Evolution and cell-type specificity of human-specific genes preferentially expressed in progenitors of fetal neocortex

Marta Florio[1§#], Michael Heide[1§], Holger Brandl[1], Anneline Pinson[1], Sylke Winkler[1], Pauline Wimberger[3], Wieland B. Huttner[1]* and Michael Hiller[1,2]*

[1]Max Planck Institute of Molecular Cell Biology and Genetics, Pfotenhauerstr. 108, D-01307 Dresden, Germany

[2]Max Planck Institute for the Physics of Complex Systems, Nöthnitzer Straße 38. D-01187 Dresden, Germany

[3]Technische Universität Dresden, Universitätsklinikum Carl Gustav Carus, Klinik und Poliklinik für Frauenheilkunde und Geburtshilfe, Fetscherstraße 74, D-01307 Dresden, Germany

[§]Joint first authors

[#]Present address: Department of Genetics, Harvard Medical School, Boston, MA 02115, USA

*Corresponding authors:

huttner@mpi-cbg.de

hiller@mpi-cbg.de

Abstract

To understand the molecular basis underlying the expansion of the neocortex during primate, and notably human, evolution, it is essential to identify the genes that are particularly active in the neural stem and progenitor cells of developing neocortex. Here, we have used existing transcriptome datasets to carry out a comprehensive screen for protein-coding genes preferentially expressed in progenitors of fetal human neocortex. In addition to the previously studied gene *ARHGAP11B*, we show that ten known and two newly identified human-specific genes exhibit such expression, however with distinct neural progenitor cell-type specificity compared to their ancestral paralogs. Furthermore, we identify 41 additional human genes with progenitor-enriched expression which have orthologs only in primates. Our study not only provides a resource of genes that are candidates to exert specific, and novel, roles in neocortical development, but also reveals that distinct mechanisms gave rise to these genes during primate, and notably human, evolution.

Introduction

The expansion of the neocortex in the course of human evolution provides an essential basis for our cognitive abilities (Azevedo et al., 2009; Borrell and Reillo, 2012; Buckner and Krienen, 2013; Dehay et al., 2015; Florio and Huttner, 2014; Kaas, 2013; Lui et al., 2011; Namba and Huttner, 2017; Rakic, 2009; Sousa et al., 2017; Striedter, 2005). This expansion ultimately reflects an increase in the proliferative capacity of the neural stem and progenitor cells in the developing human neocortex (from now on collectively referred to as cortical neural progenitor cells, cNPCs) (Azevedo et al., 2009; Bae et al., 2015; Borrell and Reillo, 2012; Dehay et al., 2015; Florio and Huttner, 2014; Lui et al., 2011; Namba and Huttner, 2017; Rakic, 2009), as well as in the duration of their proliferative, neurogenic and gliogenic phases (Lewitus et al., 2014; Otani et al., 2016). It is therefore a fundamental task to elucidate the underlying molecular basis, that is, the changes in our genome that endow human cNPCs with these neocortical expansion-promoting properties.

One approach towards this goal is to identify which of the genes that are particularly active in human cNPCs exhibit a human-specific expression pattern, or even are human-specific. We previously isolated, and determined the transcriptomes of, two major cNPC types from embryonic mouse and fetal human neocortex (Florio et al., 2015), (i) the apical (or ventricular) radial glia (aRG), the primary neuroepithelial cell-derived apical progenitor type (Götz and Huttner, 2005; Kriegstein and Götz, 2003), and (ii) the basal (or outer) radial glia (bRG), the key type of basal progenitor implicated in neocortical expansion (Betizeau et al., 2013; Borrell and Götz, 2014; Borrell and Reillo, 2012; Florio and Huttner, 2014; Lui et al., 2011). This led to the identification of 263 protein-coding human genes that are much more highly expressed in human bRG and aRG than in a neuron-enriched fraction (Florio et al., 2015). Of these, 207 genes have orthologs in the mouse genome but are not expressed in mouse cNPCs, whereas 56 genes lack mouse orthologs. Among the latter, the gene with the highest specificity of expression in bRG and aRG was found to be *ARHGAP11B*, a human-specific gene (Antonacci et al., 2014; Dennis et al., 2017; Riley et al., 2002; Sudmant et al., 2010) that we showed to be capable of basal progenitor amplification in

69 embryonic mouse neocortex and to likely have contributed to the evolutionary
70 expansion of the human neocortex (Florio et al., 2015; Florio et al., 2016).
71
72 Our previous finding that, in addition to *ARHGAP11B*, there are 55 other human genes
73 without mouse orthologs that are predominantly expressed in bRG and aRG (Florio et
74 al., 2015), raises the possibility that some of these genes may be human-specific and
75 may affect the behaviour of human cNPCs. To investigate the evolution and cell-type
76 specificity of expression of such genes, we have now data-mined our previous dataset
77 (Florio et al., 2015) as well as three additional ones (Fietz et al., 2012; Miller et al.,
78 2014; Pollen et al., 2015) to carry out a comprehensive screen for protein-coding genes
79 preferentially expressed in cNPCs of fetal human neocortex. We find that, in addition to
80 *ARHGAP11B*, 12 other human-specific genes (10 previously and 2 newly identified
81 ones) show preferential expression in cNPCs. Furthermore, we identify 41 additional
82 human genes exhibiting such expression for which orthologs are found in primate but
83 not in non-primate mammalian genomes. We provide information on the evolutionary
84 mechanisms leading to the origin of several of these primate-specific genes, including
85 gene duplication and transposition. Moreover, we analyze the cell-type specific

