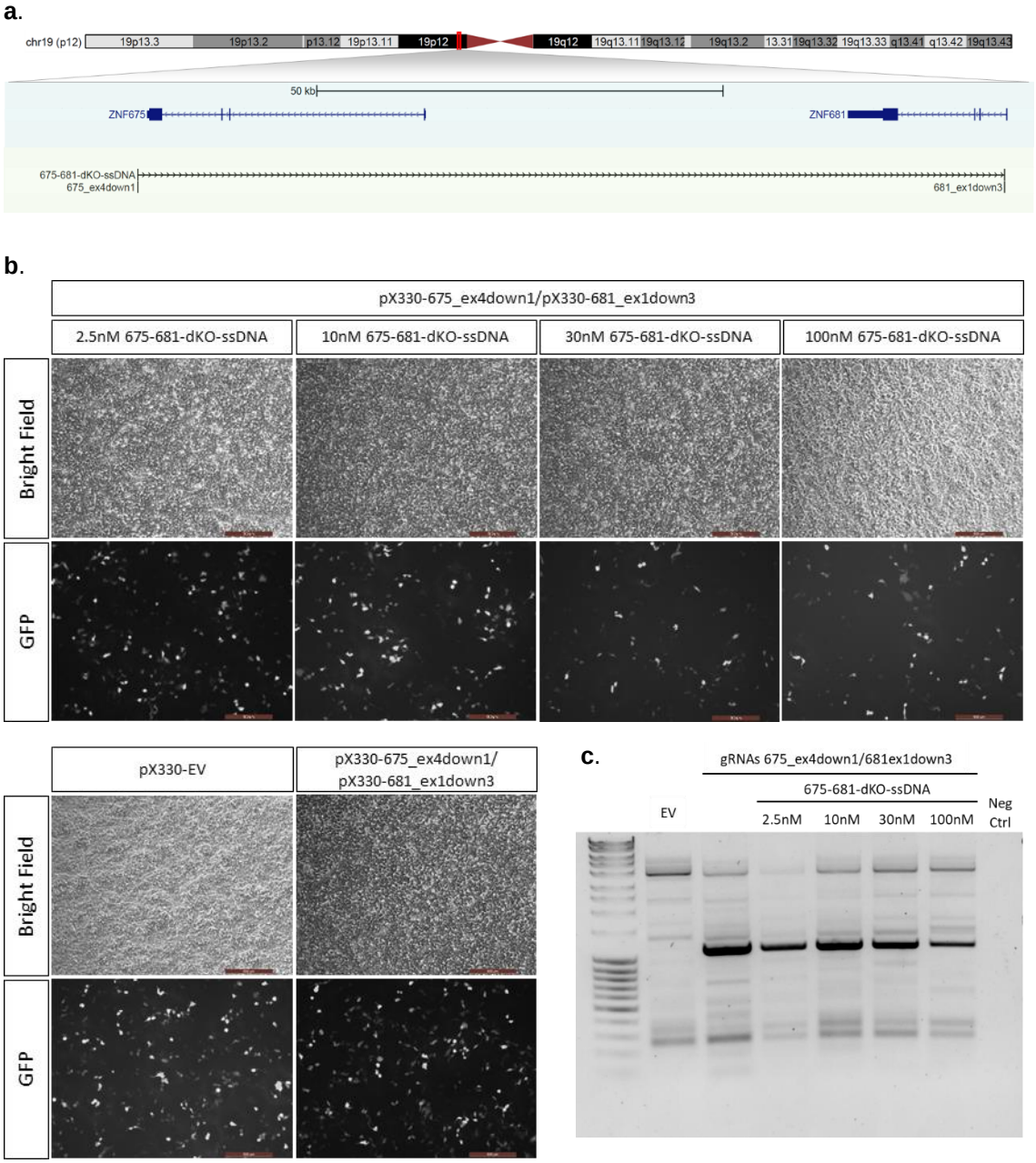


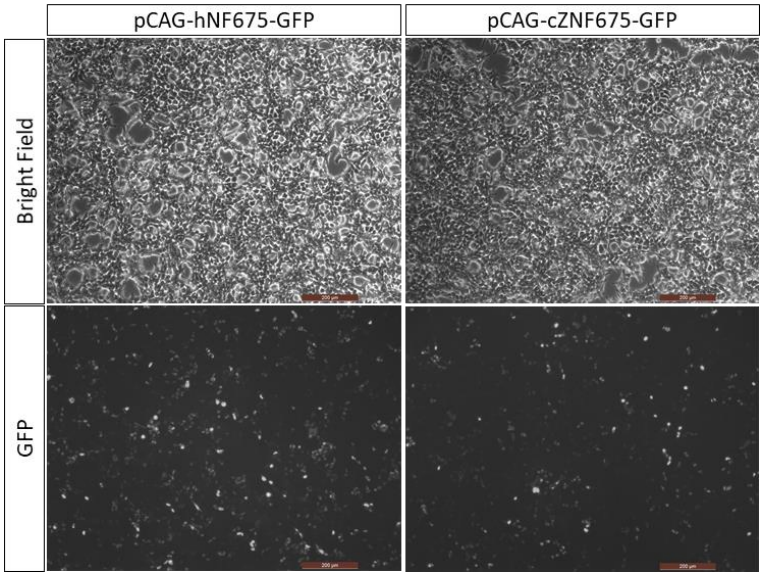
Figure 3.12 – **HDR-mediated ZNF675/ZNF681 KO using CRISPR/Cas9 system.** (a) Visualization of the gRNA pair pX330-675_ex4down1/pX330-681_ex1down3 and the oligo 675-681-dKO-ssDNA on UCSC Genome Browser. (b) HEK293T cells transfected with the pX330-EV and 675_ex4down1/pX330-681_ex1down3, and co-transfected with the gRNA pair and the 675-681-dKO-ssDNA with different concentrations: 2.5nM, 10nM, 30nM and 100nM. Magnification of 10x. (c) 1% agarose gel electrophoresis of ZNF675-ZNF681 region from the transfected cells. The DNA ladders used were GeneRuler™ High and Low Range.

