TRNP1 sequence, function and regulation co-evolve with cortical folding in mammals

Zane Kliesmete[1,§], Lucas E. Wange[1,§], Beate Vieth[1], Miriam Esgleas[2,3], Jessica Radmer[1], Matthias Hülsmann[1,4], Johanna Geuder[1], Daniel Richter[1], Mari Ohnuki[1], Magdalena Götz[2,3,5], Ines Hellmann[1,*,§], Wolfgang Enard[1,*,§]

[1] Anthropology and Human Genomics, Department of Biology II, Ludwig-Maximilians Universitaet, Munich, Germany
[2] Department of Physiological Genomics, BioMedical Center - BMC, Ludwig-Maximilians Universitaet, Munich, Germany
[3] Institute for Stem Cell Research, Helmholtz Zentrum Muenchen, German Research Center for Environmental Health, Neuherberg, Germany
[4] current address: Department of Environmental Microbiology, Eawag, 8600 Dübendorf, Switzerland & Department of Environmental Systems Science, ETH Zurich, 8092 Zürich, Switzerland
[5] SYNERGY, Excellence Cluster of Systems Neurology, BioMedical Center (BMC), Ludwig-Maximilians-Universitaet Muenchen, Planegg/Munich, Germany

[§] equal author contribution
[*] correspondence hellmann@bio.lmu.de, enard@bio.lmu.de

Abstract

Genomes can be seen as notebooks of evolution that contain unique information on successful genetic experiments[1]. This allows to identify conserved genomic sequences[2] and is very useful e.g. for finding disease-associated variants[3]. Additional information from genome comparisons across species can be leveraged when considering phenotypic variance across species. Here, we exemplify such a cross-species association study for the gene *TRNP1* that is important for mammalian brain development. We find that the rate of TRNP1 protein evolution is highly correlated with the rate of cortical folding across mammals and that TRNP1 proteins from species with more cortical folding induce higher proliferation rates in neural stem cells. Furthermore, we identify a regulatory element in *TRNP1* whose activity correlates with cortical folding in Old World Monkeys and Apes. Our analyses indicate that coding and regulatory changes in *TRNP1* have modulated its activity to adjust cortical folding during mammalian evolution and provide a blueprint for cross-species association studies.

Investigating mechanisms of molecular and cellular processes in model organisms by artificial genetic mutations is a central part of biological research. Investigating the evolution of organismic traits by analysing natural genetic mutations is another. These two areas increased their overlap due to the availability of genetic information from many organisms enabled by DNA reading technology and - more recently - also by testing genetic variants from many organisms enabled by DNA writing technology.

Here, we use these newly available resources to better understand which structural or regulatory molecular changes are necessary for a primate or mammalian brain to increase its size and/ or folding[4–8]. Several genes have been connected to the evolution of brain size by comparing the functional consequences of orthologues between pairs of species as for example human and chimpanzee[4,8–10]. While mechanistically convincing, it is unclear whether the proposed evolutionary link can be generalised. On the other side, approaches that correlate sequence changes with brain size changes across a larger phylogeny often lack mechanistic evidence[11]. Hence, a combination of mechanistic and comparative genetic approaches would reveal functional consequences of natural genetic variants and would leverage unique information to better understand the mechanisms of brain evolution as well as brain development. For example, TRNP1 promotes proliferation and NSC self-renewal in murine and ferret NSCs and its knock-down or dominant-negative forms promote the generation of cells leading to cortical folding. Importantly, regulation of its expression in a block-wise manner is critical for folding[6,12] and its N- and C-terminal intrinsically disordered domains are crucial for its function in regulating several nuclear compartments and M-phase length[13]. Thus, there is strong evidence that the regulation of a single gene, *TRNP1*, is necessary and sufficient to induce cortical folds in ferrets and mice. But it is entirely unclear whether the evolutionary diversity in cortical folding and brain size across mammals is mediated by structural and regulatory evolution of TRNP1. To answer this, we investigate genetic differences in TRNP1 coding as well as regulatory sequences and correlate them with brain phenotypes across the mammalian phylogeny.

Co-evolution between TRNP1 protein and cortical folding

We experimentally and computationally collected[14] and aligned[15] 45 mammalian TRNP1 coding sequences, including for example dolphin, elephant and 18 primates (97.4% completeness, Extended Data Fig. 1a). Using this large multiple alignment, we find that the best fitting evolutionary model includes that 8.2% of the codons show signs of recurrent positive selection[16] (i.e. $\omega > 1$, p-value< 0.001, Suppl. Table 8). Six codons with a selection signature could be pinpointed with high confidence (Suppl. Table 9) and five out of those six reside within the first intrinsically disordered region (IDR) and one in the second IDR of the protein (Fig. 1b, Extended Data Fig. 1b). The IDRs have been shown to mediate homotypic and heterotypic interactions of TRNP1 and the associated functions of phase separation, nuclear compartment size regulation and M-phase length regulation[13]. While this shows that TRNP1 evolves under positive selection, it is yet unclear whether this selection is linked to cortical folding and/or brain size.

Cortical folding is usually quantified as the gyrification index (GI), which is the ratio of the cortical surface over the perimeter of the brain surface: A GI$=1$ indicates a completely smooth brain and a GI>1 indicates increasing levels of cortical folding[17]. In addition to GI and brain size, we also consider body mass as a potential confounding variable. Larger animals often have smaller effective population sizes, which in turn reduces the efficiency of selection. Therefore, a correlation between ω and body size could also be explained by relaxation of constraint instead of directional selection[18]. Estimates for these three traits and TRNP1 sequences were available for 31 species (Fig. 1a). In order to test whether the evolution of the TRNP1 protein coding sequences is linked to any of the three traits, we used Coevol[18], a Bayesian MCMC method that jointly models the rates of substitutions and quantitative traits. The resulting covariance matrix of substitution rates (branch length λ_S, ω) and the phenotypic traits then allows for a quantitative evaluation of a potential co-evolution using the posterior probability (pp) of the correlations[18]. Considering the traits separately, we find that GI has the highest marginal correlation with ω (r=0.62, pp=0.95), followed by brain size (r=0.5, pp=0.89), and body mass (r=0.44, pp=0.85) (Suppl. Table 10). To better disentangle their effects, we then simultaneously inferred their correlations (Fig. 1c, 1d, Extended Data Fig. 1c, Suppl. Table 11). GI remained the strongest and only significant marginal correlation (r=0.69, pp=0.98) and also the strongest partial correlation (r=0.47, pp=0.87) compared to brain size (r=0.27, pp=0.75) and body mass (r=0.035, pp=0.51). Hence, these results show that TRNP1 evolved under positive selection and that its rate of sequence evolution is linked strongest to the evolution of gyrification, independent of the evolution of body mass. This indicates that TRNP1 evolved under directional selection because the degree of gyrification changed during mammalian evolution.

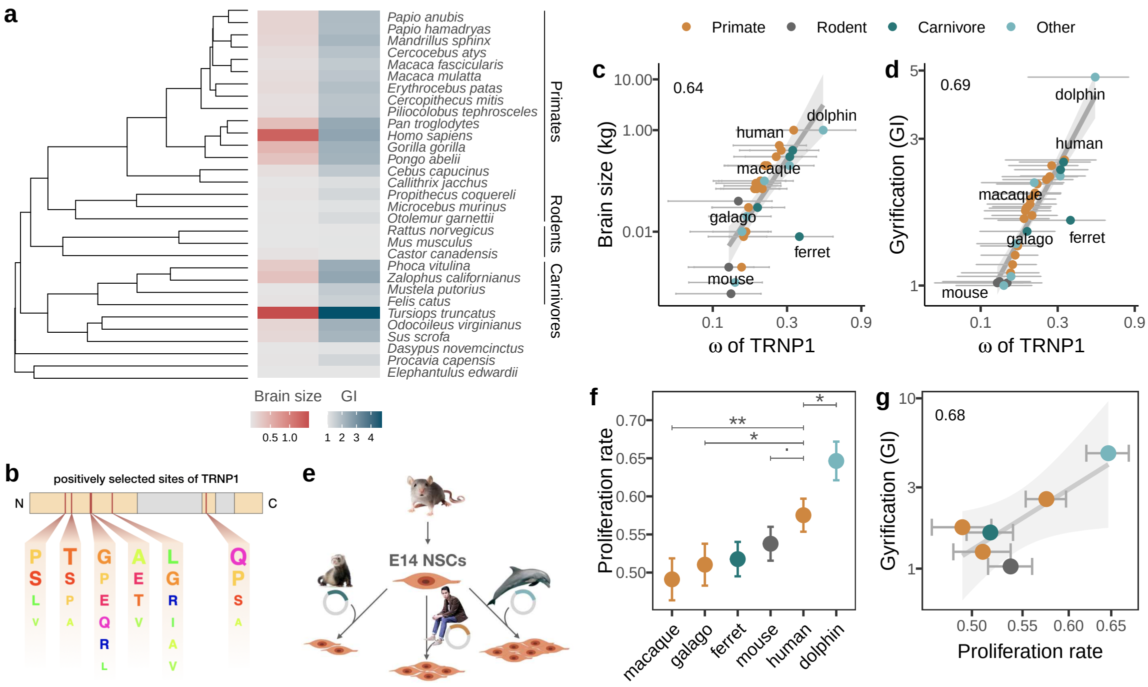


Fig. 1. TRNP1 amino acid substitution rates and proliferative activity co-evolve with cortical folding in mammals
a, Mammalian species for which brain size, GI measurements and TRNP1 coding sequences were available (n=31). The majority of the included species are primates (n=18). **b,** Scheme of the mouse TRNP1 protein (223 AAs) with intrinsically disordered regions (orange) and sites (red lines) subject to positive selection in mammals (ω >1, pp > 0.95; Extended Data Fig. 1b). The letter size of the depicted AAs represents the abundance of AAs at the positively selected sites. **c,** TRNP1 protein substitution rates (ω) correlate non-significantly with brain size (r = 0.64, pp=0.89). Grey horizontal lines represent 95% confidence intervals of ω. **d,** ω significantly correlates with GI (r = 0.69, pp=0.98). **e,** Six different TRNP1 orthologues were transfected into neural stem cells (NSCs) isolated from cerebral cortices of 14 day old mouse embryos and proliferation rates were assessed after 48 h using Ki67-immunostaining as proliferation marker in 7-12 independent biological replicates. **f,** Proliferation rate estimates according to TRNP1 orthologues with bars indicating standard errors of logistic regression and asterisks indicating the significance of pairwise comparisons (Tukey test, p-value: .<0.1, *<0.01, **<0.001). **g,** Proliferation rates are a significant predictor for GI in the respective species (PGLS LRT: χ^2=6.76, df=1, p-value< 0.01; β=3.91 ±1.355, R^2 = 0.68, n=6). Error bars indicate standard errors.

Next, we explored functional properties of TRNP1 that could be affected by these evolutionary changes. As previous studies had shown that TRNP1 transfection increases the proliferation of neural stem cells (NSCs)[12,13], we compared this property among six TRNP1 orthologues, covering the observed range of GI and ω (Fig. 1d). We quantified the proportion of transfected (GFP+) and proliferating (Ki67+) primary mouse NSCs from embryonic day 14 for each construct in 7-12 independent transfections. We confirmed that TRNP1 transfection does increase proliferation compared to a GFP-only control (p-value< 2×10^{-16}; Extended Data Fig. 1d). Remarkably, the proportion of proliferating cells was highest in cells transfected with dolphin TRNP1 followed by human, which was significantly higher than the two other primates, galago and macaque (Fig. 1f; Suppl. Tables 14,15). Notably, the effect of mouse TRNP1 (0.54) was slightly higher than expected given its ω and GI, possibly caused by an increased oligomerisation with the endogenous mouse TRNP1[13]. Nevertheless, even when including the mouse TRNP1, the proliferative activity of TRNP1 is a significant predictor of the gyrification of its species of origin (Phylogenetic generalised least squares PGLS, Likelihood Ratio Test (LRT) p-value< 0.01, $R^2 = 0.68$; Fig. 1g). These results provide strong evidence that the evolution of cortical folding is tightly linked to the evolution of the TRNP1 protein.

Co-evolution of *TRNP1* regulation and cortical folding

We next investigated whether changes in *TRNP1* regulation may also be associated with the evolution of cortical folding and/or brain size by analyzing co-variation in the activity of *TRNP1* associated cis-regulatory elements (CREs). The recently developed massively parallel reporter assays (MPRAs) allow to measure regulatory activity for thousands of sequences at the same time. To this end, a library of putative regulatory sequences is cloned into a reporter vector and their activity is quantified simultaneously by the expression levels of element-specific barcodes[19].

To identify putative CREs of *TRNP1*, we used DNase Hypersensitive Sites (DHS) from human fetal brain[20] and found three upstream CREs, the promoter-including exon 1, an intron CRE, one CRE overlapping the second exon, and one downstream CRE (Fig. 2a). The orthologous regions of the first five CREs were also identified as open chromatin in fetal brains of mice[20] (Extended Data Fig. 3), suggesting that those regions are likely to be CREs also in other mammals. We obtained additional orthologous sequences to the human CRE sequences either from genome databases or by sequencing yielding a total of 351 putative CREs in a panel of 75 mammalian species (Fig. 2b; Extended Data Fig. 3).