86 expression of most of the human-specific genes, including their splice variant
87 expression patterns. Finally, by comparing the expression of the human-specific genes
88 with their respective ancestral paralog, we show a substantial degree of gene expression
89 divergence upon gene duplication, suggesting potential neofunctionalization. Our study
90 thus provides a resource of genes that are candidates to exert specific roles in the
91 development and evolution of the primate, and notably human, neocortex.
92

Results

Screen of distinct transcriptome datasets from fetal human neocortex for protein-coding genes preferentially expressed in neural stem and progenitor cells

To identify genes preferentially expressed in the cNPCs of the fetal human neocortex, we analyzed four distinct, published transcriptome datasets obtained from human neocortical tissue ranging from 13 to 19 weeks post conception (wpc). First, the genome-wide RNA-Seq dataset obtained from specific neocortical zones isolated by laser capture microdissection (LCM) (Fietz et al., 2012), which we screened for all protein-coding genes that are more highly expressed in the VZ, iSVZ and/or oSVZ than the cortical plate (CP) (as determined by differential gene expression (DGE) analysis, $p < 0.01$). This yielded 2758 genes (Fig. 1A, B). Second, the Allen Brain Institute microarray dataset obtained from LCM-isolated specific neocortical zones (Miller et al., 2014), which we screened for all protein-coding genes with positive laminar correlation (correlation coefficient >0.5) with either the VZ, iSVZ or oSVZ as compared to the zones enriched in postmitotic cells (intermediate zone (IZ), subplate, CP, marginal zone, subpial granular zone). This yielded 4555 genes (Fig. 1A, B). Third, the genome-wide

RNA-Seq dataset obtained from specific neocortical cell types isolated by fluorescence-activated cell sorting (FACS) (Florio et al., 2015), which we screened for all protein-coding genes more highly expressed (as determined by DGE analysis, $p<0.01$) in aRG and/or bRG in S-G2-M as compared to the cell population enriched in postmitotic neurons but also containing bRG in G1. This yielded 2106 genes (Fig. 1A, B). Fourth, the dataset obtained from genome-wide single-cell RNA-Seq of dissociated cells captured from microdissected VZ and SVZ (Pollen et al., 2015), which we screened for all protein-coding genes positively correlated with either radial glial cells, bIPs or both (correlation coefficient >0.1) and negatively correlated with neurons (correlation coefficient <0.1). This yielded 5335 genes (Fig. 1A, B).

Next, we determined how many of the protein-coding genes exhibiting the above described differential expression pattern were found in all four datasets. This was the case for 780 genes (Fig. 1C, red). We also determined the number of genes found in three of the four datasets (four combinations, Fig. 1C orange), and of those found in two datasets (six combinations, Fig. 1C pink). Together this yielded a catalogue of 3,722

126 human genes with preferential expression in cNPCs (from here on referred to as
127 cNPC-enriched genes) (see Table S1).
128
129 These 3,722 genes included well-known molecular players involved in cNPC function
130 and established markers of cNPCs, notably radial glia, (e.g. *EOMES*, *FABP7, GFAP,*
131 *HOPX, NES, PAX6, SOX2, VIM*), cell proliferation (e.g. *MKI67*, *PCNA*), Notch
132 signaling (*DLL1, HES5*), and extracellular matrix and growth factor signaling (e.g.
133 *FGFR3, ITGAV*, *LUM*, *TNC*) (listed in Fig. 1D). Moreover, several genes recently
134 implicated in human-specific aspects of cNPC proliferation and neocortex formation
135 (Florio et al., 2017; Mitchell and Silver, 2017; Sousa et al., 2017) (e.g. *ARHGAP11B,*
136 *FOXP2, FZD8, GPR56, PDGFD*) were found in the analyzed datasets, though not
137 necessarily in all four (Fig. 1D). The latter finding, on the one hand, likely reflects the
138 diversity of the cNPC enrichment strategies and mode of transcriptome analysis adopted
139 to obtain the four datasets, and on the other hand highlights the significance of
140 data-mining all these datasets in combination. On a general note, the catalog of 3,722
141 cNPC-enriched human genes presented here (Table S1) provides an integrative and
142 methodologically unbiased tool to interrogate the cNPC enrichment of candidate genes

of interest, and to potentially uncover new genes involved in cNPC function during fetal human corticogenesis.

Identification of primate-specific genes

Primate-specific, notably human-specific, genes expressed in cNPCs have gained increasing attention for their potential role in species-specific aspects of neocortical development, including neurogenesis (Charrier et al., 2012; Dennis et al., 2017; Florio et al., 2017; Sousa et al., 2017). To determine how many of the 3,722 human cNPC-enriched protein-coding genes had orthologs only in primates but not in non-primate species, we eliminated from this gene set all those genes with an annotated one-to-one ortholog in any of the sequenced non-primate genomes (Fig. 1E). This greatly reduced the number of genes from 3,722 to 83 genes.

Next, we examined these 83 genes to extract those that are truly primate-specific. By inspecting genomic alignments, gene neighborhoods and gene annotations in primate and non-primate mammals, we concluded that 29 of these genes likely have an ortholog in non-primate mammals and we therefore excluded them from further analysis. The

160 remaining 54 genes were considered to be truly primate-specific (Fig. 1E) and are of
161 special evolutionary interest because of their orthology to the human cNPC-enriched
162 protein-coding genes.
163
164 **Phylogenetic analysis of the primate-specific genes**
165 To trace the phylogeny of the 54 primate-specific genes and to infer their ancestry, we
166 investigated in which species these genes exhibit an intact reading frame, and used this
167 information to assign each gene to a primate clade. First, we found that 28 of these 54
168 genes exist in the genomes of tarsiers, monkeys and apes (though not necessarily in
169 every species of these clades) and thus likely predate the ape (Hominoidea) ancestor
170 that lived ~28 Mya (Kumar et al., 2017) (Fig. 2). Specifically, some of these 28 genes
171 can already be detected in haplorrhini, some in simiiformes and some at first in
172 catarrhini (Fig. 2). Strikingly, more than half of these 28 genes encode zinc finger
173 proteins (Table 1). Second, we found that of the remaining 26 primate-specific genes
174 that exist only in the genomes of apes (hominoidea), although not necessarily in every
175 ape species (Fig. 2), 13 genes exist not only in the human genome but also in
176 non-human ape genomes. It is interesting to note that 11 of these 13 genes exist only in