3.10 ChIP of human and marmoset ZNF675

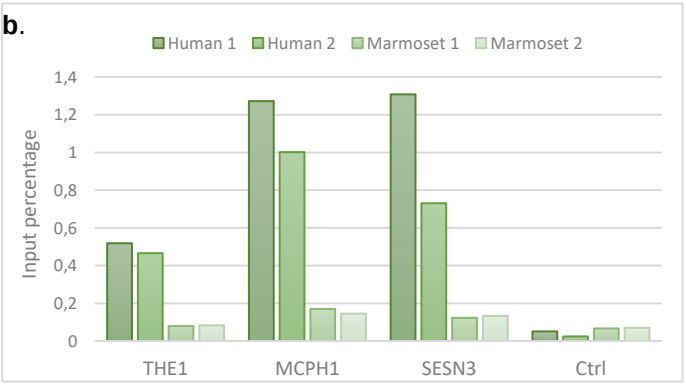
The binding site of human ZNF675 has already been described, by analysing ChIP-seq data (GSM1407624, GSM2466628). However, there is no information regarding the marmoset ZNF675. A ChIP was performed, using HEK293T cells transfected with the previously cloned pCAG-hZNF675-GFP and pCAG-cZNF675-GFP (Figure 3.13a), which was verified by GFP-positive cells. ChIP samples were

analysed by qPCR (Figure 3.13b). There is an enrichment of THE1 elements, MCPH1 and SESN3 promoter regions in the IP for hZNF675, comparing to the control. On the contrary, the IP for cZNF675 showed no substantial enrichment for either of the analysed regions. These results are supported by the western blot analysis (Figure 3.13c), since the cZNF675, with a size of 122.96 kDa, is absent in the IP samples. The hZNF675, with a size of 94.08 kDa, is present in the IP samples, though with a very faint signal. Moreover, there was degradation of both ZNF675 proteins, since there is intense signal for GFP, with a size of 25kDa. The signal at 55kDa in the IP samples is most likely the eGFP-antibody used for the ChIP.

a.



b.



c.

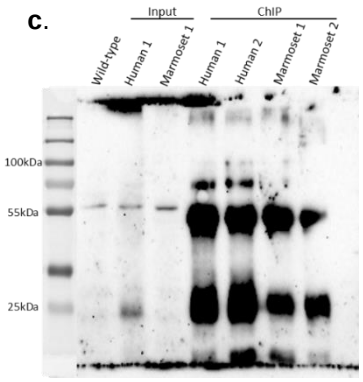
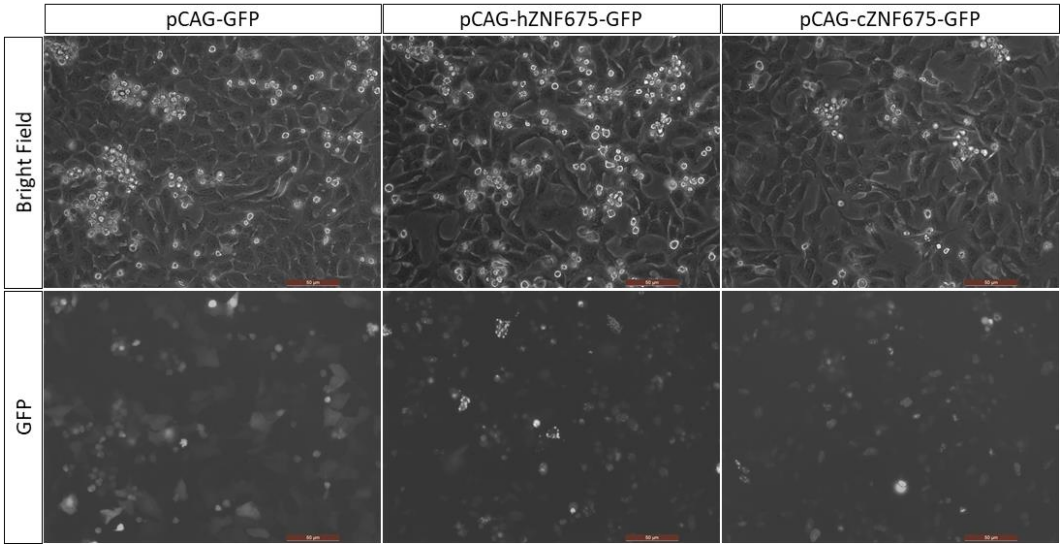


Figure 3.13 – **ChIP of human and marmoset ZNF675 on HEK293T.** (a) HEK293T cells transfected with pCAG-hZNF675-GFP and pCAG-cZNF675-GFP. Magnification of 10x. (b) qPCR with primers for THE1 elements, MCPH1 and SESN3 promoter regions, and a non-coding region on Chromosome 5 as negative control. (c) Immunoblot of IP DNA from cells transfected with pCAG-hZNF675-GFP and pCAG-cZNF675-GFP (2 samples each), correspondent input (1 sample each) and untransfected cells.

In order to improve the ChIP, it was performed following the same protocol in U2OS cells transfected with pCAG-GFP, pCAG-hZNF675-GFP and pCAG-cZNF675-GFP (Figure 3.14a). The transfection was verified by GFP-positive cells. ChIP samples were analysed by western blot (Figure 3.14b), which shows the same results as previously, though without protein degradation. While low levels of hZNF675 are detected in the IP, the cZNF675 remains undetectable. Therefore, these samples are not suitable to assess genome-wide DNA binding of cZNF675.

a.



b.

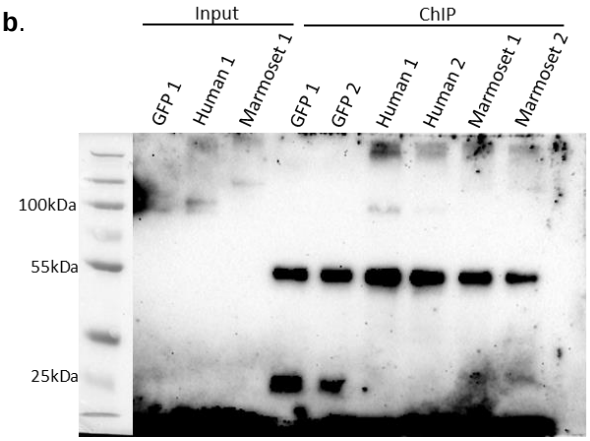


Figure 3.14 – **ChIP of human and marmoset ZNF675 on U2OS cells.** (a) U2OS cells transfected with pCAG-GFP, pCAG-hZNF675-GFP and pCAG-cZNF675-GFP. Magnification of 10x. (b) Immunoblot of IP DNA from cells transfected with pCAG-GFP, pCAG-hZNF675-GFP and pCAG-cZNF675-GFP (2 samples each), correspondent input (1 sample each).

