



UvA-DARE (Digital Academic Repository)

Structural Evolution of Gene Promoters Driven by Primate-Specific KRAB Zinc Finger Proteins

Farmiloe, G.; van Bree, E.J.; Robben, S.F.; Janssen, L.J.M.; Mol, L.; Jacobs, F.M.J.

DOI

[10.1093/gbe/evad184](https://doi.org/10.1093/gbe/evad184)

Publication date

2023

Document Version

Final published version

Published in

Genome Biology and Evolution

License

CC BY-NC

[Link to publication](#)

Citation for published version (APA):

Farmiloe, G., van Bree, E. J., Robben, S. F., Janssen, L. J. M., Mol, L., & Jacobs, F. M. J. (2023). Structural Evolution of Gene Promoters Driven by Primate-Specific KRAB Zinc Finger Proteins. *Genome Biology and Evolution*, 15(11), Article evad184. <https://doi.org/10.1093/gbe/evad184>

General rights



It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations

If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: <https://uba.uva.nl/en/contact>, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.

UvA-DARE is a service provided by the library of the University of Amsterdam (<https://dare.uva.nl>)

Structural Evolution of Gene Promoters Driven by Primate-Specific KRAB Zinc Finger Proteins

Grace Farmiloe ^{1,2,†}, Elisabeth J. van Bree^{1,2,†}, Stijn F. Robben¹, Lara J.M. Janssen¹, Lisa Mol¹, and Frank M.J. Jacobs ^{1,2,*}

¹Swammerdam Institute for Life Sciences, Evolutionary Neurogenomics, University of Amsterdam, Amsterdam, The Netherlands

²Complex Trait Genetics, Amsterdam Neuroscience, Amsterdam, The Netherlands

[†]The first two authors contributed equally to the study.

*Corresponding author: E-mail: f.m.j.jacobs@uva.nl.

Accepted: October 09, 2023

Abstract

Krüppel-associated box (KRAB) zinc finger proteins (KZNFs) recognize and repress transposable elements (TEs); TEs are DNA elements that are capable of replicating themselves throughout our genomes with potentially harmful consequences. However, genes from this family of transcription factors have a much wider potential for genomic regulation. KZNFs have become integrated into gene-regulatory networks through the control of TEs that function as enhancers and gene promoters; some KZNFs also bind directly to gene promoters, suggesting an additional, more direct layer of KZNF co-option into gene-regulatory networks. Binding site analysis of *ZNF519*, *ZNF441*, and *ZNF468* suggests the structural evolution of KZNFs to recognize TEs can result in coincidental binding to gene promoters independent of TE sequences. We show a higher rate of sequence turnover in gene promoter KZNF binding sites than neighboring regions, implying a selective pressure is being applied by the binding of a KZNF. Through CRISPR/Cas9 mediated genetic deletion of *ZNF519*, *ZNF441*, and *ZNF468*, we provide further evidence for genome-wide co-option of the KZNF-mediated gene-regulatory functions; KZNF knockout leads to changes in expression of KZNF-bound genes in neuronal lineages. Finally, we show that the opposite can be established upon KZNF overexpression, further strengthening the support for the role of KZNFs as bona-fide gene regulators. With no eminent role for *ZNF519* in controlling its TE target, our study may provide a snapshot into the early stages of the completed co-option of a KZNF, showing the lasting, multilayered impact that retrovirus invasions and host response mechanisms can have upon the evolution of our genomes.

Key words: Genomics, primate evolution, gene regulation, KRAB zinc finger proteins.

Significance

Previous research has investigated the role of KRAB zinc finger genes as repressors of transposable elements, recent data revealed that a number of KZNFs also bind to gene promoters but their role at these genomic sites is not well understood. In this study, we investigate the impact of this KZNF-promoter relationship and how the emergence of new KZNFs that recognize promoters has shaped gene-regulatory networks. We observed a higher-than-average sequence turnover at bound promoters and an increase in the expression of genes with promoter KZNF sites when the binding KZNF was removed. We conclude that members of this gene family have a subtle yet important influence on the shaping of primate gene-regulatory networks and shed light upon a transitional state in genome evolution which may contribute to primate-specific traits such as increased brain size.

© The Author(s) 2023. Published by Oxford University Press on behalf of Society for Molecular Biology and Evolution.

This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial License (<https://creativecommons.org/licenses/by-nc/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

Introduction

Krüppel-associated box (KRAB) zinc finger proteins (KZNFs) are a type of transcription factor (TF) encoded by a large gene family consisting of more than 400 loci. These genes have evolved over time through tandem duplication followed by sequence divergence to create a large collection of DNA-binding proteins that mostly function as transcriptional repressors (Emerson and Thomas 2009; Thomas and Schneider 2011; Bruno et al. 2019; Wolf et al. 2020). A typical KZNF protein contains a highly conserved KRAB domain that interacts with the co-repressor protein TRIM28 (KAP1) and a zinc finger domain, containing a highly variable array of zinc fingers that recognize and bind to specific DNA sequences (Gröner et al. 2010). Members in the KZNF gene family are primarily distinguished by differences in the composition and size of the zinc finger domain, although other structural variations, including in the KRAB domain, are also represented (Looman et al. 2002). DNA interaction is encoded in just four amino acids in each zinc finger and a single nucleotide substitution can alter the DNA binding capacities of the zinc finger unit, and the KZNF protein in total (Looman et al. 2002). The structure of KZNF genes, amplified by the fact that most KZNFs reside in highly unstable genomic loci, provides a source of rapidly flexible gene-regulatory mechanisms that have high potential to fall under evolutionary mechanisms (Nowick et al. 2010).

In recent years it was established that a large number of individual KZNFs recognize and suppress different classes of transposable elements (TEs) (Thomas and Schneider 2011; Jacobs et al. 2014; Najafabadi et al. 2015; Schmitges et al. 2016; Imbeault et al. 2017; Helleboid et al. 2019). Although this was shown in detail for just a few specific examples, the general concept is that the initial role of KZNF genes is to target and prevent newly arising TE families from replicating in genomes in order to safeguard the integrity of the host genome (Helleboid et al. 2019). However, due to their repetitive nature, TEs are prone to rapidly mutate and sometimes escape binding by the KZNF which can give rise to a new wave of TE insertions (Jacobs et al. 2014). Over time, a new or adapted KZNF gene will evolve to repress this new class of TEs, only to force it to develop new mutations to escape its repressor once again. This pattern of genomic evolution, remarkably similar to a classical evolutionary arms race, has been taking place independently in many different species, to the extent that many species have their own species-specific KZNF genes (Thomas and Schneider 2011; Castro-Diaz et al. 2014; Jacobs et al. 2014; Ecco et al. 2017).

The regulatory effects of KZNFs are not only limited to the repression of TEs. After, or in parallel to their initial role, KZNFs are often co-opted for other functions. Some remain associated with TEs to form KZNF-controlled TE-mediated regulatory elements (Ecco et al. 2016; Pontis

et al. 2019). These transposable element-embedded regulatory sequences (TEeRS) and their associated KZNFs have been shown to regulate gene expression during neuronal development in humans (Turelli et al. 2020). KZNFs also play a role in the regulation of TE-mediated cryptic gene promoters (Sundaram and Wysocka 2020; Haring et al. 2021). Here, the KZNF regulates gene expression by binding directly to a TE-derived gene promoter. *ZNF57* and *ZNF445* have been shown to play important roles in imprinting (Li et al. 2008; Quenneville et al. 2011; Takahashi et al. 2019) and *ZNF568* has been shown to regulate *IGF2* in the placenta (Yang et al. 2017).

It was recently shown that in addition to their prime TE target sites, a subgroup of KZNFs also has the ability to recognize and bind to gene promoters independent of TEs (Frietze et al. 2010; Schmitges et al. 2016; Imbeault et al. 2017; Helleboid et al. 2019; Farmiloe et al. 2020). The frequency of promoter occupancy varies widely between KZNFs, with some KZNFs showing binding capacity to over 2,000 gene promoters. Comparative analysis of the KZNF binding sites in TEs and the TE-independent binding sites in gene promoters revealed clear similarities in DNA sequence (Farmiloe et al. 2020). The binding of the KZNF to gene promoters may have been an inevitable side effect of the acquired recognition potential of the KZNF to the target TE, but once established, this secondary and coincidental regulatory involvement of the KZNF could have become indispensable for normal gene regulation. In fact, after the TE itself has lost its capacity to retrotranspose, the KZNF may have become redundant for the control of the TE, but not for the genes it has evolved to regulate directly through their promoters (Ecco et al. 2017). We previously showed that in general, the binding of KZNFs to promoters correlates well with brain-developmental gene expression patterns (Farmiloe et al. 2020), indicating that binding of KZNFs influences neuronal gene expression.

From an evolutionary perspective, the recent emergence of many primate-specific KZNFs raises the important question of how genes have coped with these new regulators that may have created a temporary imbalance creating an evolutionary impetus for stabilization. Such newly emerged regulatory roles of KZNFs on gene regulation would likely have evoked structural changes in the gene promoters they developed the ability to bind, which may have required further “stabilizing” genomic adaptations to cope with the newly acquired KZNF-mediated level of gene regulation. We therefore hypothesize that gene promoters bound by KZNFs have been under a similar pressure to change and “escape” binding by KZNFs as TEs. In this study, we investigated the relationship between KZNFs and KZNF-bound gene promoters to elucidate the extent of the evolutionary pressure that KZNFs have exerted on human gene expression patterns.

Results

ZNF519, ZNF441, and ZNF468 are Widely Expressed Primate-specific KZNFs, Binding to Thousands of Gene Promoters Genome-Wide

51 KZNFs that bind to 100 or more gene promoters were outlined in Farmiloe et al. (2020). To investigate genomic adaptations of the host genome in response to the emergence of the KZNF-mediated level of gene regulation, we focused on recently evolved promoter-binding KZNFs that are expressed in a variety of tissues. Out of the set of promoter-binding KZNFs, we selected three KZNFs based on the following criteria; 1) these KZNFs emerged specifically in primates, 2) they interacted directly with a high number of gene promoters, and 3) they showed high levels of expression in hESCs and/or hESC-derived cortical tissues (fig. 1A and B). The KZNFs that fit our criteria best were *ZNF519*, *ZNF441*, and *ZNF468* and they were selected for further in-depth investigation. All three KZNFs are expressed in the brain and, to a lesser degree in a range of human tissues (supplementary fig. 1, Supplementary Material online). These selection criteria led to the exclusion of some KZNFs, including those that were shown to bind to the highest number of gene promoters (*ZNF202*, *ZNF534*).

Comparative genomics analysis confirmed that all three KZNFs are primate-specific and not present in any of the species that diverged before the last common ancestor with new world monkeys (NWMs) (Imbeault et al. 2017) (supplementary fig. 2, Supplementary Material online). Traces of the *ZNF519* locus can be found in the squirrel monkey genome which suggests the presence of a *ZNF519* gene in the common ancestor of humans and NWMs that has been lost in the earlier lineages. Now *ZNF519* is only present in gibbons and great apes including humans (fig. 1E). *ZNF441* is detected in the genomes of NWMs, old world monkeys (OWMs), gibbons, and great apes. *ZNF468* emerged later and is only detected in the genomes of OWMs, gibbons, and great apes.