177 great apes (hominidae) and thus likely arose after the lineage split (~18 Mya, Kumar et
178 al., 2017) leading to the gibbon vs. the great apes, and 8 of these 11 genes do not exist
179 in the orangutan and thus likely arose after the great ape ancestor split into the
180 orangutan and Homininae lineage ~15 Mya (Kumar et al., 2017) (Fig. 2). Of the two
181 hominoidea-specific genes already detected in the gibbon, *PTTG2* deserves special
182 comment and is further discussed below.
183
184 Finally, we found that 13 of the 54 primate-specific genes were only present in the
185 human genome, and thus arose (or evolved to their present state) in the human lineage
186 after its split from the lineage leading to the chimpanzee (Fig. 2) (~5-7 Mya, Brunet et
187 al., 2005; Brunet et al., 2002; Vignaud et al., 2002). These 13 human-specific genes
188 include *ARHGAP11B*, a gene that we reported previously to have a key role in cNPC
189 proliferation and neocortex expansion (Florio et al., 2015; Florio et al., 2016) and that
190 was also present in the archaic genomes of Neandertals and Denisovans (Antonacci et
191 al., 2014; Florio et al., 2015; Meyer et al., 2012; Prüfer et al., 2014; Sudmant et al.,
192 2010). Similar to *ARHGAP11B*, 11 of the remaining 12 human-specific genes existed
193 also in the genomes of Neandertals and Denisovans (Dennis et al., 2017; Sudmant et al.,

194 2010) and present data, see Table 1) and thus arose before the split of the lineages
195 leading to modern humans vs. Neandertals/Denisovans ~500,000 years ago (Meyer et
196 al., 2012; Prüfer et al., 2014). Of note, *SMN2* is the only gene in this set that has been
197 reported to have arisen in the lineage leading to modern humans after its divergence
198 from the lineage leading to Neandertals and Denisovans (Dennis et al., 2017).
199
200 **Analysis of the evolution of selected primate-specific genes reveals distinct**
201 **mechanisms**
202 Next, we sought to determine how these 54 primate-specific genes evolved. With regard
203 to the primate-specific genes that are not human-specific, we focused on three genes
204 that we selected in light of their potential biological role, *MICA*, *KIF4B* and *PTTG2*.
205
206 *MICA* (*MHC class I polypeptide-related sequence A*) is a paradigmatic example of a
207 gene arising by gene duplication (Bailey et al., 2002; Eichler et al., 2004; Fortna et al.,
208 2004; Hurles, 2004), a well-known driving force of genome evolution (Lynch and
209 Conery, 2000). *MICA* arose by duplication of the widely occurring *MICB* gene. As
210 *MICA* is found in the genomes of apes and Old-World monkeys (Catarrhini), but not

New-World monkeys (Fig. 2), this gene duplication presumably occurred after the separation of the lineages leading to New-World monkeys vs. and Catarrhini ~47 Mya (Kumar et al., 2017). *MICA* is the only gene among the 54 primate-specific genes analyzed in the present study that has an established relationship to the MHC locus (Bahram et al., 1994), pointing to a possible primate-specific interaction between cNPCs and cells of the immune system.

Besides gene duplication, however, other mechanisms were found to underlie the evolution of primate-specific genes. A notable example is *KIF4B* (*Kinesin Family Member 4B)*, a gene encoding a kinesin involved in spindle organization during cytokinesis (Zhu et al., 2005). In fact, *KIF4B* is the only member of the kinesin superfamily among the 54 primate-specific genes. *KIF4B* is specific to apes, Old-World monkeys and New-World monkeys (Simiiformes, Fig. 2) and evolved by retroposition of *KIF4A*, a gene with a near-ubiquitous occurrence in the animal kingdom (Hirokawa et al., 2009). This retroposition involved the reverse transcription of a spliced *KIF4A* mRNA followed by insertion of the DNA into the genome as an intronless copy of *KIF4A*.

Similar to *KIF4B*, the primate-specific gene *PTTG2* (*pituitary tumor transforming 2*) arose by retroposition. Specifically, *PTTG2* arose by reverse transcription of the spliced mRNA of *PTTG1*, a gene encompassing five protein-coding exons conserved in reptiles, birds and mammals and implicated in promoting proliferation of pituitary tumor cells (Dominguez et al., 1998; Vlotides et al., 2007; Zhang et al., 1999). However, while *KIF4B* inserted into an intergenic locus (that however allowed its transcription), the intron-less protein-coding *PTTG2* inserted into intron 2 of the *TBC1D1* gene (Fig. 3A), which encodes a Rab-GTPase activating protein (Roach et al., 2007). This *PTTG2* retroposition event presumably occurred in the ancestor of New-World monkeys, Old-World monkeys and apes (simiiformes). Remarkably, after retroposition, the *PTTG2* gene underwent two principally different lines of evolution. In all non-hominoidea simiiformes (see Fig. 2), consistent with neutral evolution, *PTTG2* accumulated frameshifting deletions and translational stop codon mutations that cause premature termination of the open reading frame (Fig. 3B). In contrast, in hominoidea (apes and humans, see Fig. 2), the *PTTG2* reading frame remained open, with one noticeable change. This is a 1-bp insertion (T, see Fig. 3A) near the 3' end of the *PTTG2* open reading frame that causes a shift in the reading frame, resulting in a new