4 Discussion

The arms race between transposable elements and KZNF proteins have been shaping the genome throughout primate evolution, affecting mostly regulatory networks (Imbeault et al., 2017; Jacobs et al., 2014; Najafabadi et al., 2015). In fact, both TEs and KZNF proteins have been described to be involved in important processes, including development. The study of developmental mechanisms can shed light on evolutionary modifications in primates. Interestingly, within the transcription factors that are differentially expressed in the human brain, comparing to that of the chimpanzee, KZNF proteins are over-represented and have been associated, at least in part, to its size and complexity (Nowick et al., 2009).

ZNF675 is expressed in the brain, being associated with multiple neurological disorders. It arose in the LCA of NWM and OWM. It got a nonsense mutation followed by a duplication event in the LCA of OWM and apes, giving rise to ZNF681. Both new KZNF proteins are significantly different from the marmoset ZNF675, recognizing new DNA sequences due to modifications in several ZNF motifs (Figure 3.01). However, ZNF675 protein structure remained conserved from the OWM onwards, implying that crab-eating macaque and human ZNF675 have the same DNA binding site. On the contrary, ZNF681 underwent some changes in the ZNF domain. When comparing the crab-eating macaque and the human ZNF681, there are two ZNF motifs that recognize different DNA sequences. These two proteins have acquired new DNA binding sites, which might imply new regulatory properties.

ZNF675 represses THE1 and MST elements, by binding to the internal part of the retrotransposons. The drastic changes in the ZNF675 protein structure happened in the LCA of OWM and apes, which was after the great expansion of THE1 and MST elements, in the LCA of NWM and OWM, around 35 Mya (Lodewijk et al., unpublished). In fact, the marmoset ZNF675 has shown no repressive effect on an MSTA element, member of the MST family (Figure 3.06, *bottom*). The human ZNF675 repressed the MSTA element by 89% in the reporter assay (Figure 3.06, *bottom*). According to the model described by Jacobs and colleagues, since the marmoset ZNF675 did not have affinity for these elements when they were active, a new ZNF675 arose in order to silence them. In the NWMs, even though THE1 and MST elements were not repressed by the marmoset ZNF675, other KZNF proteins are found to bind in the LTR region, which might have a repressive effect. Moreover, ZNF586 emerged in the LCA of NWM and OWM and it binds to these elements, in the internal part (Lodewijk et al., unpublished). However, the binding affinity may have not been strong enough, therefore the THE1 and MST elements remained active and a new repressor was required.

THE1B elements, members of the THE1 family, are repressed by ZNF675, though there are some of these elements that the ZNF675 does not bind to. The main difference between them is the region of the predicted ZNF675 binding site, where several changes have occurred (Figure 3.02). Since KZNF proteins bind to consecutive DNA base pairs (Persikov & Singh, 2014), changes in these DNA motifs, such as indels, will affect the KZNF protein binding capacity. THE1 elements, together with MST elements, have been inactive in our genome for millions of years, so their silencing to prevent them from

transposing is no longer needed. However, by binding to these retrotransposons, the ZNF675 repressive effect might have a function in the establishment of a heterochromatin state in important regions of the genome.

KZNFs are known to bind to gene promoters, which is thought to be a collateral consequence of the arms race against TEs. ZNF675 binds to genes that are involved in cell cycle, signalling, metabolism and also some related to disease (Table 3.01). Of these genes, 25% are associated with development and neuronal processes, including HES1, MCPH1 and SESN3. The binding site of ZNF675, in 54% of the cases, is in the same region as that of transcription factors. This means that the ZNF675 might indeed be involved in the regulation of gene expression. The repression mechanism of KZNF proteins, through the recruitment of KAP1, is most likely similar when there is also binding of another TFs in the same region. However, the repression effect might change since there is probably competition with the other factors. Furthermore, ZNF675 binds to a TE in 48% of the cases, in agreement with the hypothesis that these elements function as regulatory platforms to regulate nearby gene expression (Imbeault et al., 2017; Najafabadi et al., 2015). In the NWM, when these elements inserted themselves in promoter regions, the gene expression levels might have changed because of its enhancer/repressive potential or the different binding of TFs. When the ZNF675 arose in LCA of OWM and apes, it started to repress these elements, thus changing the genes' expression once more. This regulatory modifications to important genes related to development or neuronal processes might have influenced the evolution of primates.

Conservation of the ZNF675 target genes across primate species is important in the study of the ZNF675 regulatory role during evolution. HES1, a gene involved in brain development, has remained conserved throughout primates' lineage, and so has its promoter region (data not shown). However, it is known to be up-regulated in neuroblastoma cells, after induced differentiation with RA (Jögi et al., 2002). The same experiment was performed in a neuroblastoma cell line (SK-N-SH), confirming the up-regulation of HES1 upon RA treatment (Figure 3.03). This experiment's goal was to compare the difference of HES1 expression in ZNF675 KO cell lines, with and without induction via RA. Since the $ZNF675^{-/-}$ cells were in fact heterozygotic and that the SK-N-SH cells were composed of two different cell subtypes, the experiment was not concluded.

MCPH1, also a gene involved in brain development, on the contrary, has been under positive selection during primates' evolution. Changes in its protein-coding region occurred mainly in the LCA of OWMs and apes, and continued to be positively selected in the humans' lineage (Evans et al., 2004; Y. Q. Wang & Su, 2004). The MCPH1 promoter has been overall conserved throughout primates' evolution, since the LCA of NWM and OWM. The basal primates' promoter region, however, is extremely different from that of the evolutionarily later species (Figure S1). For that reason, their sequence was not part of the analysis. ZNF675 binding site on the MCPH1 promoter is one of the regions that underwent more modifications during primates' evolution (Figure 3.04, *bottom*). Specifically, apes and humans have a deletion of 3-4 nucleotides, when comparing to the sequence of NWMs. This change could have had a drastic effect on ZNF675's binding capacity. Furthermore, the binding site of ZNF675 in the MCPH1

promoter of OWM has also undergone modifications, which coincides with the structural changes of ZNF675.