Due to limitations in the length of oligonucleotide synthesis, we cut each orthologous putative CRE into 94 base pairs highly overlapping fragments, resulting in 4950 sequence tiles, each synthesised together with a unique barcode sequence. From those, we successfully constructed a complex and unbiased (Extended Data Fig. 2a, 2b) lentiviral plasmid library containing at least 4251 (86%) CRE sequence tiles. Next, we stably transduced this library into neural progenitor cells (NPCs) derived from two humans and one macaque[21]. We calculated the activity per CRE sequence tile as the read-normalised reporter gene expression over the read-normalised input plasmid DNA (Fig. 2a). Finally, we use the per-tile activities (Extended Data Fig. 2c) to reconstruct the activities of the putative CREs from the 75 species. To this end, we summed all tile sequence activities for a given CRE while correcting for the built-in sequence overlap (Fig. 2b, Methods). CRE activities correlate well across cell lines and species (Pearson's r 0.85-0.88; Extended Data Fig. 2d). The CREs covering exon1, the intron and downstream of *TRNP1* show the highest total activity across species and the upstream regions the lowest (Extended Data Fig. 4a). Next, we tested whether CRE activity can explain part of the variance in either brain size or GI across the 45 of the 75 mammalian species for which these phenotypes were available. None of the CREs showed any association with brain size (PGLS, LRT p-value> 0.1). In contrast, we found that the CRE activity of the intron sequence had a slight positive association with gyrification (PGLS, LRT p-value< 0.1, Fig. 2c left, Suppl. Table 16). These associations are much weaker than those of the TRNP1 protein evolution analysed above. Part of the reason might be that CREs have a much higher evolutionary turnover rate than coding sequences[22,23]. This also results in shorter orthologous sequences in species more distantly related to humans that defined the open chromatin regions (Extended Data Fig. 3). Therefore, we restricted our analysis to the catarrhine clade that encompasses Old World Monkeys, great apes and humans. Here, the association between intron CRE activity and GI becomes considerably stronger (PGLS, LRT p-value< 0.003, Fig. 2c right; Suppl. Table 17). Moreover, the intron CRE activity-GI association was consistently detectable across all three cell lines including the macaque NPCs (Suppl. Table 17).

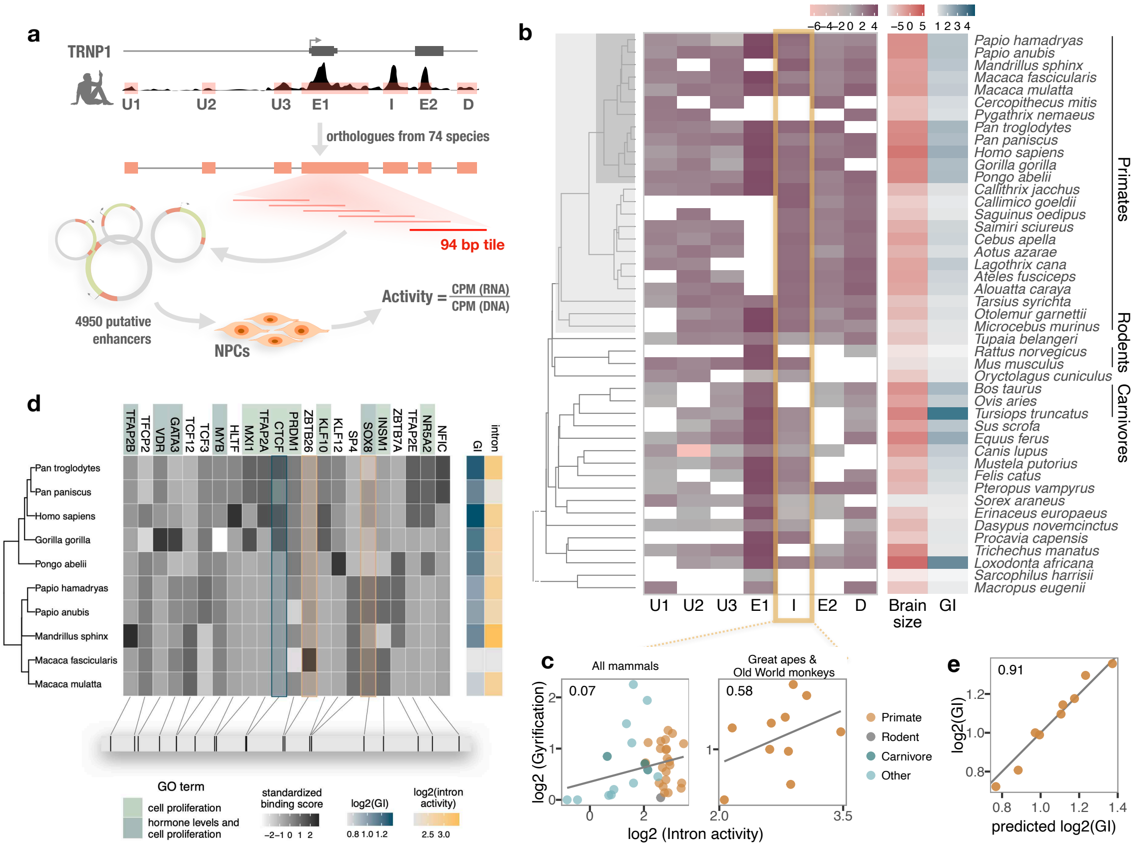


Fig. 2. Activity of a cis-regulatory element (CRE) of *TRNP1* correlates with cortical folding in catarrhines

a, Experimental setup of the MPRA assay. Regulatory activity of 7 putative TRNP1 CREs from 75 species were assayed in neural progenitor cells (NPC) derived from human and macaque induced pluripotent stem cells using lentiviral transduction. **b,** Log-transformed total regulatory activity per CRE in NPCs across species with available brain size and GI measurements (n=45). **c,** Regulatory activity of the intron CRE is moderately associated with gyrification across mammals (PGLS, LRT p-value$<$ 0.1, $R^2 = 0.07$, n=37) and strongest across great apes and Old World Monkeys, i.e. catarrhines (PGLS, LRT p-value$<$ 0.003, $R^2 = 0.58$, n=10). **d,** Variation in binding scores of 22 transcription factors (TFs) across catarrhines. Shown are all TFs that are expressed in NPCs and have their binding motif enriched (motif weight>=1) in the intron CRE sequence of catarrhines. Heatmaps indicate standardised binding scores (grey), GI values (blue) and intron CRE activities (yellow) from the respective species. TF background color indicates gene ontology assignment of the TFs to the 2 most significantly enriched biological processes (Fisher's p-value$<$ 0.05). The bottom panel indicates the spatial position of the top binding site (motif score>3) for each TF on the human sequence. Binding scores of 3 TFs (CTCF, ZBTB26, SOX8) are predictive for intron CRE activity, whereas only CTCF binding shows an association with the GI (PGLS, LRT p-value$<$ 0.05). **e,** A model combining TRNP1 protein evolution rates (ω) and intron activity as predictors can explain GI across OWMs and great apes significantly better than ω alone (PGLS, LRT p-value$<$ 0.003, $R^2 = 0.91$, n=9), indicating an additive, non-redundant effect of the TRNP1 regulatory and structural evolution on the gyrification across these primate species.

Reasoning that changes in CRE activities will most likely be mediated by their interactions with transcription factors (TF), we analysed the sequence evolution of putative TF binding sites. Among the 392 TFs that are expressed in our NPC lines, we identified 22 with an excess of binding sites[24] within the catarrhine intron CRE sequences (Fig. 2d, Extended Data Fig. 4b). In agreement with TRNP1 itself being involved in the regulation of cell proliferation[12,13,25], these 22 TFs are predominantly involved in biological processes such as regulation of cell population proliferation and regulation of hormone levels (Fig. 2d; Suppl. Table 18). To further prioritise the 22 TFs, we used motif binding scores of these TFs for each of the 10 catarrhine intron CRE sequences to predict the observed intron activity in the MPRA assay and to predict the GI of the respective species. Out of the 22 TFs that are expressed and have TFBS enrichment within the intron CRE, the inter-species variation in motif binding scores of only 3 TFs (CTCF, ZBTB26, SOX8) is predictive for intron activity and only CTCF binding scores are predictive for GI (PGLS, LRT p-value< 0.05, Extended Data Fig. 4d, 4e). In summary, we find evidence that a higher activity of the intron CRE is correlated with gyrification in catarrhines, indicating that also regulatory changes in *TRNP1* contributed to the evolution of gyrification. To gauge the combined effects of structural and regulatory changes in TRNP1 to gyrification, we combined the standardised values of the estimated protein evolution rates of TRNP1 (ω) and intron CRE activity across catarrhines within the same PGLS framework (Fig. 2e). Although ω of TRNP1 alone could already explain 75.5% of the variance in gyrification across these species, adding intron CRE activity significantly improved the model (PGLS, LRT p-value< 0.003, Suppl. Table 19), explaining in total 91% of the variance in GI across catarrhines. This suggests that in addition to the changes in the coding region, changes in regulatory regions of *TRNP1* also contribute to evolving more gyrified brains.

Discussion

Here, we have shown that the rate of protein evolution of TRNP1 correlates with gyrification in mammals and that the activity of a regulatory element of *TRNP1* co-evolves with gyrification in catarrhines. Additionally, we have shown that also the proliferative activity of TRNP1 varies across mammals, as it would be predicted if TRNP1 was indeed the target for positive selection for a more gyrified brain. Hence, while previous experimental studies have speculated that TRNP1 could be important for the evolution of gyrification[26], our analyses provide evidence that this is indeed the case.

Of note, the effect of structural changes appears stronger than the effect of regulatory changes. This is contrary to the notion that regulatory changes should be the more likely targets of selection as they are more cell-type specific[27] (but see also[28]). However, measures of regulatory activity are inherently less precise than counting amino acid changes, which will necessarily deflate the estimated association strength[22,23]. In any case, our analysis shows that evolution combined both regulatory and structural evolution to modulate and fine tune TRNP1 activity.

Moreover, our analyses generate specific hypotheses about the molecular mechanisms used to tune gyrification. They strongly suggest that an increased gyrification goes along with an increased proliferation activity in NSCs and suggests that amino acid changes in the disordered regions are responsible for this. Furthermore, we find that CTCF binding potential of the intron CRE is correlated with gyrification in cattharines. This indicates a role for CTCF in regulating gyrification, in line with its regulatory role for several developmental processes[29].

Finally, we think that our approach could serve as a blueprint to leverage the unique information stored in the evolutionary diversity among species. The fundamental principle of correlating genetic variants with phenotypes (GWAS) is a well established approach within populations and thus we believe that cross-species association studies (CSAS) will prove instrumental to understand complex phenotypes. Genome sequences for essentially all vertebrates and eukaryotes are becoming available[2,30,31], making the availability of phenotype information the only limit to cross-species association studies. On a molecular and cellular level, phenotyping of induced pluripotent stem cells and their derivatives across many species would boost this approach[32,33], further helping to tap the potential of life's diversity to understand molecular mechanisms.

References

1. Wright, S. H. Lander celebrates genome milestone in heavily attended talk. Accessed: 2020-5-22. http://news.mit.edu/2001/lander-0228 (2001).
2. Zoonomia Consortium. A comparative genomics multitool for scientific discovery and conservation. en. *Nature* **587,** 240–245 (2020).
3. Kircher, M. *et al.* A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet.* **46,** 310–315 (2014).
4. Enard, W. Comparative genomics of brain size evolution. *Frontiers in human neuroscience* **8,** 345 (2014).
5. Borrell, V. & Calegari, F. Mechanisms of brain evolution: regulation of neural progenitor cell diversity and cell cycle length. *Neuroscience research* **86,** 14–24 (2014).
6. Borrell, V. & Götz, M. Role of radial glial cells in cerebral cortex folding. en. *Curr. Opin. Neurobiol.* **27,** 39–46 (2014).
7. Lewitus, E. *et al.* An adaptive threshold in mammalian neocortical evolution. *PLoS biology* **12,** e1002000 (2014).
8. Llinares-Benadero, C. & Borrell, V. Deconstructing cortical folding: genetic, cellular and mechanical determinants. *Nature Reviews Neuroscience* **20,** 161–176 (2019).
9. Heide, M. *et al.* Human-specific ARHGAP11B increases size and folding of primate neocortex in the fetal marmoset. en. *Science* (2020).
10. Johnson, M. B. *et al.* Aspm knockout ferret reveals an evolutionary mechanism governing cerebral cortical size. en. *Nature* **556,** 370–375 (2018).
11. Montgomery, S. H. *et al.* Adaptive evolution of four microcephaly genes and the evolution of brain size in anthropoid primates. *Molecular biology and evolution* **28,** 625–638 (2010).
12. Stahl, R. *et al.* Trnp1 Regulates Expansion and Folding of the Mammalian Cerebral Cortex by Control of Radial Glial Fate. *Cell* **153,** 535–549. ISSN: 0092-8674 (2013).
13. Esgleas, M. *et al.* Trnp1 organizes diverse nuclear membrane-less compartments in neural stem cells. *The EMBO journal* **39,** e103373 (2020).
14. Camacho, C. *et al.* BLAST+: architecture and applications. *BMC bioinformatics* **10,** 421 (2009).
15. Löytynoja, A. in *Multiple sequence alignment methods* 155–170 (Springer, 2014).
16. Yang, Z. PAML: a program package for phylogenetic analysis by maximum likelihood. *Bioinformatics* **13,** 555–556 (1997).
17. Zilles, K. *et al.* Gyrification in the cerebral cortex of primates. *Brain, Behavior and Evolution* **34,** 143–150 (1989).
18. Lartillot, N. & Poujol, R. A phylogenetic model for investigating correlated evolution of substitution rates and continuous phenotypic characters. *Molecular biology and evolution* **28,** 729–744 (2010).
19. Inoue, F. & Ahituv, N. Decoding enhancers using massively parallel reporter assays. en. *Genomics* **106,** 159–164 (2015).
20. Bernstein, B. E. *et al.* The NIH roadmap epigenomics mapping consortium. *Nature biotechnology* **28,** 1045 (2010).
21. Geuder, J. *et al.* A non-invasive method to generate induced pluripotent stem cells from primate urine. *bioRxiv.* eprint: https://www.biorxiv.org/content/early/2020/08/12/2020.08.12.247619 (2020).
22. Danko, C. G. *et al.* Dynamic evolution of regulatory element ensembles in primate CD4+ T cells. en. *Nat Ecol Evol* **2,** 537–548 (2018).
23. Berthelot, C. *et al.* Complexity and conservation of regulatory landscapes underlie evolutionary resilience of mammalian gene expression. en. *Nat Ecol Evol* **2,** 152–163 (2018).