13-amino acid-long C-terminal sequence of PTTG2 in great apes (including human) (as opposed to 24 amino acids in PTTG1) (Fig. 3B). This PTTG2-specific sequence lacks the cluster of acidic residues found in the C-terminal sequence of PTTG1. In the case of the gibbon, however, the *PTTG2* gene carries (in addition to the 1-bp T insertion) a 22-bp deletion a few nucleotides 5' to this insertion. This causes yet another shift in the reading frame that results in the replacement of the C-terminal 25 amino acids of the PTTG2 of great apes (including human) by an 18-amino acid-long sequence (Fig. 3B). The potential consequences of these changes in protein sequence for the function of PTTG2 with regard to cell proliferation are discussed below.

A variety of evolutionary mechanisms gave rise to the human-specific cNPC-enriched protein-coding genes

We next investigated how the 13 human-specific cNPC-enriched protein-coding genes evolved. Nine of them arose by duplications of entire genes (Bailey et al., 2002; Eichler et al., 2004; Fortna et al., 2004; Hurles, 2004) (Figure 4A). Two arose by partial gene duplication. These genes are *ARHGAP11B* and *NOTCH2NL* (Antonacci et al., 2014; Dennis et al., 2017; Dougherty et al., 2017; Riley et al., 2002). *ARHGAP11B* arose from

partial duplication of *ARHGAP11A*, which encodes a Rho GTPase activating protein (RhoGAP). *ARHGAP11B* comprises the entire GAP domain of ARHGAP11A but, due to a single-base substitution (C–>G) creating a new splice donor site, encodes a protein with a truncated GAP domain followed by a unique, human-specific C-terminal amino acid sequence (Florio et al., 2015; Florio et al., 2016) (Fig. 4B).

The human-specific *NOTCH2NL* identified here as a cNPC-enriched gene arose from partial duplication of the *NOTCH2* gene (Fig. 4A) and comprises only those regions of the *NOTCH2* gene that give rise to a short *NOTCH2* splice variant (Ensembl transcript ENST00000602566.5). Similar to this short NOTCH2 isoform, *NOTCH2NL* encodes only a short segment of the NOTCH2 ectodomain. However, the protein encoded by the *NOTCH2NL* studied here is predicted to lack a signal peptide, which raises the issue of whether NOTCH2NL is secreted, and if so, via which pathway. Irrespective of this open question, we explored, in light of the importance of Notch signaling for cNPC behaviour (Imayoshi et al., 2013; Kawaguchi et al., 2008; Lui et al., 2011; Pierfelice et al., 2008; Wilkinson et al., 2013), a potential role of the *NOTCH2NL* mRNA and/or the NOTCH2NL protein in cNPCs by in utero electroproration of *NOTCH2NL* under the

279 control of a constitutive promoter into neocortical aRG of embryonic day (E) 13.5
280 mouse embryos. Analysis by Ki67 immunofluorescence 48 hours after *NOTCH2NL*
281 electroporation revealed an increase in cycling basal progenitors in the SVZ and IZ, but
282 not in apical progenitors in the VZ (Fig. S1A, B). This finding was further corroborated
283 by analysis of mitotic cNPCs using phosphohistone H3 immunofluorescence, which
284 showed an increase in abventricular, but not ventricular, mitoses (Fig. S1C-E). Thus,
285 forced expression of the human-specific *NOTCH2NL* gene in mouse embryonic
286 neocortex appears to promote basal progenitor proliferation.
287
288 The remaining two human-specific cNPC-enriched protein-coding genes evolved in
289 distinct ways. The *ZNF492* gene as such exists in the genomes of all non-human great
290 apes. In the case of human, however, an exon of another zinc finger protein-encoding
291 gene, *ZNF98*, inserted into the *ZNF492* locus, yielding a chimeric human-specific
292 protein containing the repressor domain of ZNF492 and the DNA binding domain of
293 ZNF98 (Fig. 4C). The *FAM182B* gene as such exists not only in human but also in
294 chimpanzee, bonobo and gorilla. However, in bonobo and gorilla, a stop codon
295 terminates the potential open reading frame soon after the initiator methionine, whereas

296 in human a single T–>G substitution abolishes this stop codon and rescues the open
297 reading frame to yield a 152-amino acid-long protein (Fig. 4D). In chimpanzee, the
298 corresponding T is missing, resulting in a reading frame shift that predicts a shorter,
299 52-amino acid-long polypeptide. Taken together, we conclude that the human-specific
300 cNPC-enriched protein-coding genes evolved by a variety of evolutionary mechanisms.
301
302 We sought to corroborate that the human-specific cNPC-enriched protein-coding genes
303 arising from complete or partial gene duplication indeed constitute additional gene
304 copies (rather than reflecting the inability of distinguishing multiple gene copies in the
305 genomes of the other great apes due to genome assembly issues). To this end, we used a
306 quantitative genomic PCR approach. The idea was that primers targeting genomic
307 regions that are identical in human, chimpanzee and bonobo and thus should amplify
308 genomic DNA of the three species proportionally to the copy number of each gene in
309 each species. As a proof of principle, we validated the known human-specific nature of
310 the partially duplicated *ARHGAP11B* by designing primers to the regions that are
311 identical between *ARHGAP11A* and *ARHGAP11B*. Using the bonobo gene as the
312 standard, this resulted in a two-fold increase of the human PCR product compared to the