The binding of human ZNF675 to the human MCPH1 promoter region was shown to have no effect in the reporter expression (Figure 3.06a, *bottom*). On the contrary, in the marmoset MCPH1 promoter region, both human and marmoset ZNF675 had a similar effect on the reporter, repressing it by 11% and 8%, respectively. Together with the fact that the marmoset ZNF675 had the same effect in both promoters, changes in the ZNF675 binding site might have resulted in the promoter of MCPH1 evading ZNF675 repression. However, the difference in repression can be due to changes in the ZNF675 binding site and/or in its structure. Therefore, it is necessary to do a reporter assay with the ZNF675 of crab-eating macaque and its MCPH1 promoter. Since crab-eating macaques share the same form of ZNF675 as humans, the only variable will be the changes in the MCPH1 promoter. Difference in binding will then clarify the cause for the evolutionary derepression of MCPH1. Since there is no accurate way of knowing the ancestral version of the MCPH1 promoter, this analysis was done considering the marmoset MCPH1 promoter as being the closest to that of the LCA of NWM and OWM. However, changes in the ZNF675 binding site might have happened in the NWM lineage, after the split.

Even though the ZNF675 was shown to bind to the MCPH1 promoter, it does not seem to have a repressive effect on gene expression. The ZNF675 might not have a regulatory function on the MCPH1 gene in humans, though it once had in NWMs, which shows the evolutionary derepression of this neurodevelopmental gene. However, since the luciferase reporter assays are performed in mESC, the results cannot be directly translated to the human context. Since the ZNF675 acts in a cell-type specific manner (Imbeault et al., 2017), it might be differentially expressed in neuronal cells, thus repressing the MCPH1 expression. Thereby, this repressive effect cannot be accurately assessed in mESCs.

MCPH1 is an important gene in brain development, because of its role in the regulation of brain size. It controls neurogenesis, by regulating the G2/M checkpoint, in order to maintain the pool of neuroprogenitors (Liu et al., 2016). In cases of MCPH1 deficiency, there is a delay in cell division. During neurogenesis, this delay, even if subtle, can have a strong impact on the final quantity of neurons, causing microcephaly (Arroyo et al., 2017). In agreement with this, a difference of 8% in MCPH1 expression could have had a big effect on brain size during primates' evolution. To further analyse the ZNF675 binding affinity in different primates, MCPH1 promoter region of 1200bp was isolated (Figure 3.07). However, further work still must be done since the isolation of the marmoset MCPH1 sequence was unsuccessful.

The genome editing CRISPR/Cas9 system has been used to modify the genome with high specificity (Ran et al., 2013). This system was chosen as a tool to replace the ZNF675 binding site in the human MCPH1 promoter by the marmoset. After a double-stranded break in the DNA, the mammalian cell repairs the damage through non-homologous end-joining (NHEJ) or homology-directed repair. The latter uses a homologous DNA sequence to repair the damage faithfully, this can be a sister chromatid or even a ssDNA oligo. Therefore, an exogenous ssDNA oligo with the marmoset binding site for ZNF675 flanked by human sequences was used as template for the HDR. The MCPH1 promoter sequence replacement was tested, although results still need to be elucidated by sequencing of the

modified region (Figure 3.09). If this strategy is successful, regulation of the MCPH1 by the ZNF675 can be studied endogenously, and reporter assays can be performed, for instance to confirm the ZNF675 binding site.

SESN3, though not a gene involved in brain development, has been associated with epilepsy, being described as the regulator of proconvulsant genes (Johnson et al., 2015). The conservation of this gene has not yet been studied, however the promoter region conservation was analysed here. Similarly to the MCPH1 promoter, the overall region has been conserved throughout primates' evolution. Although, when comparing to the sequence of basal primates, changes in the ZNF675 binding site are mainly present in the NWMs, being highly conserved in the lineage from OWMs to humans (Figure 3.05, *bottom*). The pattern of changes of the ZNF675 binding site coincides with its structural changes. This is initial evidence for a stabilization of the regulatory function of ZNF675 on the SESN3 gene.

The human ZNF675 showed a strong effect on the human SESN3 promoter, repressing the reporter expression by 48% (Figure 3.06b, *bottom*). On the contrary, the human ZNF675 had no effect on the marmoset promoter, neither the marmoset ZNF675 on the human nor the marmoset SESN3. This means that structural changes of the ZNF675 had no influence in its binding capacity, in this case. Only changes in the binding site of ZNF675 increased the binding affinity of this KZNF for the SESN3 promoter. The repression level of SESN3 was maintained throughout primates' evolution, since there were no other modifications to the ZNF675 binding site, besides those in the LCA of OWMs and apes.

ZNF675 has been associated with several neurological diseases. Even though epilepsy has not been described as one, there is a reported patient (DECIPHER ID: 289120), with a deletion of both ZNF675 and ZNF681 proteins, who has seizures. Since ZNF675 is thought to strongly repress SESN3 expression, derepression of this gene may imply an up-regulation of proconvulsant genes. This might, therefore, be one of the causes for the reported seizures.

The analysis of clinical data from DECIPHER has revealed that whenever there is a deletion or duplication of ZNF675, the ZNF681 is also included. For this reason, in order to study the effect on the transcriptome in case of deletion, a ZNF675/ZNF681 KO cell line has to be created. In this way, changes in gene expression will shed light on the regulatory function of both ZNF675 and ZNF681, that are compromised in a disease background. The CRISPR/Cas9 system was chosen, thereby gRNAs were designed for ZNF675, targeting exon 1, and for ZNF681, targeting exon 4, which codes for the ZNF domain (Figure 3.11). The ZNF675/ZNF681 KO was successful. HDR-mediated genome editing, through exogenous ssDNA oligos, has been described to improve the efficiency of the genome editing (Chu et al., 2015; Schiel, Chou, Mayer, Anderson, & van Brabant Smith, 2016). This approach was tested and although the efficiency of the deletion seemed not to have improved (Figure 3.12), to exclude an effect of the oligo the necessary controls still have to be conducted. Since the size of the deletion is 107kb, a high deletion efficiency might be reached with the CRISPR/Cas9 system alone. In fact, this approach was successful in inducing a much bigger deletion of 725kb in human induced pluripotent stem cell (hiPSC)-derived muscle cells (Young et al., 2016).