24. Frith, M. C. *et al.* Cluster-Buster: Finding dense clusters of motifs in DNA sequences. *Nucleic acids research* **31,** 3666–3668 (2003).
25. Volpe, M. *et al.* trnp: A conserved mammalian gene encoding a nuclear protein that accelerates cell-cycle progression. en. *DNA Cell Biol.* **25,** 331–339 (2006).
26. Martínez-Martínez, M. Á. *et al.* A restricted period for formation of outer subventricular zone defined by Cdh1 and Trnp1 levels. *Nature communications* **7,** 11812 (2016).
27. Carroll, S. B. Evo-devo and an expanding evolutionary synthesis: a genetic theory of morphological evolution. *Cell* **134,** 25–36 (2008).
28. Hoekstra, H. E. & Coyne, J. A. The locus of evolution: evo devo and the genetics of adaptation. en. *Evolution* **61,** 995–1016 (2007).
29. Arzate-Mejía, R. G. *et al.* Developing in 3D: the role of CTCF in cell differentiation. *Development* **145,** dev137729 (2018).
30. Koepfli, K.-P. *et al.* The Genome 10K Project: a way forward. en. *Annu Rev Anim Biosci* **3,** 57–111 (2015).
31. Lewin, H. A. *et al.* Earth BioGenome Project: Sequencing life for the future of life. en. *Proc. Natl. Acad. Sci. U. S. A.* **115,** 4325–4333 (2018).
32. Enard, W. Functional primate genomics-leveraging the medical potential. *J. Mol. Med.* **90,** 471–480 (2012).
33. Housman, G. & Gilad, Y. Prime time for primate functional genomics. en. *Curr. Opin. Genet. Dev.* **62,** 1–7 (2020).

Methods

Sequencing of *TRNP1* for primate species

Identification of cis-regulatory elements of *TRNP1*). DNase hypersensitive sites in the proximity to *TRNP1* (25 kb upstream, 3 kb downstream) were identified in human fetal brain and mouse embryonic brain DNase-seq data sets downloaded from NCBI's Sequence Read Archive (Suppl. Table 2). Reads were mapped to human genome version hg19 and mouse genome version mm10 using NextGenMap with default parameters (NGM; v. 0.0.1)[34]. Peaks were identified with Hotspot v.4.0.0 using default parameters[35]. Overlapping peaks were merged, and the union per species was taken as putative cis-regulatory elements (CREs) of *TRNP1* (Suppl. Tables 3 & 4). The orthologous regions of human *TRNP1* DNase peaks in 49 mammalian species were identified with reciprocal best hit using BLAT (v. 35x1)[36]. Firstly, sequences of human *TRNP1* DNase peaks were extended by 50 bases down and upstream of the peak. Then, the best matching sequence per peak region were identified with BLAT using the following settings: -t=DNA -q=DNA -stepSize=5 -repMatch=2253 -minScore=0 -minIdentity=0 -extendThroughN. These sequences were aligned back to hg19 using the same settings as above. The resulting best matching hits were considered reciprocal best hits if they fell into the original human *TRNP1* CREs.

Cross-species primer design for sequencing. We sequenced TRNP1 coding sequences in 6 primates for which reference genome assemblies were either unavailable or very sparse (see Suppl. Table 1). We used NCBI's tool Primer Blast[37] with the human *TRNP1* gene locus as a reference. Primer specificity was confirmed using the predicted templates in 12 other primate species available in Primer Blast. Primer pair 1 was used for sequencing library generation as it reliably worked for all 6 resequenced primate species (Suppl. Table 20).

In order to obtain *TRNP1* CREs for the other primate species, we designed primers using primux[38] based on the species with the best genome assemblies and were subsequently tested in closely related species in multiplexed PCR reactions. The species used as reference for primer design were H.sapiens (hg19), P.paniscus (PanPan1), P.troglodytes (panTro4), G.gorilla (gorGor3), P.abelii (ponAbe2), N.leucogenys (nomLeu3), M.fascicularis (MacFas), M.mulatta (rheMac3), P.anubis (papAnu2), P.hamadryas (papHam1), C.sabaeus (ChlSab), C.jacchus (calJac3) and S.boliviensis (saiBol1). A detailed list of designed primer pairs per CRE and reference genome can be found in Suppl. Table 21 and final pools of multiplexed primers per CRE and species can be found in Suppl. Table 22.

Sequencing of target regions for primate species. Primate gDNAs were obtained from Deutsches Primaten Zentrum, DKFZ and MPI Leipzig (see Suppl. Table 5). Depending on concentration, gDNAs were whole genome amplified prior to sequencing library preparation using GenomiPhi V2 Amplification Kit (Sigma). After amplification, gDNAs were cleaned up using SPRI beads (CleaNA). Both *TRNP1* coding regions and CREs were resequenced using a similar approach that included a touchdown PCR to amplify the target region followed by a ligation and Nextera XT library construction. The main difference between the two being the polymerase and the exact touchdown PCR conditions. For the CRE resequencing Q5 High-Fidelity DNA Polymerase (New England Biolabs) was used, while KAPA HiFi Polymerase (Roche) was used for the coding region. Both with their respective High GC buffer to account for the high GC content of the template. PCRs were performed as a 25 µl reaction on varying amounts of Template DNA (usually 20 ng) and 0.05 μM - 0.4 μM per primer depending on the degree of multiplexing. The thermo cycling conditions were according to manufacturers instructions for the respective polymerase. While the CREs were amplified for 20 cycles in a touchdown PCR followed by 20 cycles of standard PCR, the coding

region was amplified only in a touchdown PCR for 30 cycles. In the touchdown phase the annealing temperature was gradually decreased from 70 °C in the first cycle to 60 °C in the last cycle. Final elongation times were altered according to expected product length.
Blunt end ligation of PCR fragments and phosphorylation were done in the same reaction in T4 DNA ligation buffer with 2.5 U of T4 DNA ligase, 10 U of T4 Polynucleotide Kinase (Thermo Fisher Scientific) and up to 0.5 μg of input at 16°C overnight. The resulting products were SPRI bead purified. A Nextera XT DNA sample preparation kit (Illumina) was used to perform the library preparation from ligated PCR products according to the manufacturer's protocol with the following modifications: the initial 55°C tagmentation time was increased from 5 to 10 minutes and the whole reaction was downscaled to 1/5th. Unique i5 and i7 primers were used for each library. Fragment size distributions were determined using capillary gel electrophoresis (Agilent Bioanalyzer 2100,DNA HS Kit) and libraries were pooled in equimolar ratios. For the missing CDS, sequencing was performed as 250 bases paired end with dual indexing on an Illumina MiSeq and the CRE libraries libraries were sequenced 50bp paired end on an Illumina Hiseq 1500.

Assembly of sequenced regions. Reads were demultiplexed using deML[39]. The resulting sequences per species were subsequently trimmed to remove PCR-handles using cutadapt (version 1.6)[40]. For sequence reconstruction, Trinity (version 2.0.6) in reference-guided mode was used[41]. The reference here is defined as the mapping of sequences to the closest reference genome with NGM (version 0.0.1)[34]. Furthermore, read normalisation was enabled and a minimal contig length of 500 was set. The sequence identity of the assembled contigs was validated by BLAT[36] alignment to the closest reference *TRNP1* as well as to the human *TRNP1*. The assembled sequence with the highest similarity and expected length was selected per species.

***TRNP1* coding sequence retrieval and alignment.** Human TRNP1 protein sequence was retrieved from UniProt database[42] under accession number Q6NT89. We used the human TRNP1 in a `tblastn`[43] search of genomes from 45 species specified in Suppl. Table 1 (R-package rBLAST version 0.99.2). The following additional arguments were specified:
–`soft masking false` — Turn off applying filtering locations as soft masks
–`seg no` — Turn off masking low complexity sequences.
PRANK[44] (version 150803) was used to align TRNP1 coding sequences, using the mammalian tree from Bininda et al.[45] to guide the multiple alignment. Alignment was done using the default settings, specifying one additional parameter:
–`translate` — additionally translate the aligned nucleotide sequences to a protein.

Evolutionary sequence analysis

Identification of sites under positive selection. Program `codeml` from PAML software[46] (version 4.8) was used to infer whether a significant proportion of TRNP1 protein sites evolve under positive selection across the phylogeny of 45 species. Site models M8 and M7 were compared[47], that allow ω to vary among sites across the phylogenetic tree, but not between branches. M7 and M8 are nested with M8 allowing for sites under positive selection with ω_s. Likelihood ratio test (LRT) was used to compare these models. Naive Empirical Bayes (NEB) analysis was used to identify the specific sites under positive selection ($\Pr(\omega > 1) > 0.95$). Additional general model settings:
–tree topology from[45]
–`seqtype = 1` — codon model
–`clock = 0` — no molecular clock

–`aaDist = 0` — equal AA distances

–`CodonFreq = 2` — codon frequency: from the average nucleotide frequencies at the third codon positions (F3X4).

Inferring correlated evolution using Coevol. Coevol[48] (version 1.4) was utilised to infer the covariance between TRNP1 evolutionary rate ω and three morphological traits (brain size, GI and body mass) across species (Suppl. Table 7). `dsom` command was used to activate the codon model, in which the two a priori independent variables are dS and ω. For each model, the MCMC was run for at least 10,000 cycles, using the first 1,000 as burn-in and two runs were performed to examine global convergence. `tracecomp` was used to access the relative differences between runs ($d = 2|\mu_1 - \mu_2|/(\sigma_1 + \sigma_2)$) where μ are the means and σ are standard deviations of each parameter for the two chains) and mixing diagnostics (effective sample size) between the runs. All parameters have a relative difference < 0.1 and effective size > 300, which indicates good convergence and quantitatively reliable runs[48].

We report posterior probabilities (*pp*), the marginal and partial correlations of the full model (Suppl. Table 11) and the separate models where including only either one of the three traits (Suppl. Table 10). The posterior probabilities for a negative correlation are given by $1 - pp$. These were back-calculated to make them directly comparable, independently of the correlation direction, i.e. higher *pp* means more statistical support for the respective correlation.

Proliferation assay

Plasmids. The six *TRNP1* orthologous sequences containing the restriction sites BamHI and XhoI were synthetized by GeneScript (`www.genscript.com`). All plasmids for expression were first cloned into a pENTR1a gateway plasmid described in Stahl et al., 2013[49] and then into a Gateway (Invitrogen) form of pCAG-GFP (kind gift of Paolo Malatesta). The gateway LR-reaction system was used to then sub-clone the different TRNP1 forms into the pCAG destination vectors.

Primary cerebral cortex cultures and transfection. Cerebral cortices were dissected removing the ganglionic eminence, the olfactory bulb, the hippocampal anlage and the meninges and cells were mechanically dissociated with a fire polish Pasteur pipette. Cells were then seeded onto poly-D-lysine (PDL)-coated glass coverslips in DMEM-GlutaMAX with 10% FCS (Life Technologies) and cultured at 37°C in a 5% CO^2 incubator. Plasmids were transfected with Lipofectamine 2000 (Life technologies) according to manufacturer's instruction 2h after seeding the cells onto PDL coated coverslips. One day later cells were washed with phosphate buffered saline (PBS) and then fixed in 4% Paraformaldehyde (PFA) in PBS and processed for immunostaining.

Immunostaining. Cells plated on poly-D-lysine coated glass coverslips were blocked with 2% BSA, 0.5% Triton-X (in PBS) for 1 hour prior to immunostaining. Primary antibodies (GFP and Ki67) were applied in blocking solution overnight at 4°C. Fluorescent secondary antibodies were applied in blocking solution for 1 hour at room temperature. DAPI (4'.6-Diamidin-2-phenylindol, Sigma) was used to visualize nuclei. Stained cells were mounted in Aqua Polymount (Polysciences). All secondary antibodies were purchased from Life Technologies. Images were taken using an epifluorescence microscope (Zeiss, Axio ImagerM2) equipped with a 20X/ 0.8 N.A and 63X/1.25 N.A. oil immersion objectives. Post image processing with regard to brightness and contrast was carried out where appropriate to improve visualization, in a pairwise manner.

Proliferation rate calculation using logistic regression. The proportion of successfully transfected cells that proliferate under each condition (Ki67-positive/GFP-positive) was modelled using logistic regression (R-package `stats` (version 4.0.3), `glm` function) with logit link function $logit(p) = log(\frac{p}{1-p})$, for $0 \leq p \leq 1$, where p is the probability of success. The absolute number of GFP-positive cells were added as weights. Model selection was done using LRT within `anova` function from `stats`. Adding the donor mouse as a batch improved the models (Suppl. Tables 12, 13). To back-calculate the absolute proliferation probability (i.e., rate) under each condition, intercept of the respective model was set to zero and the inverse logit function $\frac{e^{\beta_i X_i}}{1+e^{\beta_i X_i}}$ was used, where i indicates condition (Suppl. Table 14). Two-sided multiple comparisons of means between the conditions of interest were performed using `glht` function (Tukey test, user-defined contrasts) from R package `multcomp` (version 1.4-13) (Suppl. Table 15).

Phylogenetic modeling of proliferation rates using generalized least squares (PGLS). The association between the induced proliferation rates for each TRNP1 orthologue and the GI of the respective species was analysed using generalised least squares (R-package `nlme`, version 3.1-143), while correcting for the expected correlation structure due to phylogenetic relation between the species. The expected correlation matrix for the continuous trait was generated using a Brownian motion[50,18] (`ape` (version 5.4), using function `corBrownian`) on the mammalian phylogeny from Bininda et al. (2007)[45] adding the missing species (Fig 1a). The full model was compared to a null model using the likelihood ratio test (LRT). Residual R^2 values were calculated using `R2.resid` function from R package RR2 (version 1.0.2).