Like many other KZNFs, *ZNF519*, *ZNF441*, and *ZNF468* are associated with a specific class of TEs (Table 1, Imbeault et al. 2017). For each TE class studied we see a rise in the number of elements at the same time as we see the emergence of each KZNF, suggesting their emergence is linked to the respective TE invasions. The number of MER52 elements, a class bound by *ZNF519* (Imbeault et al. 2017), increases in NWMs and OWMs and then remains relatively consistent in the great apes indicating a lack of transposition activity (fig. 1D, supplementary fig. 2A, Supplementary Material online). *ZNF441* has been shown to target AluY elements, specifically subclasses AluY and AluYa5 (Imbeault et al. 2017). AluY elements are still active in the human genome (Bennett et al. 2008). Their numbers show large species-specific expansions in OWMs (fig. 1D, supplementary fig. 2C, Supplementary Material online). MER11A elements, bound by *ZNF468*

(Imbeault et al. 2017), follow a similar evolutionary pattern as MER52 elements. They emerge in OWMs and we see a rapid expansion in these species followed by stabilization of numbers in apes and humans (fig. 1D, supplementary fig. 2C, Supplementary Material online).

Further analysis of the ChIP data for these KZNFs shows a defined peak at associated TEs (fig. 1C). It is possible that the KZNF remains important to control the TE's regulatory potential, even if the TE itself has lost the capability to retrotranspose. To address this possibility we tested the activity of MER52 elements in a luciferase assay in the absence and presence of *ZNF519*, its partner KZNF. The luciferase assay was performed in mouse embryonic stem cells (mESCs), which lack all primate-specific KZNFs including *ZNF519*. A MER52D element cloned upstream of luciferase had a mild repressive effect on luciferase expression. (supplementary fig. 3, Supplementary Material online, two-way analysis of variance (ANOVA), Tukey's multiple comparison test $P < 0.0001$). When *ZNF519* was ectopically expressed, there was no change to the regulatory effect of the MER52D element on luciferase activity (supplementary fig. 3, Supplementary Material online), noting that this may be related to the very limited regulatory potential of MER52 elements. Further analysis of endogenous KAP1 binding data shows a depletion of KAP1 at KZNF-bound gene promoters suggesting that KAP1 is not recruited at these loci (fig. 1C). Previous analyses published in Farmiloe et al. (2020) supported the endogenous binding of KZNFs at gene promoters in the absence of KAP1 suggesting endogenous binding of KZNFs at these loci. The fact that *ZNF519* is fixed in gibbons and great ape-lineages raises the possibility that *ZNF519* was co-opted for a TE-independent role in our genome.

High Similarity Between ZNF Binding Sites in TEs and Promoters

De novo motif discovery analysis of the *ZNF519* binding motif in highly bound MER52 elements and TE-independent promoters revealed high compatibility of *ZNF519* core binding motifs (fig. 2A). Mapping of *ZNF519* ChIP-seq data to a consensus MER52 sequence also showed the presence of the *ZNF519* transcription start site (TSS) motif at the summit of the MER52 binding peak (fig. 2B). A similar pattern was observed for *ZNF441* and *ZNF468*; de novo motif prediction for TE-independent TSS binding sites for both *ZNF468* and *ZNF441* closely matches the motif found in TE binding sites (fig. 2C and E). These promoter motifs can also be found at the summit of the *ZNF441* peaks mapped to the AluY consensus sequence and for *ZNF468* summits mapped to the MER11A consensus sequence (fig. 2D and F). Therefore, *ZNF519*, *ZNF441*, and *ZNF468* seem to be clear examples of KZNFs that were initially recruited to control transposable elements, but because the binding domain for each KZNF also recognized numerous gene promoters, they have become

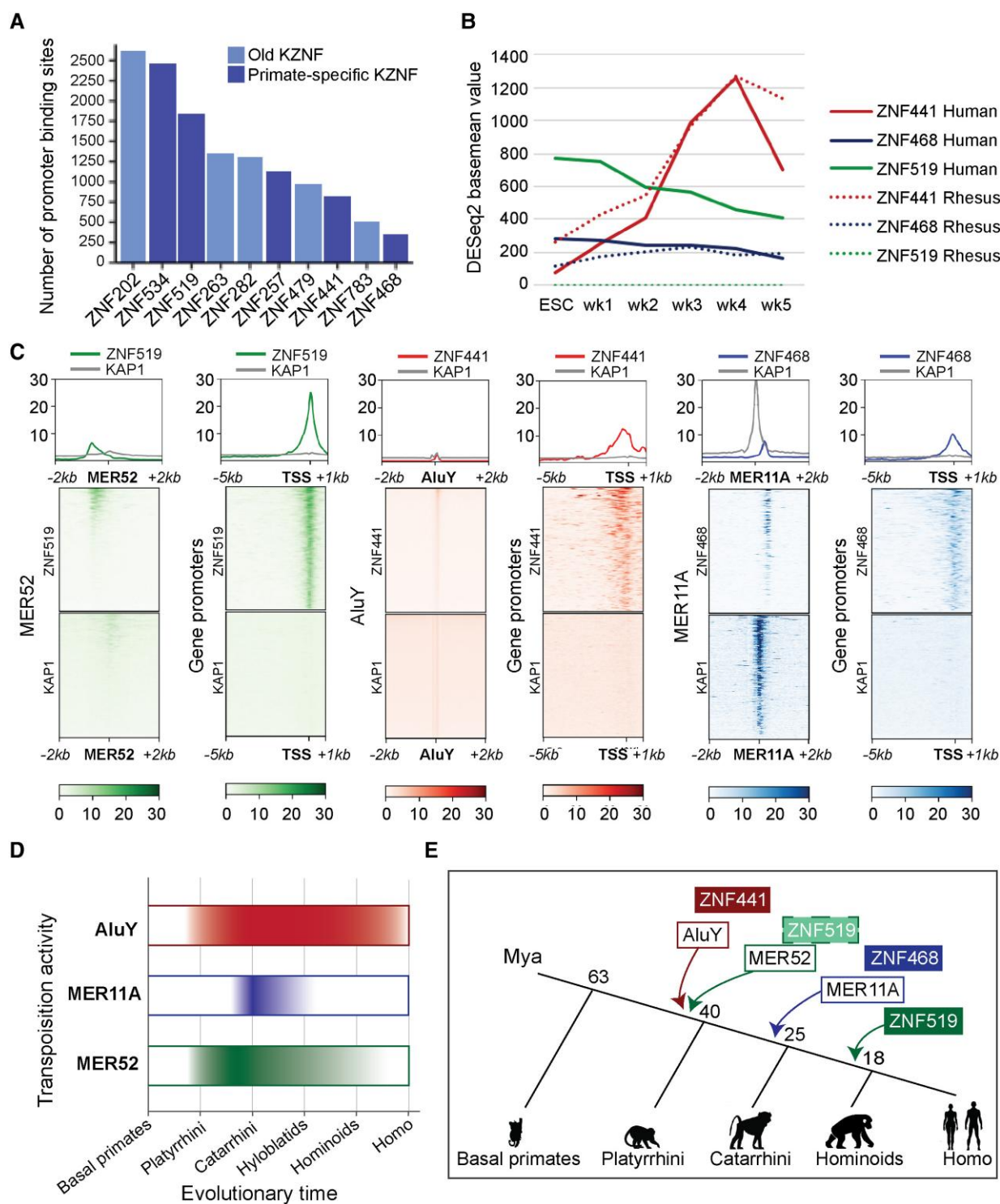


Fig. 1.—Co-evolution of three KZNFs with TEs is paralleled by co-option for direct gene-regulatory properties. **A**, The top ten promoter binding KZNFs by number of binding sites present in promoter regions (promoter defined as 5,000 bp upstream and 1,000 bp downstream of transcription start site). Adapted from Familoe et al. (2020). **B**, DESeq2 basemean expression values of KZNFs in human and rhesus embryonic stem cells (ESC) and weeks 1–5 wk1–wk5) of cortical organoid development from differential analysis of RNA-seq data. Data points show the average of two replicates for each species and time point. **C**, ChIP-Seq density plots showing ZNF and KAP1 binding at TEs and gene promoters (ZNF519—MER52; ZNF441—AluY, AluYa5; ZNF468—MER11A). **D**, The emergence and expansion of the TE classes associated with ZNF441, ZNF468, and ZNF519. Colored sections of the bars show approximately when the TEs emerged and density of color shows peak transposition activity **E**, evolutionary tree showing approximate time of emergence of TE classes and KZNFs.

Downloaded from https://academic.oup.com/gbe/article/15/11/evad184/7319543 by Universiteit van Amsterdam user on 16 February 2024

integrated into our gene-regulatory networks through the phenomenon of co-option.

In theory, it is possible that after the KZNFs became redundant for TE control, its binding specificity changed under pressure of some of the genes it evolved to regulate

Table 1

Promoter Binding KZNFs and the TE Families They Recognize

ZNF	# Promoters Bound	TE Family Recognized
ZNF519	1843	Mer52
ZNF441	816	AluY
ZNF468	3545	MER1A

as part of its co-opted function. This does not seem to be the case however: A multiple sequence alignment of the protein sequences shows very few differences in the zinc finger (ZNF) domains and contact residues of all three KZNFs between the oldest lineages and humans (supplementary figs. 5–7, Supplementary Material online). Despite these minor differences in their DNA binding domains, the predicted binding motifs of all three KZNFs are identical between the different orthologs (supplementary fig. 4, Supplementary Material online). The observed high level of conservation suggests that these KZNFs still serve an essential purpose in the human genome and may have been doing so since their emergence. In addition, the