Current information regarding the human ZNF675 is enough to analyse its binding throughout the genome. The evolutionary history, however, still needs to be completed. ChIP-seq of the marmoset ZNF675 is necessary to reveal if or which retrotransposons, or even gene promoters, it binds to. The ChIP was performed in HEK293T, for both marmoset and human ZNF675, the latter used as control. There was an enrichment for the MCPH1 and SESN3 promoter regions, and for THE1 elements, in the ChIP of human ZNF675 (Figure 3.13b). However, the ChIP of marmoset ZNF675 showed no significant enrichment. Since sonication of the cross-linked DNA is done before the ChIP, fragmentation of accessible regions, such as promoters, is more frequent than of inaccessible regions. Therefore, it is usual for promoter regions to be enriched in a ChIP sample. Since the control for the qPCR was a non-coding region, the promoters' enrichment could not be corrected. Through the analysis of the immunoblot, the qPCR results were confirmed (Figure 3.13c), the marmoset ZNF675 is absent, whilst the human ZNF675 is present, though with a very low signal. Even though the ChIP was repeated in U2OS cells, the results were the same (Figure 3.14). A protocol optimization is required, firstly by decreasing the number of washes after the IP, which might be the explanation for the loss of signal. However, the lack of enrichment for the cZNF675 can also be explained by the fact that this KZNF protein does not bind to the regions the hZNF675 binds to. Even though it would be expected, by the analysis of the luciferase assay (Figure 3.06a, *bottom*), for the cZNF675 to bind to the MCPH1 promoter.

5 Conclusion and Future Perspectives

This project aimed to elucidate the evolutionary history of ZNF675, its effect on the genes that were caught in the arms race, and how it relates to the human brain. The burst of THE1 and MST elements have occurred in the LCA of NWM and OWM, which was before the structural changes the ZNF675 has undergone. Together with the lack of effect by the marmoset ZNF675 towards THE1 and MST elements' silencing, the human ZNF675 might have arisen in order to prevent these elements from transposing. When the novel ZNF675 was created, it started to bind to gene promoters, changes to which resulted in the evasion of repression (Figure 5.01, *top*), such as MCPH1, or in the stabilization of the regulation (Figure 5.01, *bottom*), such as in SESN3. Today, several neurological disorders are associated with mutations in both ZNF675 gene and in its regulatory targets. Thus, throughout evolution, modification of these genes' expression levels by ZNF675 might have had a significant impact on the human brain.

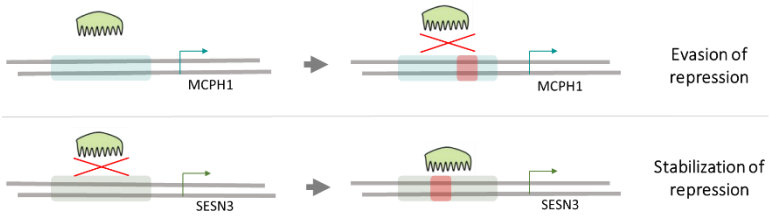


Figure 5.01 – **Representation of the changes in binding capacity of ZNF675 during primate evolution**. Two examples of promoter responses to the human ZNF675 (green) binding: evasion of repression (*top*) and stabilization of repression (*bottom*).

Further research remains to be done, the priority being luciferase report assays with the crab-eating macaque MCPH1 promoter and its ZNF675, to clarify what led to the derepression of MCPH1 during primates' evolution. The ChIP-seq of the marmoset ZNF675 will unravel the transposable elements it binds to, completing the evolutionary history of ZNF675. The replacement of the ZNF675 binding site in the human MCPH1 promoter by the marmoset is yet to be confirmed. If this strategy is successful, it can also be done for SESN3 promoter or any other interesting gene.

Furthermore, it would be interesting to investigate how ZNF675 affects cortical progenitors in a disease context, i.e. overexpression or knockout of ZNF675. Its effect could be assessed with a cell division assay in cortical progenitors, since MCPH1 is associated with the neuroprogenitors self-renewal. Moreover, cortical organoids could be created and the MCPH1 gene expression monitored in different developmental stages. It is expected that the MCPH1 gene is highly expressed in the early stages, though in the ZNF675 overexpression condition the MCPH1 is expressed in a lower level. However, the MCPH1 effect would be assessed in the final developmental stages, since its deficiency affects the number of neurons produced, thus the size of the organoids. These assays could also be performed on cortical progenitor cells with the marmoset binding site for the ZNF675 in the MCPH1 promoter, in order to confirm the changes in regulation are originated by the ZNF675.

Moreover, ZNF675 KO neurons can be used to investigate the ZNF675 role in the adult brain. Since the SESN3 gene is described as the master regulator of pro-epileptic genes, the effect of the knockout or overexpression of ZNF675 can be assessed with electrophysiology. In parallel, the SESN3 gene expression can be analysed in order to investigate if the high activity of neurons, in a ZNF675 KO condition, is correlated with a high expression of SESN3. This result might further associate the occurrence of seizures, in the patient with a deletion of ZNF675, to the SESN3 activity.

References

Abràmoff, M. D., Magalhães, P. J., & Ram, S. J. (2004). Image processing with imageJ. *Biophotonics International*, *11*(7), 36–41.

Altemose, N., Noor, N., Bitoun, E., Tumian, A., Imbeault, M., Ross Chapman, J., … Myers, S. R. (2017). A map of human PRDM9 binding provides evidence for novel behaviors of PRDM9 and other zinc-finger proteins in meiosis. *ELife*, *6*.

Arroyo, M., Kuriyama, R., Trimborn, M., Keifenheim, D., Cañuelo, A., Sánchez, A., … Marchal, J. A. (2017). MCPH1, mutated in primary microcephaly, is required for efficient chromosome alignment during mitosis. *Scientific Reports*, *7*(1), 1–12.