Massively Parallel Report Assay (MPRA)

MPRA library design. *TRNP1* CRE sequences identified in human fetal brain, mouse embryonic brain as well as orthologous regions in 73 mammalian species were considered for the Massively Parallel Reporter Assay (MPRA). In total 351 sequences were included where a sliding window per sequence entry was applied moving by 40 bases for the sequences that are longer than 94 bases, resulting in 4,950 oligonucleotide sequences which are flanked upstream by a first primer site (ACTGGCCGCTTCACTG), downstream by a KpnI/Xbal restriction cut site (GGTACCTCTAGA), a 10 base long barcode sequence as well as a second primer site (AGATCGGAAGAGCGTCG). Barcode tag sequences were specifically designed so that they contain all four nucleotides at least once, do not contain stretches of four identical nucleotides, do not contain microRNA seed sequences (retrieved from microRNA Bioconductor R package version 1.28.0) and do not contain restriction cut site sequences for KpnI nor Xbal (5'-GGTACG-3', 3'-CCATGG-5').

MPRA plasmid library construction. We modified the original protocol by Melnikov et al.[52]: We used a lentiviral delivery system as previously described[53] and introduced green fluorescent protein (GFP) instead of nano luciferase. All DNA purification and clean up steps were performed using SPRI beads unless stated otherwise, all plasmid DNA isolations were done using the standard protocol of a column-based kit (PureYield Plasmid Midiprep System, Promega). Primer sequences and plasmids used in the MPRA can be found in Suppl. Table 23 and 24 respectively. The *TRNP1* enhancer oligonucleotide library was synthesised on an oligo array by Custom Array. In a first step the single stranded oligos were double stranded and restriction sites flanking these oligos were introduced via PCR and subsequently used for directional cloning. Emulsion PCR using the commercially available Micellula DNA Emulsion & Purification Kit (roboklon) was performed in this step to avoid loss of individual variants and ensure unbiased amplification. Restriction digest using SfiI (New England Biolabs) and cloning of the variant library into the pMPRA1 plasmid (Addgene, #49349) was performed according to the original protocol. To ensure maximum complexity this first step was carried out twice and the initial emulsion PCR was performed in quadruplicates each time. Transformation of the ligation products was performed in triplicates where one fourth of the ligation product was used to transform 50 μl of chemically competent *E. coli* (5-alpha High Efficiency, New England Biolabs). Next a constant sequence, transcribed under the influence of the library, needed to be introduced into the generated plasmid pool. In the original publication a nano luciferase with a minimal promoter is introduced however we decided to use a GFP reporter here. To this end we used pNL3.1 and replaced the nano luciferase with an EGFP ORF using Gibson assembly. The resulting plasmid carried the same restriction sites as the ones used in the original publication and hence cloning was performed as described previously[52]. Electroporation into electro competent *E. coli* (10-beta, New England Biolabs) was performed to maximise transformation efficiency. All transformations were carried out in triplicates that were pooled and grown in 150 ml liquid cultures. For the final cloning step, the enhancer library including GFP and the minimal promoter were inserted into a lentiviral backbone (pMPRAlenti1, Addgene #61600). Both plasmids were digested with SfiI (New England Biolabs) to allow for directional cloning of the whole construct. The lentiviral backbone pMPRAlenti1 was treated similarly with the addition of shrimp alkaline phosphatase (rSAP,New England Biolabs). Ligation was performed with a 3:1 molar ratio of backbone to insert using 1 U of T4 DNA Ligase (Thermo Fisher Scientific) and 1 mM ATP for 3 h at 20°C. Ligation reactions were cleaned up and used to transform electrocompetent E.coli (10-beta, New England Biolabs). All transformations were pooled and used to inoculate a 200 ml bacterial culture and plasmids were isolated as before.

MPRA lentiviral particle production. Lentiviral particles were produced according to standard methods in HEK 293T cells[54]. The MPRA library was co-transfected with third generation lentiviral plasmids carrying the env, rev, gag and pol genes (pMDLg/pRRE, pRSV-Rev; Addgene #12251, #12253) as well as the VSV glycoprotein (pMD2.G,Addgene #12259) using Lipofectamine 3000. The lentiviral particle containing supernatant was harvested 48 hrs post transfection and filtered using 0.45 μm PES syringe filters. Viral titer was determined by infecting Neuro-2A cells (ATCC CCL-131) and counting GFP positive cells. To this end, N2A cells were infected with a 50/50 volume ratio of viral supernatant to cell suspension with addition of 8 μg/ml Polybrene. Cells were exposed to the lentiviral particles for 24 hrs until medium was exchanged. After additional 48 hrs, infected cells were positively selected using Blasticidin.

Culture of neural progenitor cells. Neural progenitor cells were cultured on Geltrex (Thermo Fisher Scientific) in DMEM F12 (Fisher scientific) supplemented with 2 mM GlutaMax-I (Fisher Scientific), 20 ng/µl bFGF (Peprotech), 20 ng/µl hEGF (Miltenyi Biotec), 2% B-27 Supplement (50X) minus Vitamin A (Gibco), 1% N2 Supplement 100X (Gibco), 200μM L-Ascorbic acid 2-phosphate (Sigma) and 100 μg/ml penicillin-streptomycin with medium change every second day. For passaging, NPCs were washed with PBS and then incubated with TrypLE Select (Thermo Fisher Scientific) for 5 min at 37°C. Culture medium was added and cells were centrifuged at 200 × g for 5 min. Supernatant was replaced by fresh culture medium and cells were transferred to a new Geltrex coated dish. The cells were split every two to three days in a ratio of 1:3.

MPRA lentiviral transduction. The transduction of the MPRA library was performed in triplicates on two *Homo sapiens* and one *Macaca fascicularis* NPC lines[55] (see Suppl. Table 6). 2.5×10^5 NPCs per line and replicate were dissociated, dissolved in 500 µl cell culture medium containing 8 μg/ml Polybrene and incubated with virus at MOI 12.7 for 1 h at 37°C in suspension[56].

Thereafter cells were seeded on Geltrex and cultured as described above. Virus containing medium was replaced the next day and cells were cultured for additional 24 hrs. Cells were collected, lysed in 100 μl TRI reagent and frozen at -80°C.

MPRA sequencing library generation. As input control for RNA expression, DNA amplicon libraries were constructed using 100 - 500 pg plasmid DNA. Library preparation was performed in two subsequent PCRs. A first PCR termed Adapter PCR introduced the 5' transposase mosaic end, this was used in the second PCR (Index PCR) to add a library specific index sequence and Illumina Flow cell adapters. The Adapter PCR was performed in triplicates using DreamTaq polymerase (Thermo Fisher Scientific). PCR products were cleaned up using SPRI beads (1/1 ratio) and quantified. 1-5 ng were subjected to the Index PCR using Q5 polymerase. After cleaning up libraries using SPRI beads (2/1 ratio), amplified DNA was quantified and quality control was performed using capillary gel electrophoresis (Agilent Bioanalyzer 2100). Total RNA from NPCs was extracted using the Direct-zol RNA Microprep Kit (Zymo Research). 500ng of RNA were subjected to reverse transcription using Maxima H Minus RT (Thermo Fisher Scientific) with Oligo-dT primers. 50 ng of cDNA were used for library preparation as described for plasmid DNA, with the alteration that Q5 DNA polymerase was used in both PCRs. 15-20 ng of the Adapter PCR product were subjected to the second library PCR and further treated as described for plasmid libraries. Plasmid and cDNA libraries were pooled and quality was evaluated using capillary gel electrophoresis (Agilent Bioanalyzer 2100). Sequencing was performed on an Illumina HiSeq 1500 instrument using a single-index, 50bp, paired-end protocol.

MPRA data processing and analysis. MPRA reads were demultiplexed with `deML`[39] using i5 and i7 adapter indices from Illumina. Next, we removed barcodes with low sequence quality, requiring a minimum Phred quality score of 10 for all bases of the barcode (zUMIs, `fqfilter.pl` script[57]). Furthermore, we removed reads that had mismatches to the constant region (the first 20 bases of the GFP sequence `TCTAGAGTCGCGGCCTTACT`). The remaining reads that matched one of the known CRE-tile barcodes were tallied up resulting in a count table. Next, we filtered out CRE tiles that had been detected in only one of the 3 input plasmid library replicates (4202/4950). Counts per million (CPM) were calculated per CRE tile per library (median counts: $\sim$ 900k range: 590k-1,050k). Macaque replicate 3 was excluded due its unusually low correlation with the other samples (Pearson's r: $\bar{r} - 1.5 \times \sigma_r$). The final regulatory activity for each CRE tile per cell line was calculated as:

$$a_i = \frac{median(CPM_i)}{median(CPM_i)_p}, \tag{1}$$

where a is regulatory activity, i indicates CRE tile and p is the input plasmid library. Median was calculated across the replicates from each cell line.

Given that each tile was overlapping with two other tiles upstream and two downstream, we calculated the total regulatory activity per CRE region in a coverage-sensitive manner, i.e. for each position in the original sequence mean per-bp-activity across the detected tiles covering it was calculated. The final CRE region activity is the sum across all base positions.

$$a_r = \sum_{b=1}^{k} \frac{1}{n} \sum_{i=1}^{n} \frac{a_i}{l_i}, \tag{2}$$

where a_r is regulatory activity of CRE region r, $b = 1, ..., k$ is the base position of region r, $i, ..., n$ are tiles overlapping the position b, a_i is tile activity from equation 1 and l_i is tile length. CRE activity and brain phenotypes were associated with one another using PGLS analysis (see above). The number of species varied for each phenotype-CRE pair (brain size: min. 37 for exon1, max.

48 for intron and downstream regions; GI: min. 32 for exon2, max. 37 for intron), therefore the activity of each of the seven CRE regions was used separately to predict either GI or brain size of the respective species.

Combining protein evolution rates and intron activity to predict GI across catharrines

PGLS model fits were compared either including only ω of TRNP1 protein from Coevol[48] or including ω and intron CRE activity as predictors. For this, the standardised values of either measurement were used, calculated as $\frac{x_i - \bar{x}}{\sigma}$, where x_i is each observed value, $\bar{x}$ is the mean and σ is the standard deviation.

Transcription Factor analysis

RNA-seq library generation RNA sequencing was performed using the prime-seq method, a bulk derivative of the single cell RNA-seq method SCRB-seq[58]. The major features of this method are early pooling enabled by the introduction of a cell barcode and a unique molecular identifier (UMI) in the reverse transcription reaction followed by full length cDNA amplification and enrichment of 3 prime ends in the library preparation. The full prime-seq protocol including primer sequences can be found at protocols.io (`https://www.protocols.io/view/prime-seq-s9veh66`). Here we used 10 ng of the isolated RNA from the MPRA experiment and subjected it to the prime-seq protocol with minor modifications. As sequencing of the MPRA transcripts contained in the RNA of the infected NPCs, may lead to problems in sequencing due to a duplicated read start, we determined the amount of contamination caused by MPRA transcripts in the transcriptome library. Using an additional primer (Suppl. Table 23) in the pre-amplification which generates a small MPRA amplicon, followed by a size selection we found the contamination to be negligible and proceeded with the standard prime-seq protocol. After reverse transcription all samples were pooled, the pool was cleaned up, Exonuclease 1 (Thermo Fisher Scientific) digested and finally subjected to cDNA amplification using Kapa HiFi polymerase (Roche). Nextera XT (Illumina) library preparation was performed in triplicates of 0.8 ng of amplified cDNA each. Instead of an i5 index primer a custom 3 prime enrichment primer was added to the Library PCR reaction and annealing temperature was increased to 62°C. The replicates of the sequencing library were pooled and size selected (300 -900 bp) using an 2% agarose gel followed by gel extraction to ensure optimal sequencing quality. Finally the size distribution and molarity of the library was measured using capillary gel electrophoresis (Agilent Bioanalyzer 2100). Sequencing was performed on an Illumina HiSeq 1500 instrument with an unbalanced paired end layout, where read 1 was 16 base pair long and read 2 was 50 base par long, additionally an 8 base pair index read was performed.

RNA-seq data processing Bulk RNA-seq data was generated from the same 9 samples (3 cell lines, 3 biological replicates each) that were transduced and assayed in the MPRA. This was done to detect which TFs are expressed in the assayed cell lines and might be responsible for the observed intron CRE activity. Raw read fastq files were pre-processed using zUMIs[57] together with STAR[59] to generate expression count tables for barcoded UMI data. Reads were mapped to human reference genome (hg38, Ensembl annotation GRCh38.84). Further filtering was applied keeping genes that were detected in at least 7/9 samples and had on average more than 7 counts. For further analysis, we used normalised and variance stabilised expression estimates as provided by DESeq2[60].

TFBS motif analysis on the intron CRE sequence TF Position Frequency Matrices (PFM) were retrieved from JASPAR CORE 2020[61], including only non-redundant vertebrate motifs (746 in total). These were filtered for the expression in our NPC RNA-seq data, leaving 392 TFs with 462 motifs in total).

A Hidden Markov Model (HMM)-based program `Cluster-Buster`[62] (compiled on Jun 13 2019) was used to infer the enriched TF binding motifs on the intron sequence. First, the auxiliary program `Cluster-Trainer` was used to find the optimal gap parameter between motifs of the same cluster and to obtain weights for each TF based on their motif abundance per kb across catharrine intron CREs from 10 species with available GI measurements. Weights for each motif suggested by `Cluster-Trainer` were supplied to `Cluster-Buster` that we used to find clusters of regulatory binding sites and to infer the enrichment score for each motif on each intron sequence. The program was run with the following parameters:

−`g3` — gap parameter suggested by `Cluster-Trainer`

−`c5` — cluster score threshold

−`m3` — motif score threshold.

To identify the most likely regulators of *TRNP1* that bind to its intron sequence and might influence the evolution of gyrification, we filtered for the motifs that were most abundant across the intron sequences (`Cluster-Trainer` weights >1). These motifs were distinct from one another (mean pairwise distance 0.72, Extended Data Fig. 4c). Gene-set enrichment analysis contrasting the TFs with the highest binding potential with the other expressed TFs was conducted using the Bioconductor-package `topGO`[63] (version 2.40.0) (Suppl. Table 18).

PGLS model was applied as previously described, using `Cluster-Buster` binding scores across catharrine intron CRE sequences as predictors and predicting either intron activity or GI from the respective species. The relevance of the three TFs that were associated with intron activity was then tested using an additive model and comparing the model likelihoods with reduced models where either of these were dropped.