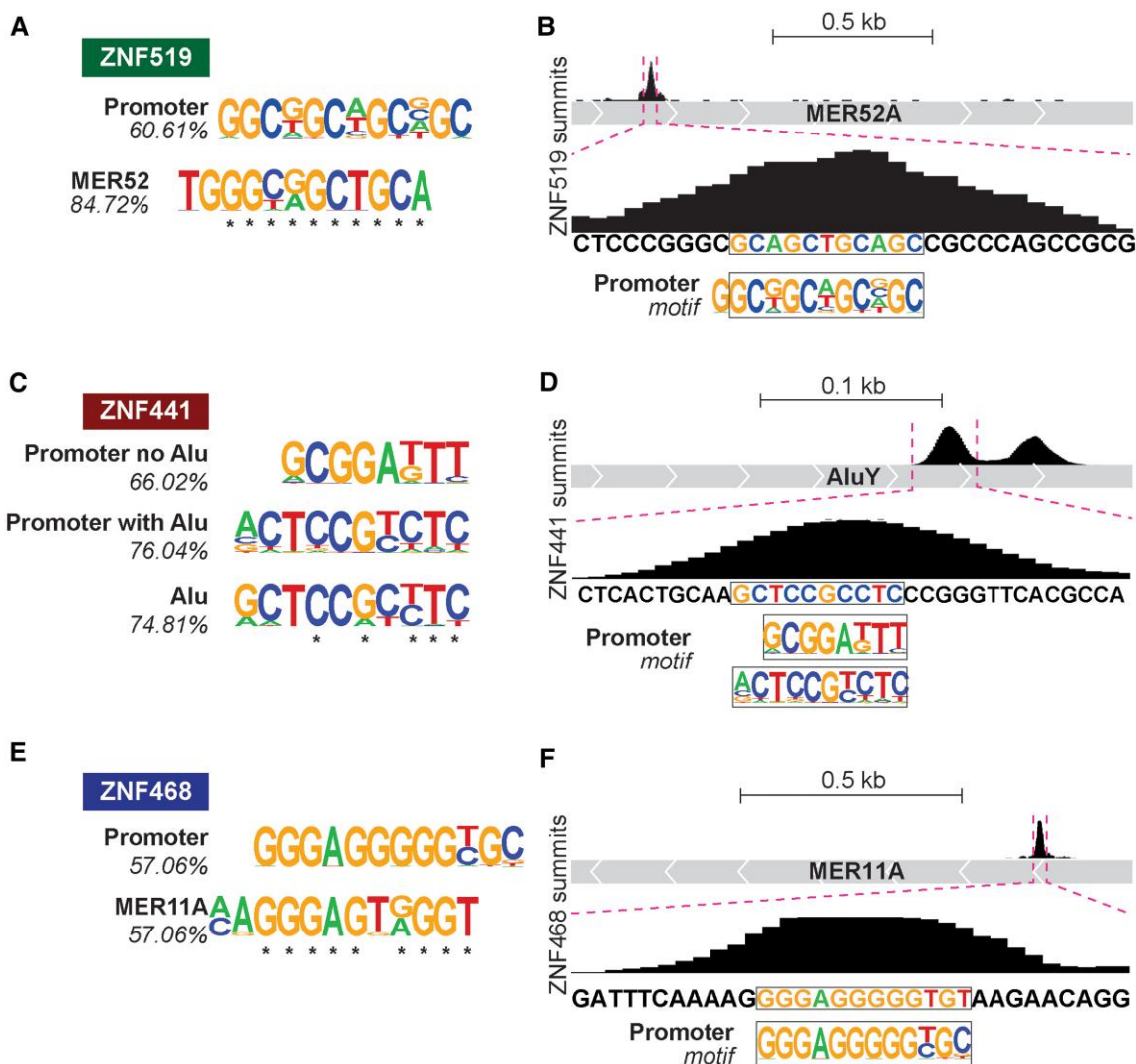


Fig. 2.—Transposable element motifs recognized by KZNFs are also seen in TSS binding sites. A, C, E, HOMER de novo motif discovery in DNA sequences ± 50 bp from the KZNF summits in TEs and gene promoters. (A) The ZNF519 summits in MER52 elements and promoter regions. (C) The ZNF441 summits in Alu elements and promoters with and without an Alu element. (E) The ZNF468 summits in MER11A elements and promoter regions. B, D, F, KZNF summits ± 7 bp lifted over to the UCSC repeat browser, shown at their recognized repeat family's consensus sequence show the presence of the KZNF-bound motif recognized in promoters for (B) ZNF519, (D) ZNF441, and (F) ZNF468.

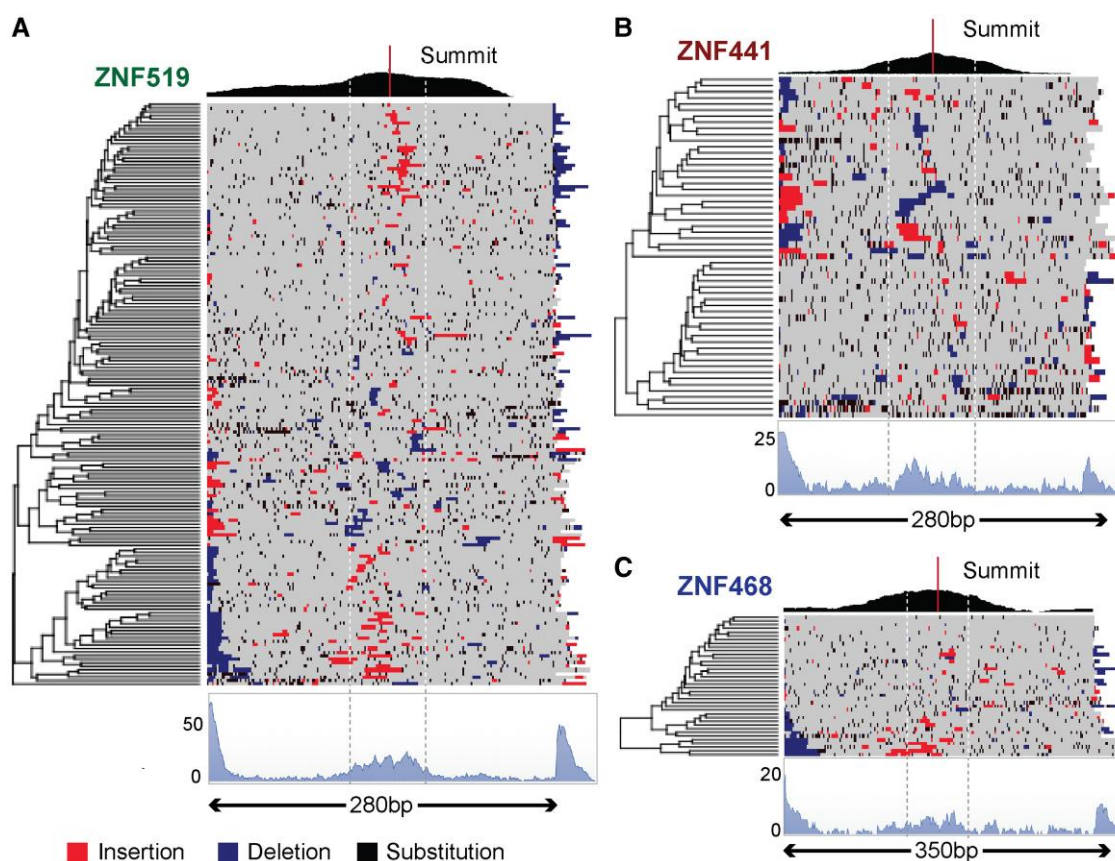


FIG. 3.—ZNF-bound promoters with hominid-specific insertions or deletions around the summit. Showing ZNF summits and adjacent bases containing indels specific to the hominid line after alignment of the human sequence with that of rhesus, green monkey, and marmoset. Composite values of all promoters are shown in the profile plots below each heatmap (A) ZNF519, (B) ZNF441, (C) ZNF468 (gray = no change, red = insertion in human, blue = deletion in human, black = substitution in human).

high occupancy of these three KZNFs on gene promoters suggests that any subsequent adaptive changes to restore the balance of gene expression must have come from modifications to the KZNF-bound promoters rather than the KZNFs themselves. We, therefore, expect that a genome that is confronted with a newly emerged KZNF that binds to a lot of gene promoters will show additional adaptations in these promoters to balance out the influence it has on gene regulation.

Accumulation of Indels and Substitutions at the KZNF Target Site in Gene Promoters

The consequences of a TE invasion are clearly not limited to the impact of TE insertions. Alongside the consequences of KZNFs which have an effect on gene regulation through TE-mediated insertions, the impact of KZNFs directly on gene promoters in the absence of TEs may be substantial as well. We investigated the presence of nucleotide substitutions and indels around the summits of the KZNF peaks in promoters. Because promoters often show high levels of

conservation, the presence of mutations at the site of KZNF binding that corresponds to the time point of emergence of the KZNF gene could be a signature of local evolutionary adaptations. To identify promoters with hominid-specific insertions or deletions, multiple sequence alignments of KZNF-bound promoter sequences were made between humans and two OWM species: Rhesus Monkey and Green Monkey. To distinguish between OWM-specific indels or hominoid-specific indels, the NWM species marmosets were taken along in the alignments as a outgroup. The regions analyzed centered around the summit of the KZNF peak, generated from ChIP-seq data, in the human genome (Imbeault et al. 2017), flanked by 140 bp upstream and downstream promoter sequence for *ZNF519* and *ZNF441* and 175 bp for *ZNF468*. These alignments were then coded for substitutions, insertions, deletions, and identical sequences, respectively. Only alignments that had complete sequence data for each species were included, which led to the exclusion of promoters with incomplete coverage in any one of the four species. Despite an overall high sequence similarity between the promoter sequences of each species, a large

number of promoters that remained after the screening process displayed a hominoid-specific insertion or deletion around the ZNF-bound promoter sequence while the adjacent sequence did not show this level of mutation (fig. 3). This was the case *ZNF519* in 164 out of 256 promoters (~64%) (fig. 3A), for *ZNF441* in 56 out of 86 (~65%) promoters (fig. 3B) and for *ZNF468* in 38 out of 59 (~64%) promoters (fig. 3C). Plotting of the relative coverage of indels along the promoters showed an increased likelihood for an indel to be present around the core binding site of the ZNF compared to the adjacent sequence (coverage graphs under each of the heatmaps, fig. 3). These data suggest that indels in ZNF-bound promoters are prevalent, and often occur at the core site of ZNF binding.

Promoter Regions Bound by KZNFs Show Higher Sequence Turnover at the Site of KZNF Binding

We further expanded our comparative analysis of KZNF-bound promoters by analyzing the levels of evolutionary sequence conservation using PhyloP 100 way conservation data available through the University of California Santa Cruz (UCSC) genome browser; 31 KZNFs that bind to >100 gene promoters were included in the analysis. For each individual KZNF, the average PhyloP value was taken at each base pair around summits in gene promoters. The regions analyzed centered around the ZNF-peak-summit and included 100 bp flanking sequences downstream and upstream. The values were then normalized and visualized in a heatmap for all of the KZNFs (fig. 4A). As a control, the analysis was repeated for the same sized-region 500 bp downstream from each of the KZNF summits (fig. 4A). The expectation is that because of the vicinity to the KZNF summit, these control regions are still in the vicinity of the TSS and still have overall high conservation.

Twenty-two out of the 31 KZNFs included in this analysis show a pattern of reduction in the PhyloP conservation value at the binding site of the promoter-bound KZNF compared to the flanking sequences up and downstream (fig. 4A). A lower conservation value is an indicator of increased sequence turnover. Whereas in the control region, 500 bp downstream of the KZNF-binding site the range of conservation values was very similar, the pattern was completely homogenous, indicating no local increase or decrease of conservation in the control region. For 10 KZNFs this pattern is particularly pronounced; the region approximately 30 bp up and downstream of the summit of the ZNF-binding site shows a clear reduction in conservation when compared to the surrounding region. Whereas the reduced conservation was clearest for *ZNF519* (fig. 4A and B) and some other KZNFs a more modest reduction of conservation values was observed for *ZNF441* (fig. 4A and C) and *ZNF468* (fig. 4A and D). To determine the number of promoters under the influence of positive (positive phyloP score) versus purifying (negative phyloP score)

selection, the bound promoters for *ZNF519*, *ZNF441*, and *ZNF468* were split based on their average score across the 200 bp (supplementary fig. 8A–C, Supplementary Material online). The summit-focused reduction in scores was maintained for both groups of promoters. These results support our initial findings from the indel analysis and show that the summits of the KZNF binding sites in gene promoter regions show an increased sequence turnover compared to the surrounding bases in the same gene promoters, or a control region downstream. Because this pattern is observed for most of the KZNFs that bind to gene promoters directly, the focal reduction in sequence conservation precisely at the site of KZNF binding seems to be a more general phenomenon that may point to local evolutionary adaptations as a direct response to KZNF binding.

To validate these findings, we repeated the analysis for a subset of known TFs using ChIP-Seq data from the Encyclopedia of DNA Elements (ENCODE) in either HEK293 or H1 cells (supplementary fig. 9, Supplementary Material online). In well-documented TFs we expect to see the opposite pattern to that observed in the KZNF data, with high levels of sequence conservation at the summits. Indeed for four out of seven TFs, *AFT2*, *EP300*, *REST*, and *CTCF* there is a strong peak showing very high conservation centered around the summit of the binding site (supplementary fig. 9, Supplementary Material online). Only one of the other TFs, *POLR2A*, shows a similar pattern to the KZNFs (supplementary fig. 9C and E, Supplementary Material online). *POLR2A* is known to bind to transcription start sites in gene promoters which show a high rate of turnover in primates (Taylor et al. 2006), this is reflected in the reduced conservation scores for *POLR2A* binding sites. It is unlikely that the turnover at KZNF sites is explained by turnover at *POLR2A* sites as only 0.02% of the KZNF summits ± 100 bp overlapped 50% or more with a *POLR2A* summit ± 100 bp (supplementary fig. 9F, Supplementary Material online). Taken together, the analysis of relative conservation values around the binding sites of KZNFs and known TFs supports our hypothesis that the emergence of KZNFs can exert selective pressure on the KZNF binding sites in gene promoters.