313 bonobo and chimpanzee, corroborating that *ARHGAP11B* is indeed a human-specific
314 partial gene duplication (Fig. 4E).
315
316 We used the same approach to validate the human-specific genes arising from complete
317 gene duplication. For four of these nine genes (*ANKRD20A2, ANKRD20A4, CBWD5,*
318 *DHRS4L2*) and for *NOTCH2NL*, we could not design primers that uniquely target these
319 genes as the respective genomic loci are not well resolved in the non-human great ape
320 genomes. Thus, the final validation of these putative human-specific genes awaits
321 improved genome assemblies. For the other five human-specific cNPC-enriched genes
322 (*FAM72B/C/D*, *GTF2H2C*, *SMN2*) and for the human-specific gene *GTF2H2B* for
323 which primers could be designed, genomic qPCR resulted in an estimated four human
324 copies of *FAM72*, three human copies of *GTF2H2* and two human copies of *SMN* (Fig.
325 4E), compared to only one copy in both chimpanzee and bonobo. This validated the
326 human-specific nature of these genes.
327

Spatial mRNA expression analysis in fetal human neocortex of the human-specific cNPC-enriched protein-coding genes and of three selected primate- but not human-specific protein-coding genes

Given that the 13 human-specific genes had emerged from a screen for cNPC-enriched genes, it was of interest to examine their spatial expression pattern in the various zones of the fetal human cortical wall. To this end, we performed in-situ hybridization (ISH) on 13 wpc human neocortex to determine the localization of their mRNAs. Depending on the gene under study, this analysis detected expression of either the human-specific gene only, or (if the ISH probe used could not distinguish between paralogs due to their sequence similarity) the mRNA of the ancestral paralog from which the duplication arose and – if existing – the mRNAs of yet other paralogs. Of note, for *ARHGAP11B*, we used a specific Locked Nucleic Acid (LNA) probe, which enabled us to distinguish the mRNA of *ARHGAP11B* from that of *ARHGAP11A* (Fig. S2). We could distinguish five types of gene expression patterns.

First, referred to as "VZ", mRNA expression essentially confined to the VZ, which was the case for *ANKRD20A1-4* and *NOTCH2NL* (Fig. 5A-B'"). Second, referred to as "VZ

345 + iSVZ + oSVZ", mRNA expression in all three germinal zones but not in the CP,
346 which was the case for *ARHGAP11B* (Fig. 5C-C'"). Third, referred to as "VZ, iSVZ >
347 oSVZ, CP", mRNA expression in all zones, however with markedly stronger staining in
348 the VZ and iSVZ than in the oSVZ and CP, which was the case for *DHRS4L1-2*,
349 *FAM72A-D* and *ZNF492* (Fig. 5D-F'"). Fourth, referred to as "VZ > CP > iSVZ, oSVZ",
350 mRNA expression in all zones, however with markedly stronger staining in the VZ and
351 CP than in the iSVZ and oSVZ, which was the case for *GTF2H2A-C* (Fig. 5J-J'"). Fifth,
352 "referred to as VZ, CP > iSVZ, oSVZ", strong mRNA expression in the VZ and CP and
353 lower mRNA expression in the iSVZ and oSVZ, which was the case for *CBWD1-7*,
354 *FAM182A-B* and *SMN1-2* (Fig. 5G-I'").
355
356 The strong ISH signal in the CP observed for the latter genes is nonetheless consistent
357 with our conclusion, based on our approach of gene identification (Fig. 1), that the
358 human-specific paralogs show enriched expression in cNPCs. Indeed, the published
359 RNA-Seq data from specific LCM-isolated neocortical zones (Fietz et al., 2012)
360 confirmed the ISH pattern for *CBWD1-7*, *FAM182A-B* and *SMN1-2* in that the sum of

their mRNA levels in the three germinal zones (VZ, iSVZ, oSVZ) was greater than the mRNA level in the CP.

We also examined by ISH the spatial expression pattern in the fetal human cortical wall of the three primate-specific genes *PTTG2*, *MICA* and *KIF4B*. Due to the high degree of similarity in nucleotide sequence this analysis also included the mRNA of the respective ancestral paralog. mRNA expression for *PTTG1/2* (Fig. 6A), *MICA/B* (Fig. 6B) and *KIF4A/B* (Fig. 6C) was robust in the human VZ and iSVZ, relatively low in the oSVZ, and moderate in the CP.

Cell type-specific expression patterns of the human-specific cNPC-enriched protein-coding genes compared to the corresponding ancestral paralogs

Complete or partial gene duplications often encompass the regulatory elements that control gene expression (Bailey et al., 2002; Eichler et al., 2004; Fortna et al., 2004; Hurles, 2004). This raises the question whether the human-specific cNPC-enriched protein-coding genes identified here exhibit similar cell-type expression patterns as their

377 respective ancestral paralogs, or whether expression differences have evolved during
378 human evolution.

379

380 Given that we could not distinguish by ISH most of the human-specific genes from their
381 respective ancestral paralog, we sought an additional approach to gain insight into the
382 cell-type specificity of expression of the human-specific genes. Specifically, we used
383 our previously reported cell-type-specific gene expression data from the human aRG
384 population (aRG), the bRG population (bRG) and the neuron fraction (N) (Florio et al.,
385 2015) and re-analyzed these data using Kallisto. Kallisto is a probabilistic algorithm to
386 estimate absolute transcript abundance, which has been proven to be accurate in
387 assigning RNA-Seq reads to specific transcripts, including those originating from
388 highly similar paralog genes (Bray et al., 2016). We could confidently ascertain
389 cell-type-specific mRNA expression profiles for 11 of the 13 human-specific genes and
390 their corresponding ancestral paralog (Fig. 7, Table S4).