34. Sedlazeck, F. J. *et al.* NextGenMap: fast and accurate read mapping in highly polymorphic genomes. en. *Bioinformatics* **29,** 2790–2791. ISSN: 1367-4803, 1367-4811 (2013).
35. John, S. *et al.* Chromatin accessibility pre-determines glucocorticoid receptor binding patterns. *Nat. Genet.* **43,** 264–268. ISSN: 1061-4036, 1546-1718 (2011).
36. Kent, W. J. BLAT—The BLAST-Like Alignment Tool. *Genome Res.* **12,** 656–664. ISSN: 1088-9051, 1549-5469 (2002).
37. Ye, J. *et al.* Primer-BLAST: a tool to design target-specific primers for polymerase chain reaction. en. *BMC Bioinformatics* **13,** 134. ISSN: 1471-2105 (2012).
38. Hysom, D. A. *et al.* Skip the alignment: degenerate, multiplex primer and probe design using K-mer matching instead of alignments. en. *PLoS One* **7,** e34560 (2012).
39. Renaud, G. *et al.* deML: robust demultiplexing of Illumina sequences using a likelihood-based approach. *Bioinformatics* **31,** 770–772 (2015).
40. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. en. *EMBnet.journal* **17,** 10–12. ISSN: 2226-6089, 2226-6089 (2011).
41. Grabherr, M. G. *et al.* Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* **29,** 644–652. ISSN: 1087-0156 (2011).
42. UniProt Consortium. UniProt: a worldwide hub of protein knowledge. *Nucleic acids research* **47,** D506–D515 (2019).
43. Camacho, C. *et al.* BLAST+: architecture and applications. *BMC bioinformatics* **10,** 421 (2009).
44. Löytynoja, A. in *Multiple sequence alignment methods* 155–170 (Springer, 2014).
45. Bininda-Emonds, O. R. *et al.* The delayed rise of present-day mammals. *Nature* **446,** 507 (2007).
46. Yang, Z. PAML: a program package for phylogenetic analysis by maximum likelihood. *Bioinformatics* **13,** 555–556 (1997).
47. Yang, Z. *et al.* Codon-substitution models for heterogeneous selection pressure at amino acid sites. *Genetics* **155,** 431–449 (2000).

48. Lartillot, N. & Poujol, R. A phylogenetic model for investigating correlated evolution of substitution rates and continuous phenotypic characters. *Molecular biology and evolution* **28,** 729–744 (2010).
49. Stahl, R. *et al.* Trnp1 Regulates Expansion and Folding of the Mammalian Cerebral Cortex by Control of Radial Glial Fate. *Cell* **153,** 535–549. ISSN: 0092-8674 (2013).
50. Felsenstein, J. Phylogenies and the comparative method. *The American Naturalist* **125,** 1–15 (1985).
51. Martins, E. P. & Hansen, T. F. Phylogenies and the comparative method: a general approach to incorporating phylogenetic information into the analysis of interspecific data. *The American Naturalist* **149,** 646–667 (1997).
52. Melnikov, A. *et al.* Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay. en. *Nat. Biotechnol.* **30,** 271–277 (2012).
53. Inoue, F. *et al.* A systematic comparison reveals substantial differences in chromosomal versus episomal encoding of enhancer activity. en. *Genome Res.* **27,** 38–52 (2017).
54. Dull, T. *et al.* A third-generation lentivirus vector with a conditional packaging system. en. *J. Virol.* **72,** 8463–8471 (1998).
55. Geuder, J. *et al.* A non-invasive method to generate induced pluripotent stem cells from primate urine. *bioRxiv.* eprint: `https://www.biorxiv.org/content/early/2020/08/12/2020.08.12.247619` (2020).
56. Nakai, R. *et al.* Derivation of induced pluripotent stem cells in Japanese macaque (Macaca fuscata). en. *Sci. Rep.* **8,** 12187 (2018).
57. Parekh, S. *et al.* zUMIs-a fast and flexible pipeline to process RNA sequencing data with UMIs. *Gigascience* **7,** giy059 (2018).
58. Bagnoli, J. W. *et al.* Sensitive and powerful single-cell RNA sequencing using mcSCRB-seq. *Nat. Commun.* **9** (2018).
59. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29,** 15–21 (2013).
60. Love, M. I. *et al.* Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome biology* **15,** 550 (2014).
61. Fornes, O. *et al.* JASPAR 2020: update of the open-access database of transcription factor binding profiles. *Nucleic acids research* **48,** D87–D92 (2020).
62. Frith, M. C. *et al.* Cluster-Buster: Finding dense clusters of motifs in DNA sequences. *Nucleic acids research* **31,** 3666–3668 (2003).
63. Alexa, A. & Rahnenführer, J. Gene set enrichment analysis with topGO. *Bioconductor Improv* **27** (2009).
64. Bernstein, B. E. *et al.* The NIH roadmap epigenomics mapping consortium. *Nature biotechnology* **28,** 1045 (2010).

Data Availability

The RNA-seq data used in this manuscript are publicly available at Array Express E-MTAB-9951. The MPRA data are available at Array Express under accession number E-MTAB-9952. Additional primate sequences for TRNP1 are available at GenBank (MW373535 - MW373709).

Code Availability

A compendium containing processing scripts and detailed instructions to reproduce the analysis for this manuscript is available from the following GitHub repository: `https://github.com/Hellmann-Lab/Co-evolution-TRNP1-and-GI`.

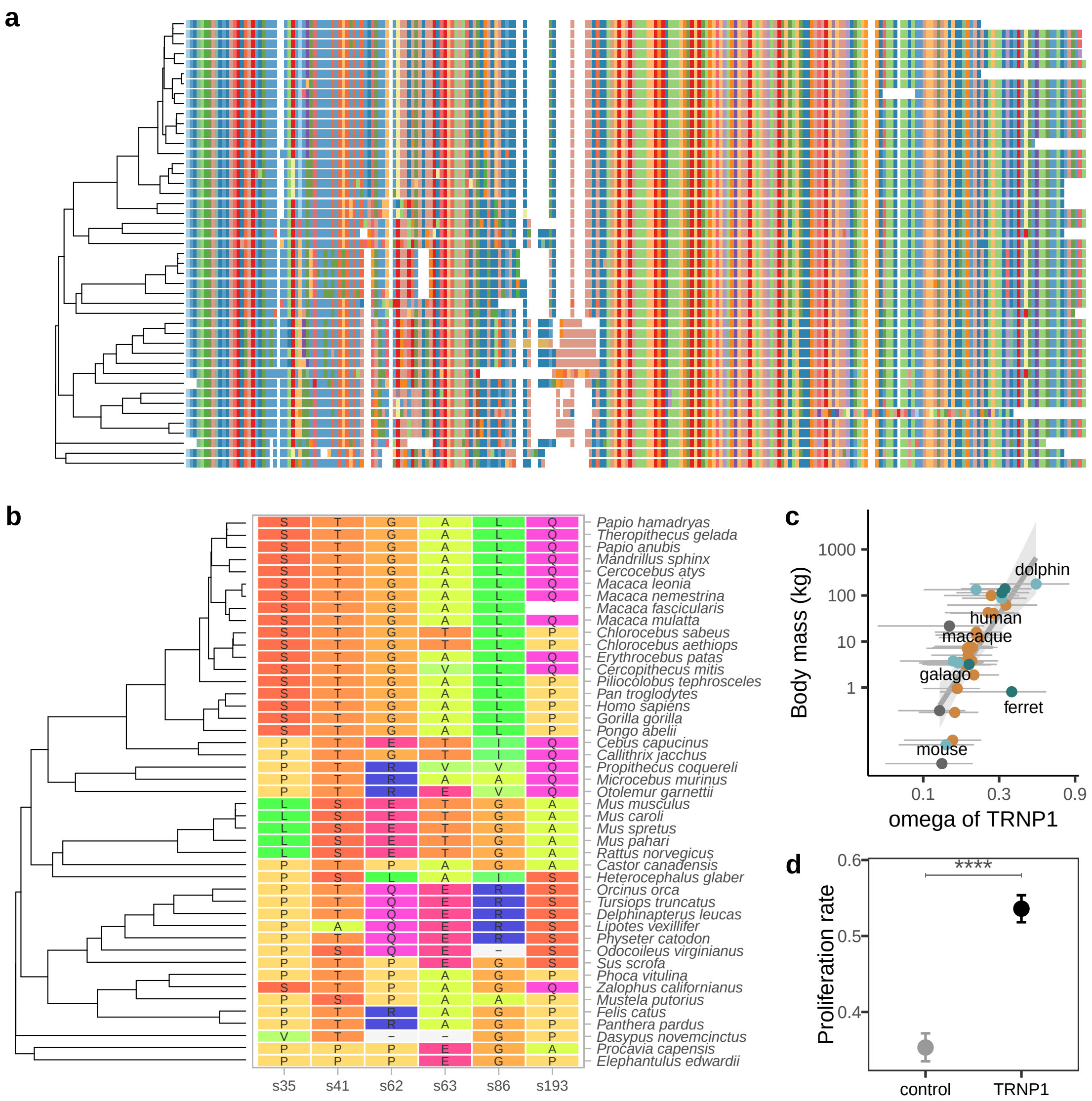
Author Contributions

M.G. proposed the project and W.E and I.H. conceived the approaches of this study. B.V. designed all initial sequence acquisitions. L.W., M.H. D.R. and J.R. conducted the MPRA assay. M.E. designed and conducted the proliferation assay. J.R., J.G. and M.O. were responsible for all primate cell culture work. Z.K. collected, integrated and analysed all data. M.G. and M.E. provided expertise on Trnp1 function throughout the study. W.E. and I.H. supervised the work and provided guidance in data analysis. Z.K., I.H., and W.E. wrote the manuscript. All authors read, corrected and approved the final manuscript.

Acknowledgements

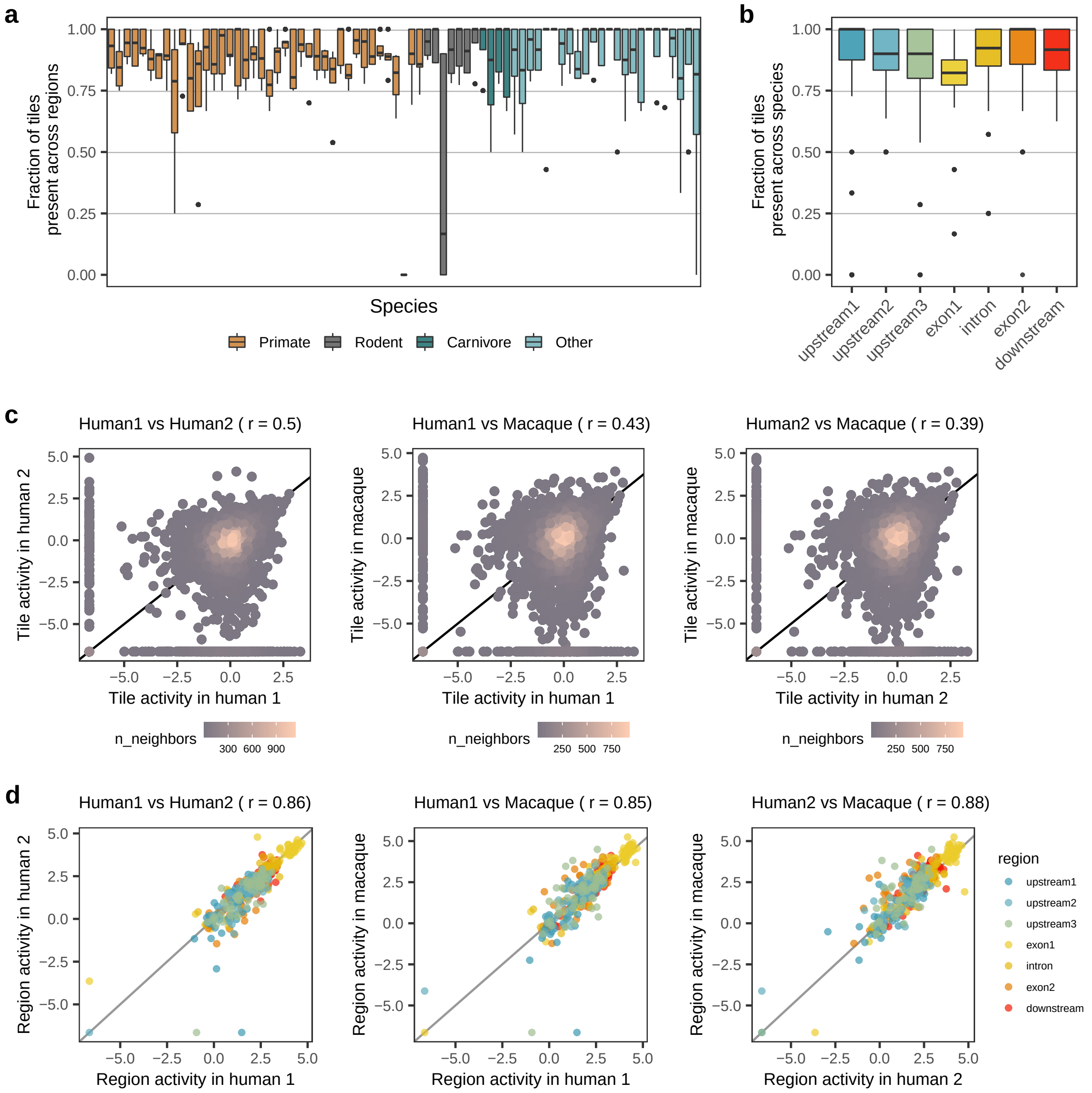
This work was supported by the Deutsche Forschungsgemeinschaft (DFG) through LMUexcellent, SFB1243 (Subproject A14/A15 to W.E. and I.H., respectively), DFG grant HE 7669/1-1 (to I.H.) and the advanced ERC grant ChroNeuroRepair (to M.G.). We want to thank Christian Roos from the German Primate Center for providing genomic DNA from primates, project students Gunnar Kuut and Fatih Sarigoel for helping to generate TRNP1 orthologous sequences, Nikola Vuković for helping to establish the MPRA assay, Reza Rifat for helping with the proliferation assays and Christoph Neumayr for helping in data analysis.