Genetic Deletion of KZNF Genes Reveals Widespread Effects on Gene Expression in Neuronal Tissues

We previously found a correlation in expression profiles of KZNFs and KZNF-bound gene promoters in the brain, suggesting KZNFs are widely integrated in neuronal gene-expression networks. Indeed, gene ontology analysis of genes bound by each of the KZNFs showed a significant enrichment of genes expressed in the brain for *ZNF519* and *ZNF468* (supplementary Tables S1 and S2, Supplementary Material online). To test this relationship further, we used Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR)/Cas9 to generate genetic

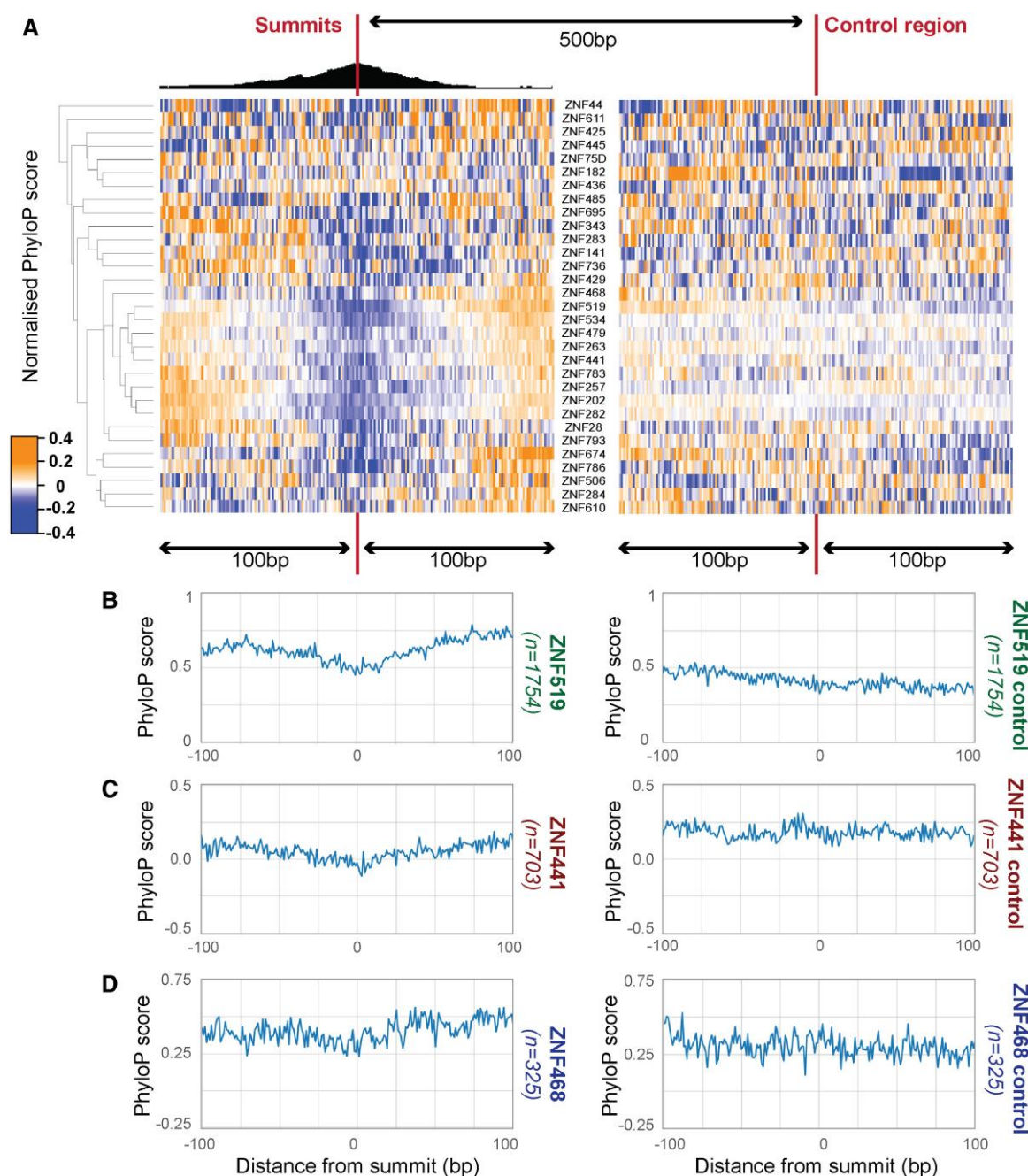


FIG. 4.—Collective conservation scores at KZNF summits in bound promoters. (A) Normalized PhyloP conservation scores averaged across all KZNF summits in gene promoter regions ± 100 bp and control regions 500 bp downstream of summits. blue = lower, orange = higher normalized PhyloP score than the average across summits for each KZNF. (B–D) Averaged, unnormalized PhyloP scores for ZNF519, ZNF441, and ZNF468 around the ChIP-seq summit and around control regions.

deletions of *either* ZNF519, ZNF441, and ZNF468 in hESCs. If the KZNFs have a regulatory effect on genes where they bind, we would expect to see changes in the expression of bound genes when the binding KZNF is removed from the genomic context.

To assess the impact of KZNF binding on gene expression in the brain, the three KZNF knockout (KO) hESC cell lines

were directed into a neuronal fate by generating cortical organoids. Correct genetic deletion and complete absence of expression for each KZNF were confirmed by RNA-seq in hESCs for ZNF519 and ZNF468. Confirmation of ZNF441 KO was done in D35 cortical organoids because of low ZNF441 expression levels in hESCs (fig. 5A, C and E). ZNF519 and ZNF441 cortical organoids were grown for

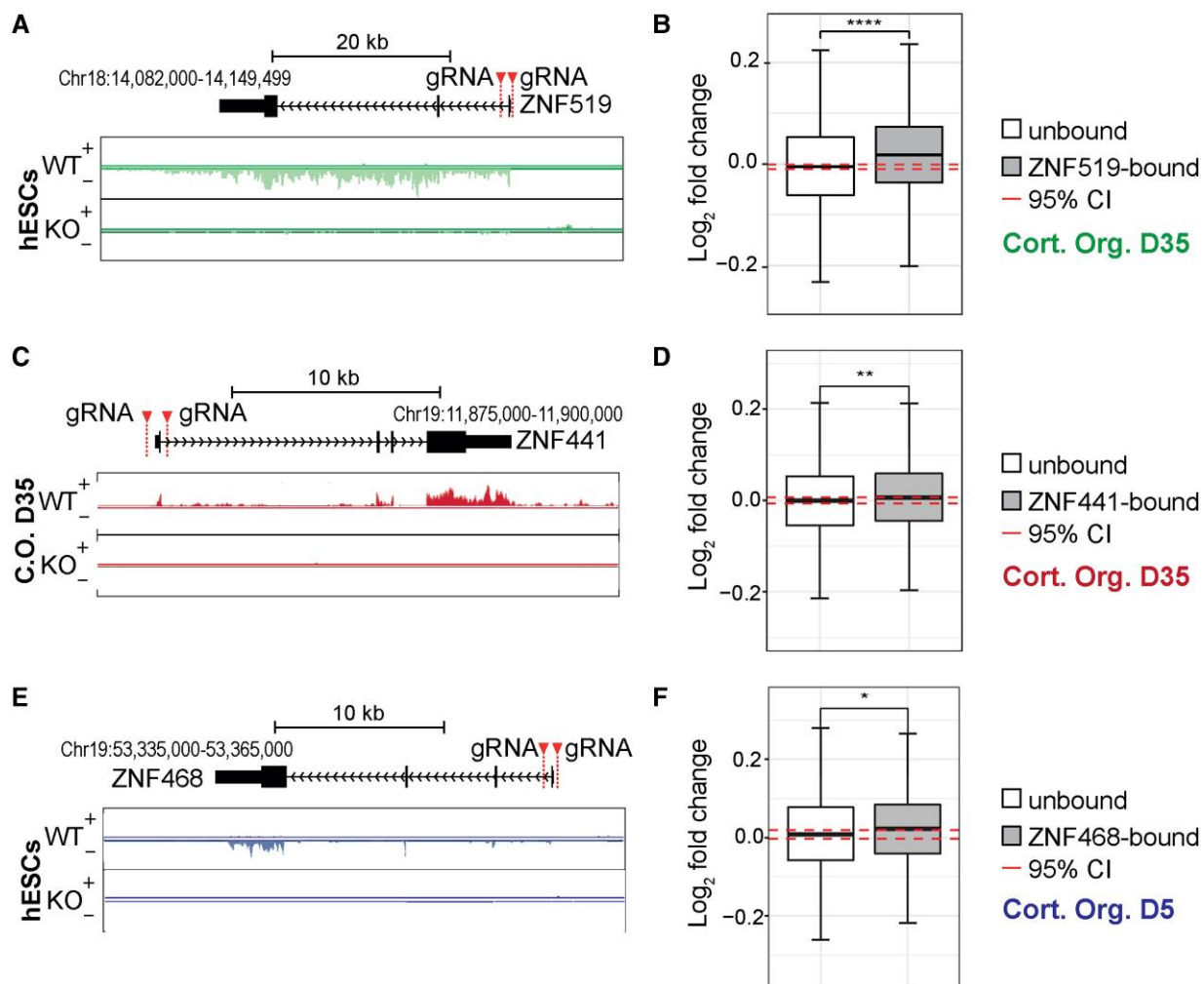


Fig. 5.—Knock out of KZNFs results in changes inbound gene expression. A, C, E, Overview of ZNF519, ZNF441 and ZNF468 loci, gRNAs used for CRISPR-Cas9 KO and RNA of WT and KO hESCs/Day 35 cortical organoids, scaling based on the number of mapped reads. B, D, F, Boxplot showing comparison of log₂ fold change of expressed (baseMean >10), high-confident KZNF-bound genes (gray) compared to unbound genes (white) after (B) ZNF519 KO in 5-week old cortical organoids (bound *n* = 2,311, unbound *n* = 10,598), (D) ZNF441 KO in 5-week old cortical organoids (bound *n* = 850, unbound *n* = 11,170) (F) ZNF468 KO in 5-day old cortical organoids (bound *n* = 397, unbound *n* = 11,345). **** = *P* < 0.0001, ** = *P* < 0.01, * = *P* < 0.05, Wilcoxon rank sum test with continuity correction. Red dashed lines were calculated independently and show the 95% CI of a 10,000 times bootstrapped median of a set of unbound genes with the same sample size as the target genes. Individual data points are not shown.

35 days before harvest. *ZNF468* KO cortical organoids began to display phenotypic changes compared to wild type (WT) at day 8 and were only grown for 5 days before harvest, a time point at which the *ZNF468*-KO and WT organoids were morphologically comparable.