391

392 We first focused on changes in the total mRNA levels between the human-specific
393 genes and their ancestral paralogs. With the exception of *CBWD5* and *NOTCH2NL,* for

394 which the total mRNA levels in aRG, bRG and N were in the same range as the
395 corresponding ancestral paralog, we found that the majority of the human-specific genes
396 showed markedly different mRNA expression levels compared to their ancestral
397 paralog, which were either reduced (*ARHGAP11B*, *FAM72B/C/D*, *GTF2H2C*) or
398 increased (*ANKRD20A2*, *ANKRD20A4*, *DHRS4L2*, *ZNF492*) (Fig. 7A, B). This reflects
399 either changes in mRNA expression levels per cell, changes in the proportions of
400 mRNA-expressing cells, or both. Irrespective of which is the case, this finding indicates
401 that expression of these human-specific genes has indeed changed compared to their
402 ancestral paralogs during human evolution – a potential indication of
403 neofunctionalization.
404
405 Next, we asked whether the human-specific genes diverged in their pattern of
406 expression in aRG vs. bRG vs. N from that of their ancestral paralogs. For five of the
407 human-specific genes (*CBWD5*, *FAM72B/C/D* and *NOTCH2NL*), the pattern of relative
408 mRNA levels in these cell types was similar to that of the respective ancestral paralog
409 (Fig. 7A). In the case of the other six human-specific genes, we observed, relative to the
410 mRNA level in aRG, either decreases (*ZNF492*) or increases (*ARHGAP11B*, *DHRS4L2*,

411 *GTF2H2C*) in the bRG mRNA level compared to respective ancestral paralog. We also
412 observed decreases (*ANKRD20A2*, *ANKRD20A4*) or increases (*DHRS4L2*, *GTF2H2C*)
413 in the N fraction mRNA level compared to the respective ancestral paralog (Fig. 7A).
414 Of note, the increase in the *ARHGAP11B* mRNA level in bRG as compared to aRG is
415 consistent with the previously reported function of this gene in basal progenitor
416 amplification (Florio et al., 2015; Florio et al., 2016). These findings suggest that these
417 six human-specific genes underwent changes in regulatory elements at the
418 transcriptional and/or post-transcriptional level.
419
420 To complement these data, we performed a second type of analysis. We identified
421 paralog-specific sequencing reads (Suppl. Files S1-S8; see Suppl. Fig. S3A for
422 illustration of a hypothetical example) using our previously reported RNA-Seq dataset
423 (Florio et al., 2015), and then determined the number of paralog-specific sequencing
424 reads for the 11 human-specific genes and their corresponding ancestral paralog in aRG,
425 bRG and N (Suppl. Fig. S3B). This analysis largely corroborated the results shown in
426 Fig. 7A, further pointing to expression changes in aRG vs. bRG vs. N for the
427 human-specific genes in comparison to their corresponding ancestral paralogs.

We finally explored the complexity in cell-type-specific expression patterns by examining the differential mRNA expression of protein-coding splice variants of the human-specific genes. Specifically, we analyzed our aRG vs. bRG vs. N RNA-Seq data (Florio et al., 2015) for cell-type-specific gene expression and relative abundance of sequencing reads diagnostic of specific protein-coding splice variants of 10 of the 11 human-specific cNPC-enriched genes shown in Fig. 7 (Suppl. Fig. S4). This showed, for most of these human-specific genes (*ANKRD20A2*, *ANKRD20A4*, *CBWD5*, *FAM72B/C*, *DHRS4L2*, *GTF2H2C*, *NOTCH2NL*), the preferential expression of certain splice variants. Moreover, this analysis revealed splice variants with preferential expression in either aRG, bRG or N for some of these human-specific genes (e.g., *CBWD5*, *GTF2H2C, ARHGAP11B*). A notable case was *ARHGAP11B*, of which one splice variant (Ensembl transcript ENST00000428041.2), endowed with a shorter 3'-UTR, was exclusively expressed in bRG whereas the other splice variant was enriched in aRG (Suppl. Fig. S4).

In summary, these analyses show that after duplication, the expression pattern of most of the resulting new, human-specific cNPC-enriched protein-coding genes evolved

445 differences in both the levels and cell-type specificity of their mRNAs compared to their

446 respective ancestral paralog.

447

448 **Discussion**

449

450 Our study not only provides a resource of genes that are candidates to exert specific
451 roles in the development and evolution of the primate, and notably human, neocortex,
452 but also has implications regarding (i) the emergence of these genes during primate
453 evolution and (ii) the maintenance vs. modification of the cell-type specificity of their
454 expression. As to their emergence during primate evolution, two aspects of our findings
455 deserve comment. First, while entire or partial gene duplications were the underlying
456 mechanism that gave rise to the majority of the human-specific cNPC-enriched
457 protein-coding genes, as noted previously (Bailey et al., 2002; Eichler et al., 2004;
458 Fortna et al., 2004; Hurles, 2004), our data reveal also other mechanisms of gene
459 evolution such as exon duplication and replacement (*ZNF492*) and translational stop
460 codon removal (*FAM182B*).

461

462 The latter notion is further underscored by our observation that of the three primate- but
463 not human-specific genes studied here in greater detail, two (*KIF4B*, *PTTG2*) arose by
464 retroposition (Brosius, 1991; Long et al., 2003; Marques et al., 2005) rather than gene

duplication. Here, *PTTG2* is a particularly interesting case in that its initially presumably open reading frame became closed during the evolution of non-hominoidea simiiformes but remained open during the evolution of hominoidea. This suggests that the functional role of PTTG2 may be essential for the development of the neocortex of apes and human but not for that of New-World and Old-World monkeys. Given the expression of *PTTG2* in the germinal zones of fetal human neocortex and the fact that this gene is derived from *PTTG1*, which encodes a protein exhibiting tumorigenic activity (Vlotides et al., 2007), it appears possible that PTTG2 may function to amplify cNPCs.