Ayarpadikannan, S., & Kim, H.-S. (2014). The Impact of Transposable Elements in Genome Evolution and Genetic Instability and Their Implications in Various Diseases. *Genomics & Informatics*, *12*(3), 98.

Cheng, Y., Geng, H., Cheng, S. H., Liang, P., Bai, Y., Li, J., … Tao, Q. (2010). KRAB zinc finger protein ZNF382 is a proapoptotic tumor suppressor that represses multiple oncogenes and is commonly silenced in multiple carcinomas. *Cancer Research*, *70*(16), 6516–6526.

Chu, V. T., Weber, T., Wefers, B., Wurst, W., Sander, S., Rajewsky, K., & Kühn, R. (2015). Increasing the efficiency of homology-directed repair for CRISPR-Cas9-induced precise gene editing in mammalian cells. *Nature Biotechnology*, *33*(5), 543–548.

Cordaux, R., & Batzer, M. A. (2009). The impact of retrotransposons on human genome evolution. *Nature Reviews Genetics*, *10*(10), 691–703.

de Koning, A. P. J., Gu, W., Castoe, T. A., Batzer, M. A., & Pollock, D. D. (2011). Repetitive elements may comprise over Two-Thirds of the human genome. *PLoS Genetics*, *7*(12).

Deininger, P. L., Moran, J. V., Batzer, M. A., & Kazazian, H. H. (2003). Mobile elements and mammalian genome evolution. *Current Opinion in Genetics and Development*, *13*(6), 651–658.

Emerson, R. O., & Thomas, J. H. (2009). Adaptive evolution in zinc finger transcription factors. *PLoS Genetics*, *5*(1).

Evans, P. D., Anderson, J. R., Vallender, E. J., Choi, S. S., & Lahn, B. T. (2004). Reconstructing the evolutionary history of microcephalin, a gene controlling human brain size. *Human Molecular Genetics*, *13*(11), 1139–1145.

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. *American Journal of Human Genetics*, *84*(4), 524–533.

Gifford, W. D., Pfaff, S. L., & MacFarlan, T. S. (2013). Transposable elements as genetic regulatory substrates in early development. *Trends in Cell Biology*, *23*(5), 218–226.

Groner, A. C., Meylan, S., Ciuffi, A., Zangger, N., Ambrosini, G., Dénervaud, N., … Trono, D. (2010). KRAB-zinc finger proteins and KAP1 can mediate long-range transcriptional repression through heterochromatin spreading. *PLoS Genetics*, *6*(3).

Hagenbuchner, J., Kuznetsov, A., Hermann, M., Hausott, B., Obexer, P., & Ausserlechner, M. J. (2012). FOXO3-induced reactive oxygen species are regulated by BCL2L11 (Bim) and SESN3. *Journal of*

Cell Science, *125*(5), 1191–1203.

Hamilton, a T., Huntley, S., Gordon, L., & Stubbs, L. (2005). Evolutionary expansion and divergence in a large family of primate-specific zinc finger transcription factor genes. *Genome Research*, *16*(5), 584–594.

Herculano-Houzel, S., Collins, C. E., Wong, P., & Kaas, J. H. (2007). Cellular scaling rules for primate brains. *Proceedings of the National Academy of Sciences*, *104*(9), 3562–3567.

Huntley, S., Baggott, D. M., Hamilton, A. T., Tran-Gyamfi, M., Yang, S., Kim, J., … Stubbs, L. (2006). A comprehensive catalog of human KRAB-associated zinc finger genes: Insights into the evolutionary history of a large family of transcriptional repressors. *Genome Research*, *16*(5), 669–677.

Imbeault, M., Helleboid, P. Y., & Trono, D. (2017). KRAB zinc-finger proteins contribute to the evolution of gene regulatory networks. *Nature*, *543*(7646), 550–554.

Jackson, A. P., Eastwood, H., Bell, S. M., Adu, J., Toomes, C., Carr, I. M., … Woods, C. G. (2002). Identification of Microcephalin, a Protein Implicated in Determining the Size of the Human Brain. *The American Journal of Human Genetics*, *71*(1), 136–142.

Jacobs, F. M. J., Greenberg, D., Nguyen, N., Haeussler, M., Ewing, A. D., Katzman, S., … Haussler, D. (2014). An evolutionary arms race between KRAB zinc-finger genes ZNF91/93 and SVA/L1 retrotransposons. *Nature*, *516*(7530), 242–245.

Jögi, A., Persson, P., Grynfeld, A., Påhlman, S., & Axelson, H. (2002). Modulation of basic helix-loop-helix transcription complex formation by Id proteins during neuronal differentiation. *The Journal of Biological Chemistry*, *277*(11), 9118–9126.

Johnson, M. R., Behmoaras, J., Bottolo, L., Krishnan, M. L., Pernhorst, K., Santoscoy, P. L. M., … Petretto, E. (2015). Systems genetics identifies Sestrin 3 as a regulator of a proconvulsant gene network in human epileptic hippocampus. *Nature Communications*, *6*.

Kazazian, H. H. (1998). Mobile elements and disease. *Current Opinion in Genetics and Development*, *8*(3), 343–350.

Kent, W. J., Sugnet, C. W., Furey, T. S., Roskin, K. M., Pringle, T. H., Zahler, A. M., & Haussler, a. D. (2002). The Human Genome Browser at UCSC. *Genome Research*, *12*(6), 996–1006.

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

Konkel, M. K., & Batzer, M. A. (2010). A mobile threat to genome stability: The impact of non-LTR retrotransposons upon the human genome. *Seminars in Cancer Biology*, *20*(4), 211–221.

Krishna, S. S., Majumdar, I., & Grishin, N. V. (2003). Structural classification of zinc fingers. *Nucleic Acids Research*, *31*(2), 532–550.

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., … International Human Genome Sequencing, C. (2001). Initial sequencing and analysis of the human genome. *Nature*, *409*(6822), 860–921.

Liu, X., Zhou, Z. W., & Wang, Z. Q. (2016). The DNA damage response molecule MCPH1 in brain development and beyond. *Acta Biochimica et Biophysica Sinica*, *48*(7), 678–685.