We next assessed the collective change in expression of genes bound or not bound by each of the KZNF in WT and KO cortical organoids. Genes bound by *ZNF519* show a small but significant increase in expression compared to unbound genes in *ZNF519*-KO organoids (median log₂ fold change (log₂FC) D35: 0.018 versus -0.0055, *P* < 2.2e-16, Wilcoxon rank sum test with continuity correction) (fig. 5B). *ZNF441*-KO organoids also show increased collective expression of *ZNF441* bound genes compared to genes without

binding of *ZNF441* in their promoters (median log₂FC D35: 0.0066 vs. 0.00011, *P* < 0.01) (fig. 5D). Finally, a similar relative increase in expression was also observed for *ZNF468*-bound genes in *ZNF468*-KO organoids (median log₂FC D5: 0.023 vs. 0.0079, *P* < 0.05) (fig. 5F). The analysis was repeated with using a random set of genes the same size as the sets of KZNF-bound genes (supplementary figs. 10 and 11, Supplementary Material online), the results of this comparison were significant for ZNF519 and ZNF441 (*P* > 0.0001, *P* < 0.05) but not for ZNF468.

The differential expression analysis was repeated for the TE classes recognized by *ZNF519* (MER52), *ZNF441* (AluY), and *ZNF468* (MER11A). No significant differences were observed between the WT and KO at individual TEs. A

collective comparison of normalized read counts showed no significant changes in the expression of MER52 or MER11A elements in the *ZNF519* and *ZNF468* knockouts respectively (supplementary fig. 12A and C, Supplementary Material online). The slight increase in expression of AluY elements after *ZNF441* KO was significant (median normalized read count WT: 2.033, KO: 2.13, $P=0.0001459$, supplementary fig. 12B, Supplementary Material online). DESeq2 results and normalized read counts can be found in supplementary data files 1 and 2, Supplementary Material online.

In summary, cortical organoids derived from all three KZNF-KO hESC lines showed a modest collective increase of expression of their respective KZNF-bound target genes, supporting our hypothesis that KZNF-binding to gene promoters influences the expression of affected genes. Furthermore, the observed increase in expression of KZNF-bound genes in KZNF-KO organoids, shows that endogenous levels of *ZNF519*, *ZNF441*, and *ZNF468* affect their target genes under normal physiological conditions.

In-Depth Analysis of *ZNF519* KO and Overexpression Experiments Shows the Opposite Trend for Changes in the Regulation of Bound Genes

To further investigate the extent of the influence of KZNFs on gene expression we continued our analyses with *ZNF519* and examined the effects of *ZNF519* KO on gene expression in hESCs and day 14 cortical organoids. Similar to the observations in day 35 organoids, the collective expression of *ZNF519*-bound genes in day 14 cortical organoids is increased (median log₂FC D14: 0.012 vs. -0.0017 , $P=1.41e-11$) (fig. 6A). Conversely we see a reduction in the collective expression of *ZNF519*-bound genes in hESCs (median log₂FC of -0.026 vs. -0.0042 , $P=4.981e-06$) (fig. 6B). These data suggest a regulatory effect of *ZNF519* on gene expression which is active at certain developmental time points and in specific tissues only. Based on the gene networks that are expressed at these times and in these tissues the regulatory effect of *ZNF519* could be subtly different.

Finally, we analyzed the effect of *ZNF519* over expression (OE) on the expression of *ZNF519*-bound genes. (fig. 6E). Ectopic expression of *ZNF519* was verified using RNA-seq (fig. 6F). Overall, we see a collective decrease in the expression of *ZNF519*-bound genes when compared to unbound genes (median log₂FC OE: -0.029 vs. -0.0022 , $P<0.0001$) (fig. 6G). This observed change mirrors the increase in expression of *ZNF519*-bound genes we see in *ZNF519*-KO cortical organoids suggesting that the effect of *ZNF519* on bound genes is related to levels of *ZNF519* in cells. These results establish that at endogenous expression levels, the promoter-bound KZNFs subject to this study are able to regulate gene expression directly by binding to gene promoters, independent of the TEs they initially evolved to recognize.

Discussion

Previous studies have established an arms-race model for some specific KZNFs and the TEs they recognize (Jacobs et al. 2014; Imbeault et al. 2017). Under this model, KZNFs could have come to bind gene promoters by chance as a side effect of the arms race with TEs. Indeed, for *ZNF519*, *ZNF441*, and *ZNF468* we found a high similarity between the binding sites recognized in TEs and gene promoters. Our data offers a view into a “transitional” state affecting gene-regulatory networks suggesting that these KZNFs are in the process of co-option for gene-regulatory functions that are unrelated to their capacity to recognize and bind TEs that have long lost their capability to retrotranspose. Data from the KO and overexpression experiments show that these KZNFs are capable of influencing gene expression and most likely exert a repressive effect upon bound genes. Our analyses show binding of the KZNFs to promoters in the absence of KAP1, this raises the question of the mechanism behind the gene-regulatory effect of the KZNFs we studied. Further study is needed to truly understand this process, however, it is possible that the physical presence of the KZNF binding at gene promoters interferes with other transcription factor activities and through this mechanism gene expression is regulated.

We further discovered a high turnover of sequence at the core ZNF-binding site in promoters that display generally high sequence conservation. The increased likelihood of insertions, deletions, and substitutions and the higher rate of sequence turnover at KZNF binding sites in gene promoters suggests that KZNFs could be driving changes in DNA sequence at promoter-binding sites. Our analysis of this process is limited in that we are only able to study promoters that are still bound by KZNFs, we do not have information on promoters that may have changed so much that they are no longer recognized by the KZNFs. Information on previously bound promoters would give us more insight into this process, however, based on our observations, we propose a model of “promoter adaptation” where a newly emerged KZNF is able to recognize motifs in gene promoters as well as TEs. The subsequent repressive effect of the KZNF on these promoters exerts selective pressure on the promoters to modulate the consequential gene-regulatory influence of the KZNF. This becomes evident by a higher rate of sequence turnover at the site of KZNF binding while the promoter adapts to the new KZNF-mediated regulatory influence (fig. 7). The collection of genomic adaptations to regain a new gene-regulatory balance, drives a wave of subtle changes in gene expression at many genetic loci simultaneously. At first sight, the effect of any particular KZNF on any particular promoter may seem noticeable but modest in absolute terms. However, it’s important to consider that our analysis shows a modest change in expression for a large set of KZNF-bound genes, meaning that the expression of many genes has been affected in a modest way. Our

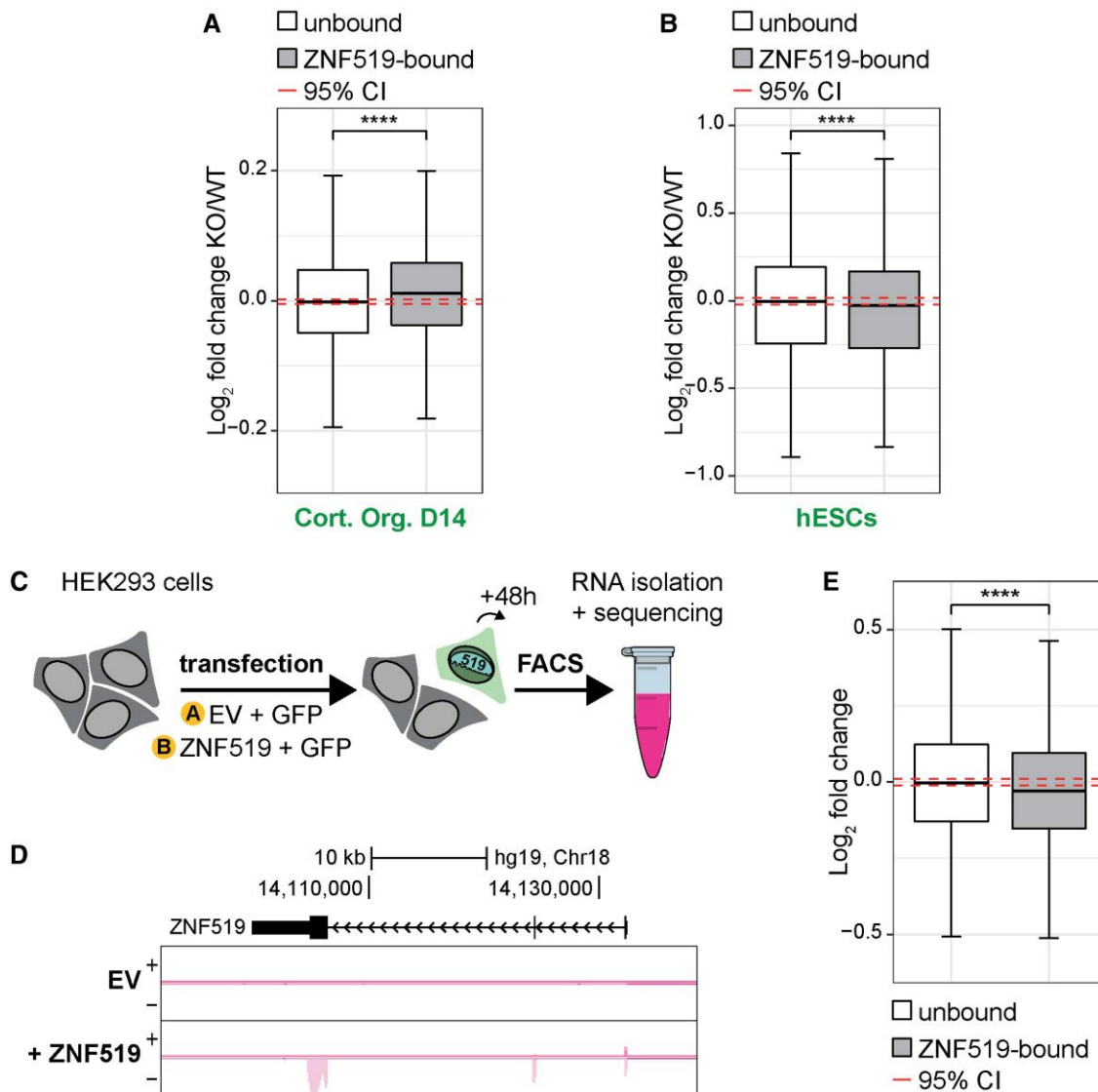


Fig. 6.—In-depth analysis of ZNF519 KO and overexpression experiments shows reciprocal changes in the regulation of bound genes. (A) Boxplot showing comparison of \log_2 fold change of expressed (baseMean >10) high-confident KZNF-bound genes (gray) compared to unbound genes (white) in ZNF519 KO cortical organoids of 2 weeks old (day 14, bound $n = 2,293$, unbound $n = 10,926$), (B) ZNF519 KO hESCs (bound $n = 2,282$, unbound $n = 9,937$). **** = $P < 0.0001$, Wilcoxon rank sum test with continuity correction. Red dashed lines were calculated independently and show the 95% CI of 10,000 times bootstrapped median of a set of unbound genes with the same sample size as the target genes. Individual data points are not shown. Mind difference in y-axis. (C) Schematic showing ZNF519 overexpression experiment set up in HEK293 cells. (D) RNA-seq at ZNF519 locus confirms overexpression in HEK293 cells. Mean of three replicates shown, scaled on number of mapped reads (excluding ZNF519 locus). (E) Boxplot showing a comparison of \log_2 fold change of expressed (baseMean >10) high-confident ZNF519-bound genes (gray, $n = 2,268$) compared to unbound genes after overexpression of ZNF519 (white, $n = 8,222$) after ZNF519 overexpression, **** = $P < 0.0001$.

analysis of these three KZNFs gives us a unique insight into evolutionary processes that are ongoing in the human and primate genomes. If this process were also to hold true for the other ~160 primate-specific KZNFs, their contribution to the evolution of gene expression in humans and primates should not be underestimated. Although evolutionary time runs slowly we highlight the importance of recognizing the genome as an entity that is constantly in flux. Our data

suggests that KZNFs are able to contribute to this change and could also be drivers behind the features that differentiate primates and humans from other species.