Extended Data Figures

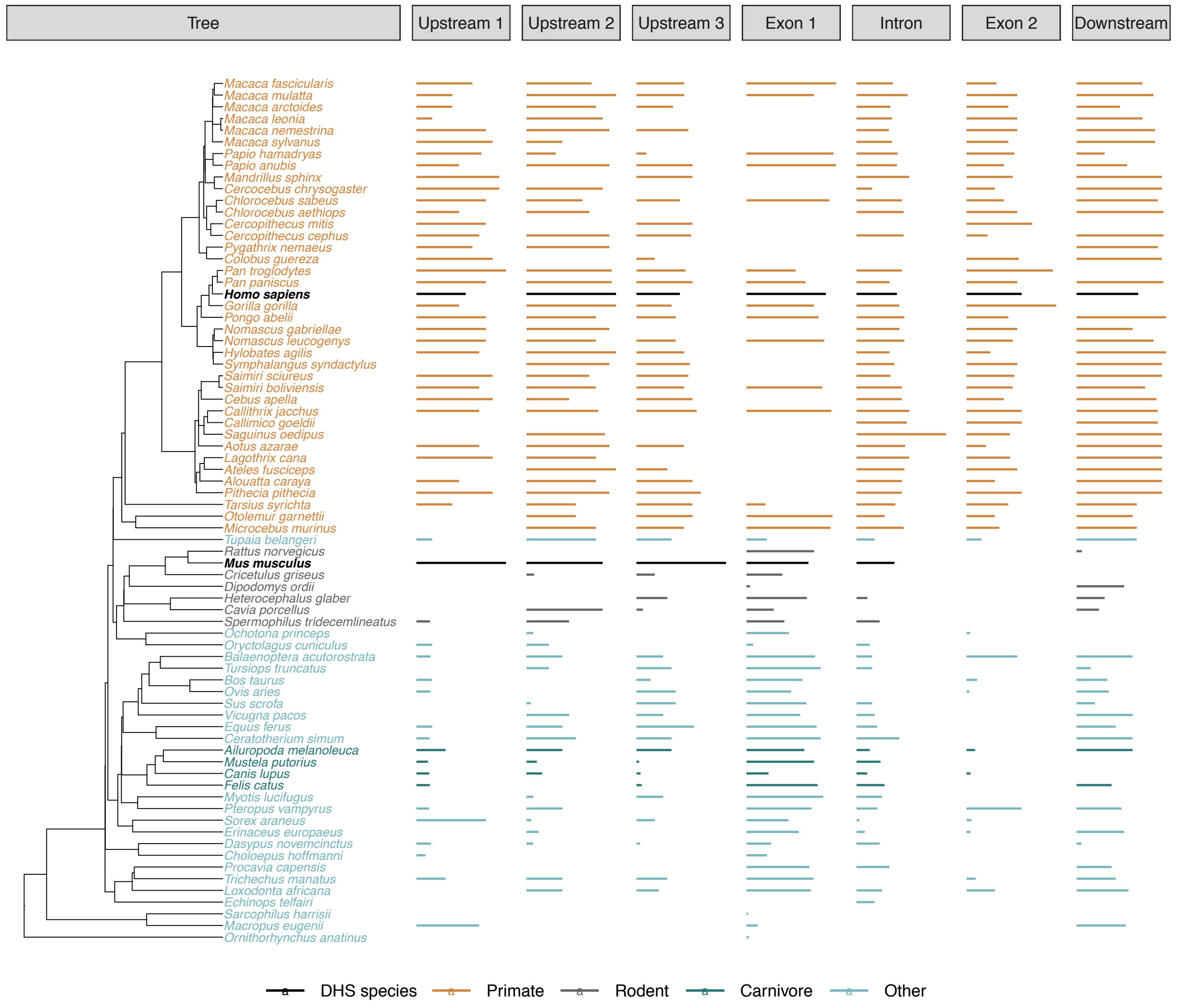


	s35	s41	s62	s63	s86	s193
Papio hamadryas	S	T	G	A	L	Q
Theropithecus gelada	S	T	G	A	L	Q
Papio anubis	S	T	G	A	L	Q
Mandrillus sphinx	S	T	G	A	L	Q
Cercocebus atys	S	T	G	A	L	Q
Macaca leonia	S	T	G	A	L	Q
Macaca nemestrina	S	T	G	A	L	Q
Macaca fascicularis	S	T	G	A	L	
Macaca mulatta	S	T	G	A	L	Q
Chlorocebus sabeus	S	T	G	T	L	P
Chlorocebus aethiops	S	T	G	T	L	P
Erythrocebus patas	S	T	G	A	L	Q
Cercopithecus mitis	S	T	G	V	L	Q
Piliocolobus tephrosceles	S	T	G	A	L	P
Pan troglodytes	S	T	G	A	L	P
Homo sapiens	S	T	G	A	L	P
Gorilla gorilla	S	T	G	A	L	P
Pongo abelii	S	T	G	A	L	P
Cebus capucinus	P	T	E	T	I	Q
Callithrix jacchus	P	T	G	T	I	Q
Propithecus coquereli	P	T	R	V	V	Q
Microcebus murinus	P	T	R	A	A	Q
Otolemur garnettii	P	T	R	E	V	Q
Mus musculus	L	S	E	T	G	A
Mus caroli	L	S	E	T	G	A
Mus spretus	L	S	E	T	G	A
Mus pahari	L	S	E	T	G	A
Rattus norvegicus	L	S	E	T	G	A
Castor canadensis	P	T	P	A	G	A
Heterocephalus glaber	P	S	L	A	I	S
Orcinus orca	P	T	Q	E	R	S
Tursiops truncatus	P	T	Q	E	R	S
Delphinapterus leucas	P	T	Q	E	R	S
Lipotes vexillifer	P	A	Q	E	R	S
Physeter catodon	P	T	Q	E	R	S
Odocoileus virginianus	P	S	Q	E	–	S
Sus scrofa	P	T	P	E	G	S
Phoca vitulina	P	T	P	A	G	P
Zalophus californianus	S	T	P	A	G	Q
Mustela putorius	P	S	P	A	A	P
Felis catus	P	T	R	A	G	P
Panthera pardus	P	T	R	A	G	P
Dasypus novemcinctus	V	T	–	–	G	P
Procavia capensis	P	P	P	E	G	A
Elephantulus edwardii	P	P	P	E	G	P

Extended Data Fig. 1. TRNP1 protein-coding sequence analysis. **a,** Multiple alignment of 45 TRNP1 coding sequences (97.4% completeness) using phylogeny-aware aligner PRANK[44]. The alignment is 744 bases long, which translates to 248 amino acids (AAs). For comparison: human TRNP1 coding sequence is 227 AA long, whereas mouse - 223 AAs. **b,** Sites under positive selection across the phylogenetic tree according to PAML[46] M8 model (in total 8.2% of sites with $\omega > 1$, LRT, p-value< 0.001). The depicted sites had a posterior probability $\Pr(\omega > 1) > 0.95$ according to Naive Empirical Bayes analysis. Colours of the amino acids indicate their relatedness in biochemical properties. Sites with light-grey background and a dash indicate gaps/indels, while a white bar indicates one missing AA. **c,** ω and body mass correlate moderately across mammal species (ω ~BM: r=0.55, pp=0.9). **d,** The overall effect of TRNP1 on proliferation rates in primary mouse NSCs. Proliferation induced by all 6 TRNP1 orthologues combined was compared to the control transfected with a plasmid carrying only GFP, but no TRNP1 coding sequence. TRNP1 presence in NSCs significantly increases the proliferation rates (TRNP1: 0.54 (±0.018), control: 0.35 (±0.019), Tukey test p-value$< 2e-16$, df=68). n=7 for galago, macaque and dolphin, n=12 for mouse, ferret, human and GFP-control.

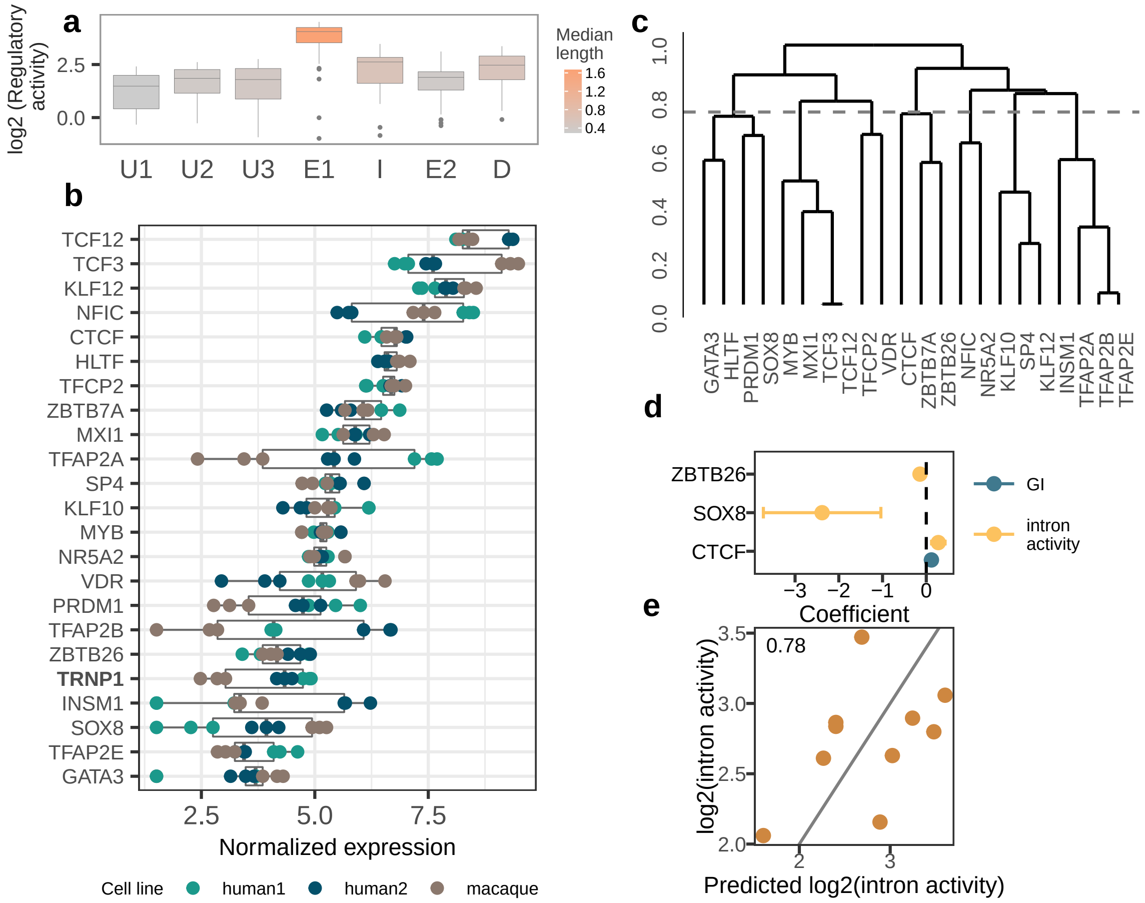


Extended Data Fig. 2. Analysis of massively parallel reporter assay (MPRA) data. **a,** Fraction of the detected CRE tiles in the plasmid library per species across regions. The detection rates are unbiased and uniformly distributed across species and clades with only one extreme outlier *Dipodomys ordii.* This is mainly due to the fact that out of 3 total orthologous regions identified in the genome of this species, upstream 3 consisted of only 1 - uncaptured - tile. **b,** Fraction of the detected CRE tiles in the plasmid library per region across species. **a,b** Each box represents the median and first and third quartiles with the whiskers indicating the most extreme data point, which is no more than 1.5 times the length of the box away from the box. Individual points indicate outliers. **c,** Pairwise correlation of the log2-transformed CRE tile activity between the three transduced cell lines: human 1, human 2 and macaque. Pearson's r is specified in the brackets of figure titles. **d,** Pairwise correlation of the log2-transformed summarized activity per CRE region between cell lines. Pearson's r is specified in the brackets of figure titles.



.

Extended Data Fig. 3. Length of the covered CRE sequences in the MPRA library across the tree. Species for which the regions were inferred based on DNase-Hypersensitive Sites (DHS) from embryonic brain [64] are marked in bold and black: human (*Homo sapiens*) and mouse (*Mus musculus*). These species do not show extreme differences in length compared to others (human: 5/7, mouse: 3/5 regions within the 10% and 90% quantiles). The orthologous CRE sequence length differs strongly between primate and non-primate species, being in average 1.8 to 2.8 times longer in the primate species than in the other mammals.



Extended Data Fig. 4. Intron-binding transcription factor analysis. **a,** Total activity per CRE across species. Exon1 (E1), intron (I) and the downstream (D) regions are more active and longer than other regions. Each box represents the median and first and third quartiles with the whiskers indicating the most extreme data point, which is no more than 1.5 times the length of the box away from the box. Individual points indicate outliers. **b,** Variance-stabilized expression of the 22 transcription factors with enriched binding sites on the intron CRE region. *TRNP1* could be consistently detected in all replicates, meaning that the TFs inducing its expression are present in this cellular system. **c,** Hierarchical clustering (average linkage) of TF Position Frequency matrices retrieved from JASPAR2020[61] for the 22 intron-enriched TFs. Dashed grey line indicates the mean pairwise binding motif distance of 0.72. **d,** Coefficients of the candidate TFs (PGLS, LRT p-value$<$ 0.05) predicting either intron activity or GI using TF binding score for the intron CRE sequence. **e,** Predicted intron activity using the additive model combining the three predictor TF binding scores compared to the observed intron CRE activity across the catarrhine species (R^2=0.78, n=10). Dropping of either predictor TF was not supported by the model (PGLS, LRT p-value$<$ 0.05).