Second, of the 54 human genes that we identified in the present study as being primate-specific, as many as 13 (i.e. almost one quarter) are human-specific. This is a far greater percentage than would be expected if the former genes arose by a constant rate during primate evolution to modern humans. This in turn suggests that the latter cNPC-enriched protein-coding genes conveyed a selection advantage specifically during the evolution of the human neocortex.

481 As to the issue of maintenance vs. modification of the cell-type specificity of expression
482 of the human-specific genes, it is striking to observe that the majority of these genes,
483 although arising by entire or partial gene duplications, show marked differences not
484 only in the level but also in the cNPC-type specificity of their mRNA expression
485 compared to their ancestral paralog. For several of the human-specific genes, the
486 corresponding spatial characteristics of their mRNA expression in the neocortical
487 germinal zones could be corroborated by specific ISH. These data suggest that during
488 human evolution these genes underwent specific changes in regulatory elements at the
489 transcriptional and/or post-transcriptional level. This in turn raises the possibility that
490 (at least some of) the human-specific genes characterized in the present study may be
491 candidates to have contributed to the evolution of human-specific features of
492 neocortical development.
493
494 In line with the latter consideration, we found that expressing the human-specific gene
495 *NOTCH2NL* studied here in mouse embryonic neocortex increased the abudance of
496 cycling basal progenitors, a hallmark of the developing human neocortex. Moreover, we
497 previously showed that the human-specific function of *ARHGAP11B* in cNPCs arose by

498 a single nucleotide substitution that generated a new splice donor site, the use of which
499 generates a novel human-specific C-terminal protein sequence that we implicate in
500 basal progenitor amplification (Florio et al., 2015; Florio et al., 2016). Importantly, this
501 single nucleotide substitution presumably occurred relatively recently during human
502 evolution (Florio et al., 2016), that is, after the partial gene duplication event ~5 million
503 years ago (Antonacci et al., 2014; Dennis et al., 2017; Riley et al., 2002). Furthermore,
504 we have identified here an *ARHGAP11B* splice variant that is specifically expressed in
505 human bRG (Fig. S4), the basal progenitor type thought to have a key role in neocortex
506 expansion (Betizeau et al., 2013; Borrell and Götz, 2014; Borrell and Reillo, 2012;
507 Florio and Huttner, 2014; Lui et al., 2011). Interestingly, in contrast to the other
508 protein-coding *ARHGAP11B* splice variant detected, which contains a long 3'-UTR with
509 predicted microRNA binding sites and which is predominantly expressed in aRG, the
510 bRG-specific *ARHGAP11B* splice variant contains only a short 3'-UTR lacking
511 predictable microRNA binding sites. This suggests that *ARHGAP11B* mRNAs may be
512 subject to differential, microRNA-mediated, regulation depending on whether
513 ARHGAP11B functions in the lineage progression from aRG to bRG or in bRG
514 amplification. Taken together, our findings reveal genomic changes at a variety of

515 levels that gave rise to novel functions and patterns of expression in cNPCs and that are

516 likely relevant for the development and evolution of the human neocortex.

517

518 ## Materials and Methods

519

520 **Human fetal brain tissue**

521 Human fetal brain tissue was obtained from the Klinik und Poliklinik für
522 Frauenheilkunde und Geburtshilfe, Universitätsklinikum Carl Gustav Carus of the
523 Technische Universität Dresden, following elective termination of pregnancy and
524 informed written maternal consent, and with approval of the local University Hospital
525 Ethical Review Committees. The gestational age of the specimen used for ISH (13 weeks
526 post conception, wpc) was assessed by ultrasound measurements of crown-rump length,
527 as described previously (Florio et al., 2015). Immediately after termination of pregnancy,
528 the tissue was placed on ice and transported to the lab. The sample was then transferred to
529 ice-cold Tyrode's solution, and tissue fragments of cerebral cortex were identified and
530 dissected. Tissue was fixed in 4% paraformaldehyde in 120 mM phosphate buffer (pH
531 7.4) for 3 hours at room temperature followed by 24 hours at 4°C. Fixed tissue was then
532 incubated in 30% sucrose overnight, embedded in Tissue-Tek OCT (Sakura), and frozen
533 on dry ice. Cryosections of 12 µm were produced using a cryostat (Microm HM 560,
534 Thermo Fisher Scientific) and stored at –20°C until processed for ISH.

535

536 **Identification of human cNPC-enriched protein-coding genes**

537 To identify genes the expression of which is enriched in human cNPCs, we screened
538 differential gene expression data from four published datasets (Fietz et al., 2012; Florio
539 et al., 2015; Miller et al., 2014; Pollen et al., 2015) generated from 12-19 wpc human
540 fetal neocortex, using diverse cortical zone or cell type-enrichment strategies and modes
541 of determination of RNA levels (summarized in Table S1).

542

543 Fietz et al., 2012 – This dataset was generated by RNA-Seq of the germinal zones (VZ,
544 iSVZ, oSVZ) and CP isolated by LCM from the neocortex of six human fetuses ranging
545 in gestational age from 12 to 16 wpc. We screened this dataset for protein-coding genes
546 more highly expressed, across all stages, in either VZ, iSVZ or oSVZ than CP (as
547 determined by DGE analysis, $p<0.01$, (Fietz et al., 2012). The resulting data-subset
548 contained 2,758 genes (Table S1, Fig. 1).