Our study emphasizes that the invasion of a genome by a new class of TEs has an even more far-reaching and long-lasting impact on a species' gene-regulatory network than was previously considered, and it happens on clearly distinguishable regulatory levels: First by the addition of new

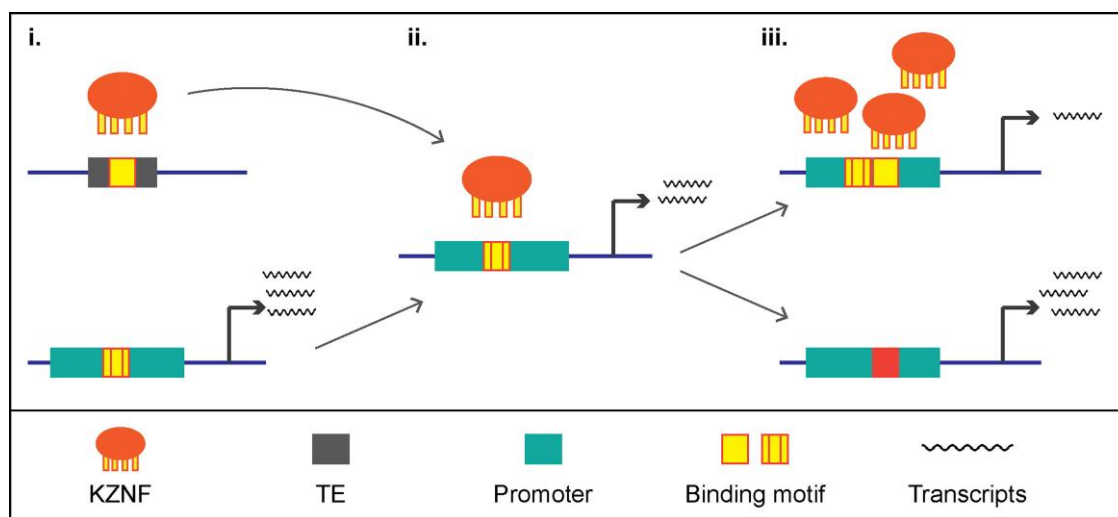


Fig. 7.—Promoter escape model of evolution at promoter KZNF binding sites. Promoter escape model of evolution at promoter KZNF binding sites: 1) Binding motif present in TE and recognized by KZNF is also present in the gene promoter region. 2) KZNF also recognizes promoter motifs and binds there, affecting gene expression. 3) The effect of KZNF binding on gene expression exerts a selective pressure at the locus, either leading to a loss of binding site or strengthened binding by the KZNF.

TE- and KZNF-TE-mediated regulatory functions; second by the evolution of KZNF-mediated control over gene promoters, and lastly by the increase sequence turnover in gene promoters at the sites of KZNF binding. Whereas it's impossible to know the sequence of events that led KZNF to bind both TEs and promoters, the most likely scenario is that most promoter-binding KZNFs evolved to bind gene promoters at the same time as the TEs, simply because the targeted TE sequence has similarity to some gene promoters. A careful dissection of the evolutionary histories of both the TEs, promoters, and the KZNFs could sometimes give clues about that: Whereas such analysis did not yield clues for the KZNFs in this study, the evolutionary history was analyzed in much detail for ZNF91/SINE-VNTR-Alu (SVA) and ZNF93/LINE1 (Jacobs et al. 2014) revealing that both KZNF gene ZNF91 and ZNF93 were around quite some time before their structural changes allowed it to recognize and repress the newly invading SVA and L1PA classes, respectively. In short, earlier forms of the KZNF were not able to repress the TE, but later versions were. This shows that with regards to KZNFs and TEs, the KZNFs were present earlier than the TEs but adopted their structure quickly once the TE invasion started. Keeping in mind that KZNF binding properties evolved as a result of TE invasions, we believe that the most likely scenario is that most KZNFs evolved to bind gene promoters at the same time as the TEs. A less likely scenario, in our view, is that KZNFs evolved to bind gene promoters before TEs because this could have happened straight after the KZNFs first emerged in primate genomes, which often preceded the emergence of TEs. We do not intend to claim that this is the case for every KZNF in our genome and some KZNFs are likely to have evolved gene-regulatory properties

independent from TEs, for which some of the older KZNFs (without clear TE targets as identified by Imbeault et al. 2017) may be good examples for that.

Altogether, our study reveals the multilayered impact of the emergence of primate-specific KZNFs on human neuronal gene expression patterns and raises the exciting question of how the combination of these regulatory effects is influencing the evolution of species-specific gene-regulatory networks.

Materials and Methods

Human and Rhesus RNA-Seq Data

Basemean values from DESeq2 differential expression analysis data published by (Field et al. 2019) were used for analyses between human and Rhesus cortical organoid development. The data plotted is the average of two replicates.

Analysis of Published ChIP-Seq Data

The ChIP-exo data for this analysis was generated by Imbeault et al. (2017) (NCBI gene expression omnibus (GEO) database accession number GSE78099). Using Trimmomatic (Galaxy v0.36.5) (Bolger et al. 2014), the reads were processed and adaptor and illumina-specific sequences were removed. The reads were mapped to the human genome (assembly GRCh37/hg19; (Lander et al. 2001) using Bowtie2 (Galaxy v2.3.4.2) (Langmead and Salzberg 2012) with single-end, very sensitive end-to-end settings. Summits were generated using Model-based Analysis of ChIP-seq 2 (MACS2) (Zhang et al. 2008) with default settings and intersected with the list of peaks in gene

promoters generated in Farmiloe et al. to produce a final list of KZNF summits in gene promoter regions. The TE families most recognized by each KZNF were taken from published data (Imbeault et al. 2017). TE family locations were extracted from the UCSC genome browser repeatmasker track for MER11A, MER52 elements and Alu subclasses AluY and AluYa5. ChIP density plots were generated using computeMatrix (Galaxy Version 3.1.2.0.0) and plotHeatmap (Galaxy Version 3.1.2.0.1) (Ramírez et al. 2016).

KZNF Evolutionary History and Binding Motif Analysis

The human exonic sequences from each KZNF transcript of interest (*ZNF519*: ENSG00000175322, *ZNF441*: ENSG00000197044, *ZNF468*: ENSG00000204604) were extracted from the UCSC genome browser hg38 using the table browser tool (Kent 2002; Karolchik et al. 2004) <http://genome.ucsc.edu>. The equivalent primate sequences from panTro5, panPan2, gorGor5, ponAbe3, nomLeu3, rheMac10, macFas5, papAnu4, calJac3, and saiBol1 were retrieved using the UCSC BLAST-like alignment (BLAT) tool (<https://genome.ucsc.edu/cgi-bin/hgBlat>) (Kent 2002). This comparison was confirmed by lifting over the coordinates of the sequences returned in the BLAT search from primate to human using the UCSC liftOver tool. If all the exons for each gene were found in proximity in the primate genome this was taken to be a paralogous gene and the sequence was extracted. The genomic sequences were then converted to peptide sequences using the ExPASy translate tool (Gasteiger et al. 2003) and a multiple sequence alignment was performed on these sequences using ClustalW (1.2.4) (Sievers et al. 2011). Predicted zinc finger domains and motifs were generated using the Persikov tool (<http://zf.princeton.edu/>, Persikov and Singh 2014).

TE Evolutionary Analysis

The TE families associated with each of the three KZNFs were identified from previously published data (Imbeault et al. 2017). Coordinates for the loci of all the TEs from each family recognized by a candidate KZNF (MER51, MER11A, AluY) were downloaded from the RepeatMasker track on the UCSC genome browser for the human and other primate genomes. The human coordinates were then converted to each of the primate genomes using the UCSC liftOver tool to find the number of homologous loci in other genomes. The minimum amount of bases set to remap was set to 0.95. The analysis was done pairwise between the human genome and each primate genome available. The flanking region, 1000 bp immediately downstream of the TE coordinates was also generated and lifted over from human to each of the primate genomes and this number was used to correct for discrepancies in annotation accuracy. Charts were made using the dotplot function of the ggplot2 package in R (Wickham 2016). Species were grouped as follows:

Great apes: Chimpanzee (panTro5), Bonobo (panPan2), Gorilla (gorGor5), Orangutan (ponAbe3)

Apes: nomLeu3

Old-world monkeys: Rhesus macaque (rheMac10), Crab eating macaque (macFas5), Baboon (papAnu4), Green monkey (chlSab2), Golden snub-nosed monkey (rhiRox1), Proboscis monkey (nasLar1)

New-world monkeys: Marmoset (calJac3), Squirrel monkey (saiBol1)

Basal primates: Tarsier (tarSyr2), Bushbaby (otoGar3), Mouse lemur (micMur1)

KZNF Expression GTEX

The tissue expression graphs for *ZNF519*, *ZNF441*, and *ZNF468* were generated on the Genotype-Tissue Expression (GTEx) Portal (<https://gtexportal.org/home/>) using GTEx Analysis Release V8 (dbGaP Accession phs000424.v8.p2).

KZNF De Novo Motif Analysis

De novo motif analysis for *ZNF519*, *ZNF441*, and *ZNF468* in highly bound MER52s, MER11As, Alus, and TSSs (coverage > 100 on the UCSC genome browser) was performed with Hypergeometric Optimization of Motif Enrichment (HOMER; Heinz et al. 2010). As input, a bed file was used with locations of the KZNF-binding summit plus/minus 50 bp.

KZNF Binding at Repeat Elements

ZNF519, *ZNF441*, and *ZNF468* summits were extended with 7 bp on each side and subsequently a lift over to the hg38 repeat browser was performed using the liftOver tool from the UCSC genome browser (<http://hgdownload.soe.ucsc.edu/admin/exe/>) using the hg19_to_hg38reps.over.chain as provided by the UCSC repeat browser (<https://repeatbrowser.ucsc.edu/>). The file was sorted using bedSort and a coverage track was generated using bedtools genomecov (-bg -split) followed and the bedGraphToBigWig tool to specifically visualize the summit of *ZNF519*, *ZNF441*, and *ZNF467* binding at the consensus MER52A, AluY, and MER11A, respectively.

Cloning of MER52 Elements into Luciferase Reporter Plasmid

A MER52D with a peak height > 100 reads was selected and amplified by polymerase chain reaction (PCR) using flanking primers (chr10:56667794-56670008, assembly GRCh37, forward 5'-3': CCATACCCTATGAAAGCTGGTC, reverse 5'-3': GGGAGATTGTACCTTGATGAC) with LongAmp® Taq DNA Polymerase (NEB) with an annealing temperature of 57 °C. Amplicons were purified using the QIAquick Gel Extraction Kit (QIAGEN), and cleaned with the DNA Clean & Concentrator™-5 Kit (ZYMO Research). Blunting of 3' ends was done using DNA Polymerase I, Large (Klenow)

Fragment (NEB) before phosphorylation of 5' ends using T4 Polynucleotide Kinase (NEB). Inserts were ligated upstream of the luciferase reporter in the pGL4.12[luc2CP]SV40 plasmid that was digested using EcoRV (Thermo Scientific™), purified using the QIAGEN, cleaned with the DNA Clean & Concentrator™-5 Kit (ZYMO Research) and treated with Shrimp Alkaline Phosphatase (rSAP) (NEB) to remove 5'- and 3'- phosphates. Inserts and vectors were concentrated using the DNA Clean & Concentrator™-5 Kit (ZYMO Research) and ligated using the Quick Ligation™ kit (NEB). Cloning resulted in one orientation of the MER52, and, therefore, another PCR was performed on generated plasmids using primers with restriction site for size-oriented cloning (forward 5'-3': AGATG AGCTCCCATACCTATGAAAGCTGGTC, reverse 5'-3': CCTCAGATCTGGGAGATTGTACCTTGATGAC). Amplicons were cleaned using the DNA Clean & Concentrator™-5 Kit (ZYMO Research) and, together with the pGL4.12[luc2CP]SV40 plasmid, digested using corresponding restriction enzymes (SacI and BglII, Thermo Scientific™). Inserts and plasmid were purified, cleaned, and ligated as described above. Sanger sequencing was performed to confirm correct cloning.

mESC Cell Culture and Transfection and Luciferase Assay

The mouse embryonic stem cell and luciferase assay protocols followed were published in van Bree et al. (2022).