Supplementary Information

TRNP1 sequence, function and regulation co-evolve with cortical folding in mammals

Zane Kliesmete[1,§], Lucas E. Wange[1,§], Beate Vieth[1], Miriam Esgleas[2,3], Jessica Radmer[1], Matthias Hülsmann[1,4], Johanna Geuder[1], Daniel Richter[1], Mari Ohnuki[1], Magdalena Götz[2,3,5], Ines Hellmann[1,*,§], Wolfgang Enard[1,*,§]

[1] Anthropology and Human Genomics, Department of Biology II, Ludwig-Maximilians Universitaet, Munich, Germany

[2] Department of Physiological Genomics, BioMedical Center - BMC, Ludwig-Maximilians Universitaet, Munich, Germany

[3] Institute for Stem Cell Research, Helmholtz Zentrum Muenchen, German Research Center for Environmental Health, Neuherberg, Germany

[4] current address: Department of Environmental Microbiology, Eawag, 8600 Dübendorf, Switzerland & Department of Environmental Systems Science, ETH Zurich, 8092 Zürich, Switzerland

[5] SYNERGY, Excellence Cluster of Systems Neurology, BioMedical Center (BMC), Ludwig-Maximilians-Universitaet Muenchen, Planegg/Munich, Germany

[§] equal author contribution

[*] correspondence hellmann@bio.lmu.de, enard@bio.lmu.de

Supplementary Table 1. Genome sources of the TRNP1 protein coding sequences

	Species	Genome Source	Genome Assembly
1	Callithrix jacchus	NCBI	ASM275486v1
2	Castor canadensis	NCBI	C.can genome v1.0
3	Cebus capucinus	Ensembl41	Cebus_imitator-1.0
4	Cercocebus atys	Ensembl41	Caty_1.0
5	Cercopithecus mitis	targeted resequencing, Enard Lab	-
6	Chlorocebus aethiops	targeted resequencing, Enard Lab	-
7	Chlorocebus sabeus	NCBI	chlSab2
8	Dasypus novemcinctus	NCBI	dasNov3
9	Delphinapterus leucas	NCBI	ASM228892v3
10	Elephantulus edwardii	NCBI	EleEdw1.0
11	Erythrocebus patas	NCBI	EryPat_v1_BIUU
12	Felis catus	Ensembl41	Felis_catus_9.0
13	Gorilla gorilla	NCBI	Susie3
14	Heterocephalus glaber	NCBI	hetGla2
15	Homo sapiens	targeted resequencing, Enard Lab	-
16	Lipotes vexillifer	NCBI	Lipotes_vexillifer_v1
17	Macaca fascicularis	Ensembl41	Macaca_fascicularis_5.0
18	Macaca leonia	targeted resequencing, Enard Lab	-
19	Macaca mulatta	NCBI	rheMac8
20	Macaca nemestrina	Ensembl41	Mnem_1.0
21	Mandrillus sphinx	targeted resequencing, Enard Lab	-
22	Microcebus murinus	NCBI	micMur2
23	Mus caroli	Ensembl41	CAROLI_EIJ_v1.1
24	Mus musculus	NCBI	mm10
25	Mus pahari	Ensembl41	PAHARI_EIJ_v1.1
26	Mus spretus	Ensembl41	SPRET_EiJ_v1
27	Mustela putorius	cDNA, Goetz Lab	-
28	Odocoileus virginianus	NCBI	Ovir.te_1.0
29	Orcinus orca	NCBI	Oorc_1.1
30	Otolemur garnettii	NCBI	otoGar3
31	Pan troglodytes	NCBI	panTro6
32	Panthera pardus	Ensembl41	PanPar1.0
33	Papio anubis	targeted resequencing, Enard Lab	-
34	Papio hamadryas	NCBI	papHam1
35	Phoca vitulina	NCBI	GSC_HSeal_1.0
36	Physeter catodon	NCBI	Physeter_macrocephalus-2.0.2
37	Piliocolobus tephrosceles	NCBI	ASM277652v1
38	Pongo abelii	NCBI	ponAbe3
39	Procavia capensis	NCBI	proCap1
40	Propithecus coquereli	Ensembl41	Pcoq_1.0
41	Rattus norvegicus	NCBI	rn6
42	Sus scrofa	NCBI	susScr3
43	Theropithecus gelada	NCBI	Tgel_1.0
44	Tursiops truncatus	NCBI	Tur_tru v1
45	Zalophus californianus	NCBI	zalCal2.2

Supplementary Table 2. DNase-seq experiments

	SRX	Species	Tissue	Stage	GEO Accession
1	SRX027085	Human	Fetal Brain	Day 85	GSM595922
2	SRX027086	Human	Fetal Brain	Day 85	GSM595923
3	SRX027089	Human	Fetal Brain	Day 96	GSM595926
4	SRX027091	Human	Fetal Brain	Day 96	GSM595928
5	SRX121276	Human	Fetal Brain	Day 101	GSM878650
6	SRX121277	Human	Fetal Brain	Day 104	GSM878651
7	SRX201815	Human	Fetal Brain	Day 105	GSM1027328
8	SRX121278	Human	Fetal Brain	Day 109	GSM878652
9	SRX040380	Human	Fetal Brain	Day 112	GSM665804
10	SRX027083	Human	Fetal Brain	Day 117	GSM595920
11	SRX026914	Human	Fetal Brain	Day 122	GSM530651
12	SRX027076	Human	Fetal Brain	Day 122	GSM595913
13	SRX062364	Human	Fetal Brain	Day 122	GSM723021
14	SRX040395	Human	Fetal Brain	Day 142	GSM665819
15	SRX188655	Mouse	Fetal Brain	E 14.5	GSM1003828
16	SRX191055	Mouse	Fetal Brain	E 14.5	GSM1014197
17	SRX191042	Mouse	Fetal Brain	E 18.5	GSM1014184

Supplementary Table 3. Human *TRNP1* DNase hypersensitive sites

	Chromosome	Start	End	ID
1	chr1	27293479	27293766	upstream1
2	chr1	27310581	27310877	upstream2
3	chr1	27318087	27318439	upstream3
4	chr1	27319449	27321900	promexon1
5	chr1	27323922	27324667	intron
6	chr1	27327174	27327461	exon2
7	chr1	27328171	27328804	downstream

Supplementary Table 4. Mouse *Trnp1* DNase hypersensitive sites

	Chromosome	Start	End	ID
1	chr4	133494338	133494835	intron
2	chr4	133495742	133496109	unique1
3	chr4	133497135	133498824	promexon1
4	chr4	133504292	133504667	unique2
5	chr4	133504895	133505292	unique3
6	chr4	133508796	133509525	upstream3
7	chr4	133511090	133511417	upstream2
8	chr4	133512990	133513416	upstream1

Supplementary Table 5. gDNA samples

	Family	Species	Source
1	Apes	Homo sapiens	DKFZ
2	Apes	Pan troglodytes	MPI Leipzig
3	Apes	Pan paniscus	MPI Leipzig
4	Apes	Gorilla gorilla	MPI Leipzig
5	Apes	Pongo abelii	MPI Leipzig
6	Apes	Symphalangus syndactylus	MPI Leipzig
7	Apes	Nomascus gabriellae	MPI Leipzig
8	Apes	Hylobates agilis	MPI Leipzig
9	Old World Monkeys	Papio anubis	Deutsches Primatenzentrum Goettingen
10	Old World Monkeys	Papio hamadryas	MPI Leipzig
11	Old World Monkeys	Mandrillus sphinx	MPI Leipzig
12	Old World Monkeys	Cercocebus chrysogaster	Deutsches Primatenzentrum Goettingen
13	Old World Monkeys	Macaca mulatta	MPI Leipzig
14	Old World Monkeys	Macaca arctoides	Deutsches Primatenzentrum Goettingen
15	Old World Monkeys	Macaca leonia	Deutsches Primatenzentrum Goettingen
16	Old World Monkeys	Macaca sylvanus	Deutsches Primatenzentrum Goettingen
17	Old World Monkeys	Macaca nemestrina	MPI Leipzig
18	Old World Monkeys	Cercopithecus cephus	Deutsches Primatenzentrum Goettingen
19	Old World Monkeys	Cercopithecus mitis	Deutsches Primatenzentrum Goettingen
20	Old World Monkeys	Chlorocebus aethiops	Deutsches Primatenzentrum Goettingen
21	Old World Monkeys	Colobus guereza	Deutsches Primatenzentrum Goettingen
22	Old World Monkeys	Semnopithecus entellus	Deutsches Primatenzentrum Goettingen
23	Old World Monkeys	Pygathrix nemaeus	Deutsches Primatenzentrum Goettingen
24	New World Monkeys	Alouatta caraya	Deutsches Primatenzentrum Goettingen
25	New World Monkeys	Lagothrix cana	Deutsches Primatenzentrum Goettingen
26	New World Monkeys	Ateles fusciceps	Deutsches Primatenzentrum Goettingen
27	New World Monkeys	Cebus apella	MPI Leipzig
28	New World Monkeys	Saimiri sciureus	MPI Leipzig
29	New World Monkeys	Aotus azarae	MPI Leipzig
30	New World Monkeys	Saguinus oedipus	Deutsches Primatenzentrum Goettingen
31	New World Monkeys	Callimico goeldii	Deutsches Primatenzentrum Goettingen
32	New World Monkeys	Callithrix jacchus	Deutsches Primatenzentrum Goettingen
33	New World Monkeys	Pithecia pithecia	Deutsches Primatenzentrum Goettingen

Supplementary Table 6. Cell lines used for the MPRA

Name	Purpose	Species	Cell Line	Source
Neuro2a cells	Lentiviral titer determination	Mus musculus	N2A	ATCC
Human embryonic kidney cells	Production of lentiviral particles	Homo sapiens	HEK293T	ATCC
Human neural progenitor cells	used in MPRA	Homo sapiens	N4_29B5	Geuder et. al.
Macaca fascicularis neural progenitor cells	used in MPRA	Macaca fascicularis	N15_39B2	Geuder et. al.

Supplementary Table 7. Phenotype data and its source publications used in this study. In cases where there are multiple sources listed, mean across the individual measurements was calculated. For 11 species, the phenotype data of their close sister species (column ”Original species”) was used. For 3 additional species with only missing GI, this information was borrowed from the indicated sister species (column ”GI source” in the brackets)

Species	Original species	Body mass(g)	Brain size(g)	EQ	Brain, body mass source	GI	GI source
Alouatta caraya		2955.00	45.60	1.80	[1]	1.47	[2] (A.seninculus)
Aotus azarae	A.trivirgatus	783.60	17.40	1.67	[2]	1.31	[2]
Ateles fusciceps		9026.50	113.60	2.12	[3]	1.68	[2] (A.paniscus)
Bos taurus		596666.67	462.00	0.52	[2]	2.53	[2]
Callimico goeldii		492.20	10.95	1.43	[2]	1.25	[2]
Callithrix jacchus		288.12	7.61	1.43	[2]	1.17	[2]
Canis lupus	C.latrans	10750.00	86.23	1.43	[2]	1.80	[2]
Castor canadensis		21750.00	41.17	0.43	[2]	1.02	[2]
Cebus apella		2589.00	71.30	3.07	[4],[5],[1]	1.60	[5]
Cebus capucinus		1879.00	70.21	3.75	[3],[6]	1.69	[7] (C.albifrons)
Cercocebus atys		3792.86	100.80	3.36	[6],[1]	1.84	[5]
Cercopithecus cephus		1915.00	57.50	3.03	[6]		
Cercopithecus mitis		5041.29	67.00	1.85	[2]	1.78	[2]
Chlorocebus aethiops		3452.67	64.13	2.28	[4],[1]		
Chlorocebus sabeus		3042.80	73.61	2.84	[6],[1]		
Colobus guereza		10281.25	83.90	1.43	[8]		
Dasypus novemcinctus		3762.00	10.75	0.36	[2]	1.07	[2]
Delphinapterus leucas		636000.00	2083.00	2.24	[9]		
Elephantulus edwardii	E.fuscipes	57.00	1.33	0.74	[2]	1.00	[2]
Equus ferus	E.caballus	367000.00	712.00	1.11	[2]	2.80	[2]
Erinaceus europaeus		801.00	3.50	0.33	[2]	1.00	[2]
Erythrocebus patas		7421.40	105.65	2.25	[2]	1.91	[2]
Felis catus		3183.33	31.18	1.17	[2]	1.50	[2]
Gorilla gorilla		99648.40	477.44	1.78	[2]	2.26	[2]
Heterocephalus glaber		41.67	0.46	0.31	[10]		
Homo sapiens		61770.32	1300.00	6.68	[2]	2.56	[2]
Hylobates agilis		5528.75	88.10	2.28	[6]		
Lagothrix cana	L.lagothricha	5959.00	95.58	2.35	[2]	1.97	[2]
Lipotes vexillifer		180000.00	558.00	1.40	[9]		
Loxodonta africana		3775000.00	5253.56	1.72	[2]	3.84	[2]
Macaca arctoides		7630.00	100.70	2.10	[4]		
Macaca fascicularis		3109.45	66.93	2.55	[4],[6],[1]	1.65	[11]
Macaca leonia		2050.00	90.00	4.53	[1]		
Macaca mulatta		7102.83	89.22	1.95	[2]	1.75	[5]
Macaca nemestrina		4456.00	110.00	3.29	[12],[1]		
Macaca sylvanus		11200.00	87.70	1.42	[1]		
Macropus eugenii		6500.00	23.70	0.55	[2]	1.13	[2]
Mandrillus sphinx		12125.00	159.40	2.44	[2]	2.14	[2]
Microcebus murinus		71.83	1.85	0.88	[2]	1.10	[2]
Mus musculus		22.25	0.55	0.57	[2]	1.03	[2]
Mustela putorius		809.00	8.25	0.77	[2]	1.63	[13]
Odocoileus virginianus		87000.00	160.00	0.65	[2]	2.27	[2]
Orcinus orca		2049000.00	5617.00	2.76	[9]		
Oryctolagus cuniculus		2000.00	6.50	0.33	[2]	1.15	[2]
Otolemur garnettii	O.crassicaudatus	952.43	10.60	0.89	[2]	1.25	[2]
Ovis aries		63966.67	125.00	0.63	[2]	1.94	[2]
Pan paniscus		39700.00	329.70	2.28	[5],[8]	2.17	[5]
Pan troglodytes		41057.09	392.06	2.65	[2]	2.46	[2]
Papio anubis		13829.29	179.10	2.51	[4],[6],[8],[14]	2.00	[7]
Papio hamadryas		16000.00	182.00	2.31	[2]	1.99	[2]
Phoca vitulina		115000.00	273.75	0.93	[2]	2.38	[2]
Piliocolobus tephrosceles	P.badius	7854.83	76.75	1.57	[2]	1.80	[2]
Pongo abelii	P.pygmaeus	42447.67	347.75	2.30	[2]	2.21	[2]
Procavia capensis		3420.00	19.17	0.68	[2]	1.37	[2]
Propithecus coquereli	P.verreauxi	3547.12	26.90	0.94	[2]	1.34	[2]
Pteropus vampyrus	P.giganteus	1038.75	9.00	0.71	[2]	1.25	[2]
Pygathrix nemaeus		8481.67	84.83	1.65	[2]	1.64	[2]
Rattus norvegicus		314.67	2.41	0.43	[2]	1.02	[2]
Saguinus oedipus		368.83	9.80	1.56	[2]	1.20	[2]
Saimiri boliviensis		750.00	24.10	2.38	[8]		
Saimiri sciureus		680.60	22.98	2.42	[2]	1.55	[2]
Sarcophilus harrisii			15.00		[7]	1.33	[13]
Semnopithecus entellus		7010.00	111.50	2.46	[1]		
Sorex araneus		9.00	0.20	0.38	[2]	1.00	[2]
Sus scrofa		133600.00	137.65	0.42	[2]	2.16	[2]
Symphalangus syndactylus		12172.00	134.80	2.06	[6],[8]		
Tarsius syrichta		112.05	3.83	1.35	[2]	1.10	[2]
Theropithecus gelada		7710.00	130.00	2.69	[6]		
Trichechus manatus		797000.00	382.00	0.35	[2]	1.02	[2]
Tupaia belangeri	T.glis	173.33	3.03	0.80	[2]	1.06	[2]
Tursiops truncatus		177500.00	1489.00	3.77	[2]	4.76	[2]
Zalophus californianus		140000.00	363.00	1.08	[2]	2.52	[2]