Lonsdale, J., Thomas, J., Salvatore, M., Phillips, R., Lo, E., Shad, S., … Moore, H. F. (2013). The

Genotype-Tissue Expression (GTEx) project. *Nature Genetics*, *45*(6), 580–585.

Lowe, C. B., Bejerano, G., & Haussler, D. (2007). Thousands of human mobile element fragments undergo strong purifying selection near developmental genes. *Proceedings of the National Academy of Sciences*, *104*(19), 8005–8010.

Lupo, A., Cesaro, E., Montano, G., Zurlo, D., Izzo, P., & Costanzo, P. (2013). KRAB-Zinc Finger Proteins: A Repressor Family Displaying Multiple Biological Functions. *Current Genomics*, *14*(4), 268–278.

Mantere, T., Winqvist, R., Kauppila, S., Grip, M., Jukkola-Vuorinen, A., Tervasmäki, A., … Pylkäs, K. (2016). Targeted Next-Generation Sequencing Identifies a Recurrent Mutation in MCPH1 Associating with Hereditary Breast Cancer Susceptibility. *PLoS Genetics*, *12*(1), 1–14.

McClintock, B. (1950). The origin and behavior of mutable loci in maize. *Proceedings of the National Academy of Sciences*, *36*(6), 344–355.

Mills, R. E., Bennett, E. A., Iskow, R. C., & Devine, S. E. (2007). Which transposable elements are active in the human genome? *Trends in Genetics*, *23*(4), 183–191.

Murata, K., Hattori, M., Hirai, N., Hirata, H., Kageyama, R., Shinozuka, Y., … Minato, N. (2005). Hes1 Directly Controls Cell Proliferation through the Transcriptional Repression of Hes1 Directly Controls Cell Proliferation through the Transcriptional Repression of p27 Kip1. *Molecular and Cellular Biology*, *25*(10), 4262–4271.

Najafabadi, H. S., Mnaimneh, S., Schmitges, F. W., Garton, M., Lam, K. N., Yang, A., … Hughes, T. R. (2015). C2H2 zinc finger proteins greatly expand the human regulatory lexicon. *Nature Biotechnology*, *33*(5), 555–562.

Nakamura, Y., Sakakibara, S. I., Miyata, T., Ogawa, M., Shimazaki, T., Weiss, S., … Okano, H. (2000). The bHLH gene Hes1 as a repressor of the neuronal commitment of CNS stem cell. *The Journal of Neuroscience*, *20*(1), 283–293.

Nogueira, V., Park, Y., Chen, C. C., Xu, P. Z., Chen, M. L., Tonic, I., … Hay, N. (2008). Akt Determines Replicative Senescence and Oxidative or Oncogenic Premature Senescence and Sensitizes Cells to Oxidative Apoptosis. *Cancer Cell*, *14*(6), 458–470.

Nowick, K., Gernat, T., Almaas, E., & Stubbs, L. (2009). Differences in human and chimpanzee gene expression patterns define an evolving network of transcription factors in brain. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(52), 22358–22363.

Nowick, K., Hamilton, A. T., Zhang, H., & Stubbs, L. (2010). Rapid sequence and expression divergence suggest selection for novel function in primate-specific KRAB-ZNF genes. *Molecular Biology and Evolution*, *27*(11), 2606–2617.

Okonechnikov, K., Golosova, O., Fursov, M., Varlamov, A., Vaskin, Y., Efremov, I., … Tleukenov, T. (2012). Unipro UGENE: A unified bioinformatics toolkit. *Bioinformatics*, *28*(8), 1166–1167.

Pace, J. K., Feschotte, C., & Ii, J. K. P. (2007). intense activity in the primate lineage The evolutionary history of human DNA transposons: Evidence for The evolutionary history of human DNA transposons: Evidence for intense activity in the primate lineage. *Genome Res. Cold Spring Harbor Laboratory Press on December Morgan Oosumi et Al. Robertson Smit and Riggs Robertson and Martos Robertson and Zumpano*, *17*(3), 422–432.

Peng, H., Begg, G. E., Harper, S. L., Friedman, J. R., Speicher, D. W., & Rauscher, F. J. (2000). Biochemical analysis of the Kruppel-associated box (KRAB) transcriptional repression domain: Spectral, kinetic, and stoichiometric properties of the KRAB·KAP-1 complex. *Journal of Biological Chemistry*, *275*(24), 18000–18010.

Persikov, A. V., & Singh, M. (2014). De novo prediction of DNA-binding specificities for Cys2His2zinc finger proteins. *Nucleic Acids Research*, *42*(1), 97–108.

Ran, F. A., Hsu, P. D., Wright, J., Agarwala, V., Scott, D. A., & Zhang, F. (2013). Genome engineering using the CRISPR-Cas9 system. *Nature Protocols*, *8*(11), 2281–2308.

Schiel, J. A., Chou, E., Mayer, M., Anderson, E. M., & van Brabant Smith, A. (2016). Homology-directed repair with Dharmacon ™ Edit-R ™ CRISPR-Cas9 reagents and single-stranded DNA oligos, (1).

Schmidt, D., Wilson, M. D., Spyrou, C., Brown, G. D., Hadfield, J., & Odom, D. T. (2009). ChIP-seq: Using high-throughput sequencing to discover protein-DNA interactions. *Methods*, *48*(3), 240–248.

Schultz, D. C., Ayyanathan, K., Negorev, D., Maul, G. G., & Rauscher Iii, F. J. (2002). SETDB1 : a novel KAP - 1 - associated histone H3 , lysine 9 - specific methyltransferase that contributes to HP1 - mediated silencing of euchromatic genes by KRAB zinc - finger proteins. *Genes and Development*, *16*, 919–932.

Schultz, D. C., Friedman, J. R., & Rauscher, F. J. (2001). Targeting histone deacetylase complexes via KRAB-zinc finger proteins: The PHD and bromodomains of KAP-1 form a cooperative unit that recruits a novel isoform of the Mi-2α subunit of NuRD. *Genes and Development*, *15*(4), 428–443.