Insertion and Deletion Analysis at Gene Promoter KZNF Binding Summits

The DNA sequence the size of average peak of each KZNF was extracted from the UCSC reference assemblies around the *ZNF519*, *ZNF441*, and *ZNF468* summits for human, rhesus, green monkey, and marmoset. Alignments were made between the human, rhesus, green monkey and marmoset using the European Molecular Biology Open Software Suite (EMBOSS):6.6.0.0 stretcher tool (Rice et al. 2000) to assess mutations specific to the human lineage. Alignments with indels specific to the human lineage were confirmed with a multiple sequence alignment using ClustalW on the ebi online portal (<https://www.ebi.ac.uk/Tools/msa/clustalo/>) (Sievers et al. 2011). The remaining human alignments were coded 1=no difference, 2=insertion in human sequence, 3=substitution, and 4=deletion in human sequence. The coded files were then visualized in R using the heatmap2 function from the gplots package using the default clustering method (Warnes et al. 2016).

KZNF Summit Conservation Score Analysis

PhyloP conservation data was accessed from UCSC genome browser and intersected with KZNF summit coordinates

±100 bp and transcription factor summits ±100 bp using Python (see [supplementary script, Supplementary Material](#) online). The transcription factor summits used for this analysis were taken from ENCODE (ENCODE Project Consortium 2004) using the hg19 narrow peaks files and selecting all peaks with a signal value greater than 100, in cases where the average signal value was lower, the top 2,500 peaks were selected. This analysis was repeated for control regions 500 bp downstream of all summits ±100 bp. The data was then compiled into tables for each KZNF, transcription factor, and control region from which averages at each base location relative to the summit were calculated across all binding sites for each KZNF. These averages were visualized in a heatmap made using Python (see [supplementary script, Supplementary Material](#) online).

Guide RNA Design and Cloning

gRNAs to KO exon 1 of *ZNF519* (upstream: ATGCTAAAAAATGACCCCT; downstream: ATGTATGGGAGCGTAGAAGT), exon 1 of *ZNF441* (upstream: AAATCAGGGGATAGCTCCAC; downstream: CTGGTGTCCCACGC GTGAG), and exon 1 of *ZNF468* (upstream: GCCCTTCTGGGCGGAACGT; downstream: AAACCCTCTTGATCGTGT) were designed using Benchling (Biology Software), with masked regions included and protospacer adjacent motif (PAM) sequence set to NGG in the search. CHOPCHOP (Labun et al. 2016), CRISPR design (<http://crispr.mit.edu/>), and the BLAT tool of the UCSC Genome Browser (Kent 2002) were used for verification of the efficiency and specificity of the gRNAs. Complement gRNA oligos were ordered with CACCG added at the 5' end while reverse-complement oligos contained AAAC at the 5' and C at the 3' end, to facilitate cloning. Equal amounts of complement and reverse-complement gRNA oligos were combined in annealing buffer (10 mM Tris pH7.5–8.0, 50 mM NaCl, 1 mM Ethylenediaminetetraacetic acid (EDTA)) and annealed using a thermocycler. The pX330-U6-Chimeric_BB-CBh-hSpCas9 plasmid (Addgene) was digested using BbsI (Thermo Scientific™) and cleaned using the DNA Clean & Concentrator™-5 Kit (ZYMO Research). gRNAs were ligated into the pX330 plasmid using the Quick Ligation™ kit (NEB).

KO of KZNFs in hESCs

KO of *ZNF519*, *ZNF441*, and *ZNF468* in hESCs was performed as described previously by our lab (Haring et al. 2021). *ZNF441* and *ZNF468* KO and WT lines were transfected with pX330 plasmid+guide RNAs. WT clones were selected from clones transfected with the complete CRISPR construct which did not show a deletion as in Haring et al. (2021). *ZNF519* WT hESCs, cells were

transfected using the pX330 plasmid without guide RNAs. Initial genotyping of hESCs was performed similarly as described above according to the protocol of Hendriks et al. (2015). Three KO and three WT hESC clones were analyzed by RNA sequencing for *ZNF519* and *ZNF468*. Due to the low expression of *ZNF441* in hESCs, this KO was validated in the cortical organoids.

Primers used for the genotyping were:

ZNF519 F: GCCTAATAAGGGCGTTTGTG; R: GAAATACAA
AAAAAAGAGGTGTTCT
ZNF441 F: CCAGACTGGTCTCGAATTCT; R: GCAGAAG
AATGCGGTTTCT;
ZNF468 F: CCTTCGTCGCAAAGATGCA; R: GGATGTCTCT
GAAGCTGAGCACT

Cortical Organoid Culture

Cortical organoids were grown from WT or KO hESCs based on the methods of Eiraku et al. (2008). In short, hESC colonies were grown in a 10 cm dish on mitomycin C-treated mouse embryonic fibroblasts (MEFs, Global Stem) in hESC medium (Dulbecco's Modified Eagle Medium, DMEM-F12 (Gibco) supplemented with 20% KO Serum Replacement (Gibco), 100 U/ml penicillin/100 µg/ml streptomycin (Gibco), 2 mM GlutaMAX (Gibco), 1× MEM Non-Essential Amino Acids solution (Gibco), 100 µM 2-mercaptoethanol (Gibco)) with 8 ng/ml fresh basic fibroblast growth factor (bFGF, Sigma). The medium was changed to differentiation medium (99% hESC medium supplemented with 1 mM sodium pyruvate (Gibco)) and colonies with a diameter of 1–2 mm were detached from the plate using a cell lifter (Corning) adjusted to approximately 2–3 mm width. Lifted colonies were collected in 5 ml medium, transferred to a 60 mm ultra-low attachment dish (Corning), and embryoid bodies were formed overnight at 37 °C, 5% CO₂. The next day, medium was refreshed for differentiation medium with freshly added 3 µM IWR-1-Endo, 1 µM Dorsomorphin, 10 µM SB-431542 hydrate, and 1 µM Cyclopamine hydrate (day 0 of differentiation), which was repeated every other day. Organoids were placed on a rocker on days 3–4, to enhance growth and prevent the merging of organoids. From day 18 onward, Neurobasal/N2 medium (Neurobasal (Gibco) supplemented with 100 U/ml penicillin/100 µg/ml streptomycin (Gibco), 2 mM GlutaMAX (Gibco), 1× N-2 supplement (Gibco)) supplemented with 1 µM Cyclopamine hydrate was used for growth or organoids. From D24, no inhibitors were added anymore to the medium until organoids were harvested at day 35 and day 5. For organoid formation, one KZNF-KO line was used for each KZNF; For the *ZNF441* and *ZNF468* organoids, two replicates (each replicate containing >10 organoids) were taken from two independent batches of organoids made at different times for a total of four replicates for each condition. For the *ZNF519*

organoids, three replicates (each replicate containing >10 organoids) were taken from a single batch of organoids for each condition.

Overexpression *ZNF519* HEK293 Cells

HEK293 (ATCC) cells were grown in DMEM, high glucose, GlutaMAX™, supplemented with 10% heat inactivated fetal bovine serum (HIFBS Gibco™), and 100 U/ml pen/strep (Gibco™). 24 h before transfection, cells were plated on a 60 mm dish at a density of 50.000 cells/cm² to ensure 70–90% confluency at the time of transfection. Transfection was performed in a complete growth medium without pen/strep using Polyethylenimine (PEI) (Polysciences) for six hours with 6.4 ng pCAGEN-*ZNF519* (Jacobs et al. 2014) or pCAGEN-empty vector (Addgene #11150), combined with 335.6 ng pCAGEN-GFP (green fluorescent protein). Transfected cells were grown in a complete growth medium for 48 h, with medium refreshment after 24 h.

Before Fluorescence-activated cell sorting (FACS) sorting, cells were washed with warm phosphate buffered saline (PBS) and harvested by incubation for five minutes at 37 °C, 5% CO₂ in 0.25% Trypsin, 0.5 mM EDTA. Trypsinization was deactivated by addition of a complete growth medium, after which cells were pelleted and resuspended in 500 µl FACS buffer (PBS supplemented with 3% HIFBS, 0.5 mM EDTA). 300.000 GFP+ cells were sorted in resuspension buffer (0.5 mM EDTA in PBS), and centrifuged. Supernatant was removed and samples were ready for RNA isolation.

RNA Isolation and Sequencing

For *ZNF519* overexpression and *ZNF519*, *ZNF441*, and *ZNF468* KO experiments, RNA was isolated in 400 µl TRIzol Reagent (Invitrogen™) according to manufacturer's recommendations. Potential DNA contamination was removed with DNaseI (Roche) and samples were cleaned using the DNA Clean & Concentrator™-5 Kit (ZYMO Research). Libraries were prepared with the TruSeq Stranded Total RNA (Illumina) with Ribo-Zero ribosomal RNA depletion, and sequenced paired-end, 75 bp on a NextSeq 550 system (Illumina) by molecular analysis department (MAD): Dutch Genomics Service & Support Provider (Swammerdam Institute for Life Sciences, Amsterdam).

RNA-Seq Data Analysis

The public Freiburg Galaxy server (Goecks et al. 2010) (use-galaxy.eu) (Afgan et al. 2018) was used for processing data. Adapters were removed and reads were trimmed with trimomatic (Bolger et al. 2014) version 0.36.5 for paired-end reads (ILLUMINACLIP TruSeq3 paired-end), cutting if the average per base quality in a four-base sliding window was below 20, dropping reads below 30 bases. Mapping of reads was performed with HISAT2 (Kim et al. 2019)

(Galaxy Version 2.1.0 + galaxy3) against the built-in reference hg19 Full genome. Reads were assigned to gencode V19 and rmsk features using featureCounts (Liao et al. 2014) (Galaxy Version 1.6.3) with `-p, -d 75 -D 900 -B -C`. Output was analyzed using DESeq2 (Love et al. 2014) (Galaxy Version 2.11.40.3) with default settings. Coverage tracks were generated using bamCoverage (Galaxy Version 3.0.2.0, with deepTools2 (Version 3.0.2) and samtools (Version 1.7)) from the deeptools2 package (Ramírez et al. 2016) (bin size 1) and scaled on UCSC with a scaling factor based on the number of uniquely assigned reads from featureCounts, or scaled using bamCoverage for merging of replicates (HEK293). Scaled coverage tracks were merged using wiggletools (Zerbino et al. 2014) mean, and wig files were transformed into bigwig files using the wigToBigWig script (<http://hgdownload.soe.ucsc.edu/admin/exe/>). DESeq2 output and read counts available in [supplementary Data Files 1 and 2, Supplementary Material](#) online.