Supplementary Table 8. Comparison between PAML[15] site models for TRNP1 protein-coding sequence using Likelihood Ratio Test (LRT). Maximum likelihood of M8 is significantly higher than that of M7, suggesting that a proportion (p_1=8.2%) of amino acid sites in TRNP1 evolve under positive selection

Model	Parameters	lnL	2(lnL(M8)-lnL(M7))	Df	χ^2 p-value
M8 (beta and ω)	$p_0(p_1 = 1 - p_0)$, $p, q, \omega_s > 1$	-5438.53	17.34	2	< 0.001
M7 (beta)	p, q	-5447.20			

Supplementary Table 9. TRNP1 amino acid sites under positive selection across the phylogeny according to Naive Empirical Bayes analysis (*: P>95%; **: P>99%)

Alignment position	Pr($\omega > 1$)	post mean ω
35	0.977*	1.11
41	0.994**	1.12
62	0.991**	1.12
63	0.971*	1.10
86	0.984*	1.12
193	0.967*	1.10

Supplementary Table 10. Pairwise correlations between the substitution rates of TRNP1 (dS: synonymous substitution rates, ω: the ratio of the non-synonymous over the synonymous substitution rates) and the rate of change in either GI, brain size or body mass estimated separately across 31 mammalian species using Coevol[16]. Partial correlations are the maximally controlled correlations, controlling for all other included variables

Parameter 1	Parameter 2	Marginal Correlation	Partial Correlation
		(Posterior Probability)	(Posterior Probability)
GI	ω	0.62 (0.95)	0.745 (0.98)
ω	dS	-0.044 (0.55)	0.291 (0.71)
GI	dS	-0.404 (0.97)	-0.411 (0.82)
brain mass	ω	0.499 (0.89)	0.667 (0.96)
ω	dS	-0.031 (0.52)	0.334 (0.74)
brain mass	dS	-0.524 (0.99)	-0.516 (0.88)
body mass	ω	0.44 (0.85)	0.587 (0.92)
ω	dS	0.008 (0.51)	0.27 (0.7)
body mass	dS	-0.436 (0.97)	-0.425 (0.86)

Supplementary Table 11. Pairwise correlations between the substitution rates of TRNP1 and the rate of change in the three morphological traits (GI, brain size, body mass) across 31 mammalian species, all estimated together in a joint framework using Coevol[16]

Parameter 1	Parameter 2	Marginal Correlation	Partial Correlation
		(Posterior Probability)	(Posterior Probability)
GI	ω	0.69 (0.98)	0.474 (0.87)
brain size	ω	0.638 (0.93)	0.273 (0.75)
body mass	ω	0.553 (0.90)	0.035 (0.51)
ω	dS	-0.242 (0.74)	0.192 (0.66)
GI	dS	-0.41 (0.97)	0.016 (0.53)
body mass	dS	-0.453 (0.98)	0.143 (0.68)
brain size	dS	-0.551 (0.99)	-0.332 (0.85)
GI	brain size	0.817 (1.00)	0.354 (0.85)
brain size	body mass	0.909 (1.00)	0.656 (0.95)
GI	body mass	0.681 (1.00)	-0.196 (0.76)

Supplementary Table 12. Model selection results between logistic regression models that predict the proportion of proliferating mouse NSCs in the presence of TRNP1 compared to a GFP control (LRT). n=donor mouse (batch). Proliferation is best predicted by the presence of TRNP1 (yes/no) together with the donor mouse to correct for the batch

Model	Resid. Df	Resid. Dev	Df	Deviance	Pr(>Chi)
0: prolif(%) ∼ 1	68	1491.35			
1: prolif(%) ∼ TRNP1	67	1258.02	1	233.33	1.1E-52
2: prolif(%) ∼ TRNP1+n	56	235.84	11	1022.18	3.2E-212

Supplementary Table 13. Model selection results between logistic regression models that predict the proportion of proliferating mouse NSCs in the presence of different TRNP1 orthologues (LRT). n=donor mouse (batch). Proliferation is best predicted by the respective TRNP1 orthologue together with the donor mouse to correct for the batch

Model	Resid. Df	Resid. Dev	Df	Deviance	Pr(>Chi)
0: prolif(%) ∼ 1	56	1067.98			
1: prolif(%) ∼ orthologue	51	968.08	5	99.91	5.5E-20
2: prolif(%) ∼ orthologue+n	40	121.49	11	846.58	1.9E-174

Supplementary Table 14. The induced NSC proliferation rates by the different TRNP1 orthologues (according to model 2 from Suppl. Table 13)

Species	Proliferation rate	Std. Error
macaque	0.49	0.028
galago	0.51	0.027
ferret	0.52	0.023
mouse	0.54	0.022
human	0.58	0.022
dolphin	0.65	0.025

Supplementary Table 15. Pairwise proliferation rate comparison between the TRNP1 orthologues of interest (Tukey test)

Comparison	Estimate	Std. Error	z value	Pr(>|z|)
human - mouse	0.15	0.069	2.188	0.0989
dolphin - human	0.30	0.089	3.378	0.0028
human - macaque	0.34	0.085	3.972	0.0003
human - galago	0.26	0.085	3.071	0.0082

Supplementary Table 16. PGLS model selection using LRT to test whether CRE activity of the 7 *TRNP1* regulatory regions is predictive for either brain size or gyrification (GI) across species. The reduced model contains intercept as the only predictor

Model	Value	Std.Error	df	L.Ratio	LRT p-value
log2(brain size) ∼ log2(upstream1)	0.01	0.152	1	0.01	0.929
log2(brain size) ∼ log2(upstream2)	-0.28	0.234	1	1.43	0.232
log2(brain size) ∼ log2(upstream3)	-0.30	0.223	1	1.92	0.166
log2(brain size) ∼ log2(exon1)	0.31	0.487	1	0.42	0.517
log2(brain size) ∼ log2(intron)	0.35	0.399	1	0.78	0.377
log2(brain size) ∼ log2(exon2)	0.28	0.286	1	0.97	0.324
log2(brain size) ∼ log2(downstream)	0.18	0.474	1	0.16	0.693
log2(GI) ∼ log2(upstream1)	0.01	0.073	1	0.01	0.929
log2(GI) ∼ log2(upstream2)	-0.04	0.060	1	0.53	0.468
log2(GI) ∼ log2(upstream3)	-0.06	0.048	1	1.51	0.219
log2(GI) ∼ log2(exon1)	0.07	0.088	1	0.62	0.431
log2(GI) ∼ log2(intron)	0.14	0.086	1	2.75	0.097
log2(GI) ∼ log2(exon2)	0.10	0.086	1	1.33	0.250
log2(GI) ∼ log2(downstream)	0.03	0.114	1	0.07	0.787

Supplementary Table 17. PGLS model selection using LRT to test whether the association between GI and intron CRE activity on the Old World monkey and great ape branch is consistent across three independent cell lines. The reduced model contains intercept as the only predictor

Model	Value	Std.Error	df	L.Ratio	LRT p-value	Cell line
log2(GI) $\sim$ log2(intron)	0.20	0.059	1	8.66	0.003	human1
log2(GI) $\sim$ log2(intron)	0.14	0.071	1	3.81	0.051	human2
log2(GI) $\sim$ log2(intron)	0.10	0.059	1	3.31	0.069	macaque

Supplementary Table 18. Enriched Gene Ontology Terms (Fisher's p-value< 0.05) of the 22 TFs with binding site enrichment on the intron CRE sequences from the 10 catarrhine species. Background: all expressed TFs included in the motif binding enrichment analysis (392)

GO.ID	Term	Annotated	Significant	Expected	Fisher's P
GO:0010817	regulation of hormone levels	26	5	1.47	0.011
GO:0042127	regulation of cell population proliferation	117	12	6.60	0.012
GO:0008285	negative regulation of cell population proliferation	61	8	3.44	0.012
GO:0043523	regulation of neuron apoptotic process	20	4	1.13	0.020
GO:1901615	organic hydroxy compound metabolic process	21	4	1.18	0.024
GO:0051402	neuron apoptotic process	23	4	1.30	0.033
GO:1903706	regulation of hemopoiesis	47	6	2.65	0.037
GO:0006325	chromatin organization	35	5	1.97	0.037
GO:0008283	cell population proliferation	135	12	7.62	0.039
GO:0090596	sensory organ morphogenesis	36	5	2.03	0.042
GO:0006259	DNA metabolic process	25	4	1.41	0.044

Supplementary Table 19. PGLS model selection using LRT where GI was predicted using either standardized TRNP1 protein evolution rates (ω) combined with standardized intron CRE activity or the standardized ω alone across the Old World monkey and great apes for which both measurements were available (n=9)

Model	Predictor	Value	Std.Error	df	logLik	L.Ratio	LRT p-value
log2(GI) $\sim$ log2(ω)	ω	0.19	0.041	3	11.27		
log2(GI) $\sim$ log2(ω) + log2(intron)	ω	0.16	0.029				
log2(GI) $\sim$ log2(ω) + log2(intron)	intron	0.05	0.015	4	15.66	8.78	0.003

1. Warnke, P. Mitteilung neuer Gehirn-und Körpergewichtsbestimmungen bei Saugern. *Psychol. Neurol* **13,** 355–403 (1908).
2. Lewitus, E. *et al.* An adaptive threshold in mammalian neocortical evolution. *PLoS biology* **12,** e1002000 (2014).
3. Crile, G. & Quiring, D. P. A record of the body weight and certain organ and gland weights of 3690 animals (1940).
4. Bronson, R. T. Brain weight-body weight relationships in 12 species of nonhuman primates. *Am. J. Phys. Anthropol.* **56,** 77–81 (1981).
5. Rilling, J. K. & Insel, T. R. The primate neocortex in comparative perspective using magnetic resonance imaging. en. *J. Hum. Evol.* **37,** 191–223 (1999).
6. Hrdlička, A. Weight of the brain and of the internal organs in American monkeys. With data on brain weight in other apes. *Am. J. Phys. Anthropol.* **8,** 201–211 (1925).
7. Zilles, K. *et al.* Gyrification in the cerebral cortex of primates. *Brain Behav. Evol.* **34,** 143–150 (1989).
8. Boddy, A. M. *et al.* Comparative analysis of encephalization in mammals reveals relaxed constraints on anthropoid primate and cetacean brain scaling. *J. Evol. Biol.* **25,** 981–994 (2012).
9. Manger, P. R. An examination of cetacean brain structure with a novel hypothesis correlating thermogenesis to the evolution of a big brain. en. *Biol. Rev. Camb. Philos. Soc.* **81,** 293–338 (2006).
10. Kverková, K. *et al.* Sociality does not drive the evolution of large brains in eusocial African mole-rats. en. *Sci. Rep.* **8,** 9203 (2018).
11. Ventura-Antunes, L. *et al.* Different scaling of white matter volume, cortical connectivity, and gyrification across rodent and primate brains. en. *Front. Neuroanat.* **7,** 3 (2013).
12. Spitzka, E. A. Brain-weights of animals with special reference to the weight of the brain in the Macaque monkey. *J. Comp. Neurol.* **13,** 9–17 (1903).
13. Brodmann, K. Neuere Forschungsergebnisse der Großhirnrindenanatomie mit besonderer Berücksichtigung anthropologischer Fragen. *Naturwissenschaften* **1,** 1120–1122 (1913).
14. Stephan, H. *et al.* New and revised data on volumes of brain structures in insectivores and primates. en. *Folia Primatol.* **35,** 1–29 (1981).
15. Yang, Z. PAML: a program package for phylogenetic analysis by maximum likelihood. *Bioinformatics* **13,** 555–556 (1997).
16. Lartillot, N. & Poujol, R. A phylogenetic model for investigating correlated evolution of substitution rates and continuous phenotypic characters. *Molecular biology and evolution* **28,** 729–744 (2010).