Shannon, M., Hamilton, A. T., Gordon, L., Branscomb, E., & Stubbs, L. (2003). Differential Expansion of Zinc-Finger Transcription Factor Loci in Homologous Human and Mouse Gene Clusters Differential Expansion of Zinc-Finger Transcription Factor Loci in Homologous Human and Mouse Gene Clusters. *Genome Research*, *19*, 1097–1110.

Thomas, J. H., & Schneider, S. (2011). Coevolution of retroelements and tandem zinc finger genes. *Genome Research*, *21*(11), 1800–1812.

Varki, A., & Altheide, T. K. (2005). Comparing the human and chimpanzee genomes: Searching for needles in a haystack. *Genome Research*, *15*(12), 1746–1758.

Wang, H., Xing, J., Grover, D., Hedges Kyudong Han, D. J., Walker, J. A., & Batzer, M. A. (2005). SVA elements: A hominid-specific retroposon family. *Journal of Molecular Biology*, *354*(4), 994–1007.

Wang, M., Xu, Y., Liu, J., Ye, J., Yuan, W., Jiang, H., … Wan, J. (2018). Recent Insights into the Biological Functions of Sestrins in Health and Disease. *Cellular Physiology and Biochemistry*, *43*(5), 1731–1741.

Wang, Y. Q., & Su, B. (2004). Molecular evolution of microcephalin, a gene determining human brain size. *Human Molecular Genetics*, *13*(11), 1131–1137.

Williams, M. F. (2002). Primate encephalization and intelligence. *Medical Hypotheses*, *58*(4), 284–290.

Wolf, D., & Goff, S. P. (2009). Embryonic stem cells use ZFP809 to silence retroviral DNAs. *Nature*, *458*(7242), 1201–1204.

Yang, P., Wang, Y., & Macfarlan, T. S. (2017). The Role of KRAB-ZFPs in Transposable Element Repression and Mammalian Evolution. *Trends in Genetics*, *33*(11), 871–881.

Ying, Q. L., Stavridis, M., Griffiths, D., Li, M., & Smith, A. (2003). Conversion of embryonic stem cells

into neuroectodermal precursors in adherent monoculture. *Nature Biotechnology*, *21*(2), 183–186.

Young, C. S., Hicks, M. R., Ermolova, N. V., Nakano, H., Jan, M., Younesi, S., … Pyle, A. D. (2016). A Single CRISPR-Cas9 Deletion Strategy that Targets the Majority of DMD Patients Restores Dystrophin Function in hiPSC-Derived Muscle Cells. *Cell Stem Cell*, *18*(4), 533–540.

Zeng, Y., Wang, W., Ma, J., Wang, X., Guo, M., & Li, W. (2012). Knockdown of ZNF268, which is transcriptionally downregulated by GATA-1, promotes proliferation of K562 cells. *PLoS ONE*, *7*(1), 1–8.

Zhou, L., Bai, Y., Li, Y., Liu, X., Tan, T., Meng, S., … Dong, Z. (2016). Overexpression of MCPH1 inhibits uncontrolled cell growth by promoting cell apoptosis and arresting the cell cycle in S and G2/M phase in lung cancer cells. *Oncology Letters*, *11*(1), 365–372.

Supplemental material

S.1 THE1 elements

Table S1 – **THE1B-int elements bound and not bound by ZNF675.** The first column corresponds to the chromosomal location of 15 THE1 elements bound by ZNF675 and 15 elements not bound. The second column corresponds to the elements' coordinates. The similarity score is the third column, the strand orientation and the length of the elements are in the fourth and fifth columns, respectively.

	Chromosome	*Coordinates*	*Score*	*Strand*	*Length*
Elements bound by ZNF675	chr5	61453737-61455317	13075	+	1580
	chr5	73819041-73820624	13046	+	1583
	chr5	55924010-55925584	12968	+	1574
	chr3	16028965-16030534	12952	-	1569
	chr5	6011618-6013192	12755	-	1574
	chr1	169840659-169842218	12689	+	1559
	chr3	26644529-26646113	12647	-	1584
	chr6	74391796-74393357	12617	-	1561
	chr2	99284480-99285631	12611	+	1151
	chr1	85714013-85715583	12604	-	1570
	chr5	145372393-145373969	12574	-	1576
	chr3	72112105-72113666	12561	-	1561
	chr1	60201958-60203533	12556	+	1575
	chr14	67917006-67918585	12552	+	1579
	chr7	144123945-144125491	12496	-	1546
Elements not bound by ZNF675	chrX	15918445-15920029	13091	-	1584
	chr4	188409175-188410750	12906	+	1575
	chr1	46523039-46524600	12852	+	1561
	chr4	76457187-76458755	12805	-	1568
	chr10	128047807-128049388	12684	-	1581
	chr9	74904612-74905701	12614	-	1089
	chr15	70883852-70885421	12594	+	1569
	chr5	81675037-81676605	12574	-	1568
	chr7	86927549-86929126	12566	-	1577
	chr18	10942179-10943758	12547	+	1579
	chr9	80305281-80306480	12522	-	1199
	chr2	118394242-118395812	12453	+	1570
	chr5	90642949-90644522	12446	-	1573
	chr3	194240024-194241601	12432	+	1577
	chr4	39372116-39373685	12409	+	1569

S.2 MCPH1 promoter sequence conservation in basal primates

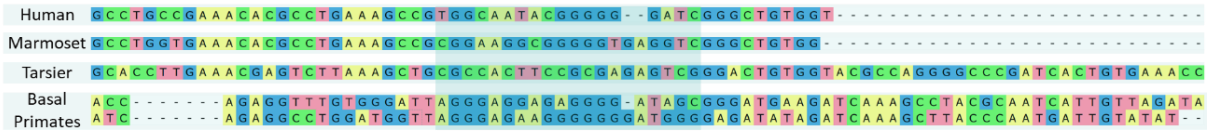


Figure S1 – **Conservation of the MCPH1 promoter region in basal primates, centred in the ZNF675 binding site.** Comparison between sequences of human, marmoset, tarsier and basal primates (mouse lemur and bushbaby), using Multiple Sequence Alignment on UGENE. Shady region in the centre is the binding site of ZNF675. Sequences are reversed (3' to 5').