KZNF Target Analysis OE and KO

Peak data provided by Imbeault et al. (2017) was extracted from NCBI (GSM2466578). Peaks with a MACS score >500 were taken as “high-confident ZNF519-bound” regions while all peaks provided by Imbeault et al. (2017) were taken as “ZNF519-bound” regions. TSSs were generated from gencodeV19 annotation by selecting the first bp of protein-coding is known transcripts. The center of the MACS peaks was calculated to perform an overlap where at least 50% of the KZNF peak is in the promoter region. Using bedtools ClosestBed, the 50 closest TSSs to the MACS peaks were selected and reported with the distance of the center of the MACS peak to the TSSs. These were subsequently filtered to only keep those peaks that overlap with the promoter regions of TSSs as defined by a window of 5,000 bp upstream and 1,000 bp downstream of the TSS.

Log₂FC of bound and unbound genes was compared using R (R Core Team 2019) and visualized using ggplot2 (Wickham 2016). The number of bound and unbound genes expressed differed greatly and so the 95% confidence interval (CI) of the median log₂FC of unbound genes was calculated by 10,000 times bootstrapping the median of a random set of unbound genes with a similar sample size as the bound genes. For comparison of log₂ fold change of expressed target genes versus nontarget genes, a Wilcoxon rank sum test with continuity correction was performed.

Supplementary Material

Supplementary data are available at *Genome Biology and Evolution* online (<http://www.gbe.oxfordjournals.org/>).

Acknowledgment

This work was supported by a European Research Council (ERC) starting grant (ERC-2016-stG-716035) to F.M.J.J. Many thanks to Gonzalo Congrains Sotomayor for FACS expertise and technical assistance; Alfredo Huaman for his help with primer design; Wim de Leeuw for his time and technical support for bioinformatics; Selina van Leeuwen and the MAD: Dutch Genomics Service & Support Provider of the University of Amsterdam for sequencing; and the many helpful discussions with Evolutionary Neurogenomics Group and others at the Swammerdam Institute for Life Sciences (SILS).

Author Contributions

Conceptualization, F.M.J.J., E.J.v.B., and G.F.; Methodology, F.M.J.J., E.J.v.B., G.F.; Investigation & Validation, G.F., E.J.v.B., S.R., L.J.M.J., L.M.; Data Curation, F.M.J.J., E.J.v.B., G.F.; Writing—Original Draft, G.F.; Writing—Review & Editing; F.M.J.J., E.J.v.B., G.F.; Visualization, G.F., E.J.v.B., S.R.; Supervision, F.M.J.J.; Project Administration, F.M.J.J.; Funding Acquisition, F.M.J.J.

Data Availability

Raw sequence reads and processed gene expression values from this study have been submitted to the NCBI Sequence Read Archive (SRA; <http://www.ncbi.nlm.nih.gov/sra>) under accession number PRJNA830860 and to the NCBI Gene Expression Omnibus (GEO; <http://www.ncbi.nlm.nih.gov/geo>) under accession number GSE201315, respectively.

Literature Cited

- Afgan E, et al. 2018. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. *Nucleic Acids Res.* 46:W537–W544.
- Bennett EA, et al. 2008. Active Alu retrotransposons in the human genome. *Genome Res.* 18:1875–1883.
- Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 30:2114–2120.
- Bruno M, Mahgoub M, Macfarlan TS. 2019. The arms race between KRAB–zinc finger proteins and endogenous retroelements and its impact on mammals. *Annu Rev Genet.* 53:393–416.
- Castro-Diaz N, et al. 2014. Evolutionally dynamic L1 regulation in embryonic stem cells. *Genes Dev.* 28:1397–1409.
- Ecco G, et al. 2016. Transposable elements and their KRAB-ZFP controllers regulate gene expression in adult tissues. *Dev Cell.* 36: 611–623.
- Ecco G, Imbeault M, Trono D. 2017. A tale of domestication: the endovirome, its polydactyl controllers and the species-specificity of human biology. *Development* 144:2719–2729.
- Eiraku M, et al. 2008. Self-organized formation of polarized cortical tissues from ESCs and its active manipulation by extrinsic signals. *Cell stem cell.* 3:519–532.
- Emerson RO, Thomas JH. 2009. Adaptive evolution in zinc finger transcription factors. *PLoS Genet.* 5:e1000325.

- ENCODE Project Consortium. 2004. The ENCODE (ENCyclopedia of DNA elements) project. *Science*. 306:636–640.
- Familioe G, Lodewijk GA, Robben SF, van Bree EJ, Jacobs FMJ. 2020. Widespread correlation of KRAB zinc finger protein binding with brain-developmental gene expression patterns. *Philos Trans R Soc Lond B Biol Sci*. 375:20190333.
- Field AR, et al. 2019. Structurally conserved primate LncRNAs are transiently expressed during human cortical differentiation and influence cell-type-specific genes. *Stem Cell Rep*. 12:245–257.
- Frietze S, Lan X, Jin VX, Farnham PJ. 2010. Genomic targets of the KRAB and SCAN domain-containing zinc finger protein 263. *J Biol Chem*. 285:1393–1403.
- Gasteiger E, et al. 2003. ExPASy: the proteomics server for in-depth protein knowledge and analysis. *Nucleic Acids Res*. 31:3784–3788.
- Goecks J, Nekrutenko A, Taylor J; Galaxy Team. 2010. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol*. 11(8):R86.
- Groner AC, et al. 2010. KRAB-zinc finger proteins and KAP1 can mediate long-range transcriptional repression through heterochromatin spreading. *PLoS Genet*. 6:e1000869.
- Haring NL, et al. 2021. ZNF91 Deletion in human embryonic stem cells leads to ectopic activation of SVA retrotransposons and up-regulation of KRAB zinc finger gene clusters. *Genome Res*. 31:551–563.
- Heinz S, et al. 2010. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol Cell*. 38:576–589.
- Helleboid P-Y, et al. 2019. The interactome of KRAB zinc finger proteins reveals the evolutionary history of their functional diversification. *EMBO J*. 38:e101220.
- Hendriks WT, Jiang X, Daheron L, Cowan CA. 2015. TALEN-and CRISPR/Cas9-mediated gene editing in human pluripotent stem cells using lipid-based transfection. *Curr Protoc Stem Cell Biol*. 2015:5B.3.1–5B.3.25.
- Imbeault M, Helleboid PY, Trono D. 2017. KRAB zinc-finger proteins contribute to the evolution of gene regulatory networks. *Nature* 543:550–554.
- Jacobs FM, et al. 2014. An evolutionary arms race between KRAB zinc-finger genes ZNF91/93 and SVA/L1 retrotransposons. *Nature* 516(7530):242–245.
- Karolchik D, et al. 2004. The UCSC Table Browser data retrieval tool. *Nucleic Acids Res*. 32:D493–D496.
- Kent WJ. 2002. BLAT—the BLAST-like alignment tool. *Genome Res*. 12:656–664.
- Kim D, Paggi JM, Park C, Bennett C, Salzberg SL. 2019. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat Biotechnol*. 37(8):907–915.
- Labun K, Montague TG, Gagnon JA, Thyme SB, Valen E. 2016. CHOPCHOP v2: a web tool for the next generation of CRISPR genome engineering. *Nucleic Acids Res*. 44:W272–W276.
- Lander ES, et al. 2001. Initial sequencing and analysis of the human genome. *Nature* 409:860–921.
- Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. *Nat Methods*. 9(4):357–359.
- Li X, et al. 2008. A maternal-zygotic effect gene, *Zfp57*, maintains both maternal and paternal imprints. *Dev Cell*. 15:547–557.
- Liao Y, Smyth GK, Shi W. 2014. Featurecounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* 30:923–930.
- Looman C, Åbrink M, Mark C, Hellman L. 2002. KRAB Zinc finger proteins: an analysis of the molecular mechanisms governing their increase in numbers and complexity during evolution. *Mol Biol Evol*. 19:2118–2130.
- Love MI, Huber W, Anders S. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol*. 15:550.
- Najafabadi HS, et al. 2015. C2h2 zinc finger proteins greatly expand the human regulatory lexicon. *Nat Biotechnol*. 33(5):555–562.
- Nowick K, Hamilton AT, Zhang H, Stubbs L. 2010. Rapid sequence and expression divergence suggest selection for novel function in primate-specific KRAB-ZNF genes. *Mol Biol Evol*. 27:2606–2617.
- Persikov A V, Singh M. 2014. De novo prediction of DNA-binding specificities for Cys2His2 zinc finger proteins. *Nucleic Acids Res*. 42:97–108.
- Pontis J, et al. 2019. Hominoid-specific transposable elements and KZFPs facilitate human embryonic genome activation and control transcription in naive human ESCs. *Cell Stem Cell*. 24:724–735.e5.
- Quenneville S, et al. 2011. In embryonic stem cells, ZFP57/KAP1 recognize a methylated hexanucleotide to affect chromatin and DNA methylation of imprinting control regions. *Mol Cell*. 44:361–372.
- Ramírez F, et al. 2016. Deeptools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Res*. 44:W160–W165.
- R Core Team. 2019. R: A language and environment for statistical computing. Vienna (Austria): R Foundation for Statistical Computing. <https://www.R-project.org/>.
- Rice P, Longden I, Bleasby A. EMBOSS: the European molecular biology open software suite. *Trends Genet*. 2000;16:276–277.
- Schmitges FW, et al. 2016. Multiparameter functional diversity of human C2H2 zinc finger proteins. *Genome Res*. 26:1742–1752.
- Sievers F, et al. 2011. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol Syst Biol*. 7:539.
- Sundaram V, Wysocka J. 2020. Transposable elements as a potent source of diverse cis-regulatory sequences in mammalian genomes. *Philos Trans R Soc Lond B Biol Sci*. 375:20190347.
- Takahashi N, et al. 2019. ZNF445 is a primary regulator of genomic imprinting. *Genes Dev*. 33:49–54.
- Taylor MS, Kai C, Kawai J, Carninci P, Hayashizaki Y. 2006. Heterotachy in Mammalian Promoter Evolution. *PLOS Genet*. 2:e30.
- Thomas JH, Schneider S. 2011. Coevolution of retroelements and tandem zinc finger genes. *Genome Res*. 21:1800–1812.
- Turelli P, Playfoot C, Grun D, Raclot C, Pontis J. 2020. Primate-restricted KRAB zinc finger proteins and target retrotransposons control gene expression in human neurons. *Sci Adv*. 6.
- van Bree EJ, et al. 2022. A hidden layer of structural variation in transposable elements reveals potential genetic modifiers in human disease-risk loci. *Genome Res*. 32:656–670.
- Warnes MGR, Bolker B, Bonebakker L, Gentleman R, Huber W. 2016. Package ‘gplots’. Various R programming tools for plotting data.
- Wickham H. 2016. ggplot2: elegant graphics for data analysis.
- Wolf G, et al. 2020. Krab-zinc finger protein gene expansion in response to active retrotransposons in the murine lineage. *Elife* 9:e56337.
- Yang P, et al. 2017. A placental growth factor is silenced in mouse embryos by the zinc finger protein ZFP568. *Science* 356:757–759.
- Zerbino DR, Johnson N, Juettemann T, Wilder SP, Flicek P. 2014. Wiggletools: parallel processing of large collections of genome-wide datasets for visualization and statistical analysis. *Bioinformatics* 30:1008–1009.
- Zhang Y, et al. 2008. Model-based analysis of ChIP-seq (MACS). *Genome Biol*. 9(9):R137.

Associate editor: Dr. Josefa Gonzalez