549

550 Miller et al., 2014 (BrainSpan Atlas of the Allen Brain Institute, Prenatal LMD
551 Microarray, http://www.brainspan.org/lcm/search/index.html) – This dataset (Miller et

552 al., 2014) was generated by microarray RNA expression profiling of germinal zones
553 (VZ, iSVZ, oSVZ) and neuron-enriched layers (IZ, subplate, CP, marginal zone, subpial
554 granular zone) isolated by LCM from fetal human neocortex (for the purpose of the
555 present analysis, only data obtained from two 15-16 wpc human fetuses were
556 considered). We screened this dataset for protein-coding genes with highest correlation
557 with either VZ, iSVZ or oSVZ (correlation coefficient >0.5) compared to all cortical
558 regions analyzed. The resulting data-subset contained 4,555 genes (Table S1, Fig. 1).
559
560 Florio et al., 2015 – This dataset was generated by RNA-Seq of human radial glia
561 subtypes (aRG and bRG) and CP neurons (N) isolated from the neocortex of two 13
562 wpc human fetuses. These cell types were differentially labeled using a combination of
563 fluorescent molecular markers, and isolated by FACS. By experimental design, only
564 cells that exhibited apical plasma membrane and/or contacted the basal lamina were
565 isolated. Moreover, the isolation of aRG and bRG was confined to cells that had
566 duplicated their DNA, and the neuron fraction contained a minority of bRG in G1
567 (Florio et al., 2015). We screened this dataset for protein-coding genes with higher

568 expression in either aRG or bRG than N (as determined by DGE analysis, $p<0.01$,
569 (Florio et al., 2015). The resulting data-subset contained 2,106 genes (Table S1, Fig. 1).
570
571 Pollen et al., 2015 – This dataset was generated by RNA-seq of single cells captured
572 from the VZ and SVZ microdissected from the neocortex of three 16.5-19 wpc human
573 fetuses. Cells were post-hoc attributed – based on gene expression profiling – to either
574 radial glia (aRG and bRG), intermediate progenitors (i.e. bIPs), or neurons (N). We
575 screened this dataset for genes positively correlated with either radial glia or bIPs
576 (correlation coefficient >0.1, (Pollen et al., 2015) and negatively correlated with N
577 (correlation coefficient <0.1, (Pollen et al., 2015). The resulting data-subset contained
578 5,335 genes (Table S1, Fig 1).
579
580 These data-subsets contain only protein-coding genes, which were identified and
581 selected using the Ensembl data-mining tool BioMart
582 (http://www.ensembl.org/biomart/martview/), implementing the Genome Reference
583 Consortium Human Build 38 (GRCh38.p10) dataset.

584 Next, we intersected the four data-subsets obtained. To do this, we converted all gene
585 IDs contained in the four original datasets to match the latest Ensembl gene annotation
586 (Ensembl v89) of the GRCh38.p10 genome assembly, and then searched for the
587 co-occurrence of genes (or lack thereof) across the four data-subsets. This resulted in
588 3,722 human cNPC-enriched protein-coding genes present in at least two of the four
589 data-subsets (listed in Table S1, see also Fig. 1).
590
591 **Screening of human cNPC-enriched protein-coding genes for primate-specific**
592 **orthologs**
593 The 3,722 human cNPC-enriched protein-coding genes were screened for the
594 occurrence of one-to-one orthologs in non-primate species, using BioMart and
595 implementing v89 Ensembl annotation of "1-to-1 orthologs". All genes that had an
596 annotated 1-to-1 ortholog in non-primate species were excluded from our four
597 data-subsets. This yielded 83 genes that were candidates to be primate-specific.
598
599 We visualized whole genome alignments in the UCSC genome browser (Tyner et al.,
600 2017) to manually analyze each of the 83 candidate primate-specific genes. To this end,

601 we inspected co-linear chains of local alignments (Kent et al., 2003) between the human
602 hg38 genome assembly and the assemblies of non-primate mammals to check if the
603 human gene locus aligned to non-primate mammals. For the genes that aligned to
604 non-primate mammals, regardless of whether they aligned in a conserved or in a
605 different context, we used gene annotations of the aligning species to assess which gene
606 is annotated in the respective locus. For this purpose, we made use of gene annotations
607 from Refseq, Ensembl (Aken et al., 2017) and CESAR (Sharma et al., 2016) (a method
608 that transfers human gene annotations to other aligned genomes if the gene has an intact
609 reading frame), and removed those candidate genes that likely have an aligning gene in
610 non-primate mammals. This reduced the list of the 83 candidates to 54 genes that we
611 considered as primate-specific.
612
613 **Tracing the evolution of the primate-specific genes in the primate lineage**
614 We traced the evolution of these 54 primate-specific genes in the primate lineage to
615 determine which of these have orthologs, in non-human primates, to the corresponding
616 54 human cNPC-enriched protein-coding genes, and which do not, and therefore are
617 human-specific. To this end, we inspected co-linear alignment chains and a multiple

618 genome alignment that includes 17 non-human primate genomes (Sharma and Hiller,
619 2017). For the genes that aligned to other primates, we used the CESAR annotations to
620 check if a gene of interest has an intact reading frame in other species. We only
621 considered a gene to be conserved if an intact reading frame is present in the respective
622 species. For example, while *FAM182* aligns in a conserved context to chimpanzee and
623 gorilla, CESAR did not find an intact reading frame and did not annotate the gene;
624 indeed, inspecting the multiple genome alignment revealed a frameshift in chimpanzee
625 and a stop codon mutation in gorilla, showing that *FAM182B* is likely a non-coding
626 gene in non-human primates. Then, we assigned each gene to a node in the primate
627 phylogeny (clade), based on the descending species that likely have an intact coding
628 gene. Note that this inferred ancestry does not imply that all descending species have an
629 intact gene. This is exemplified by *TMEM99*, which aligns to all great apes and has an
630 intact reading frame in human and orangutan, but encodes no or a truncated protein in
631 chimpanzee/bonobo (due to a frameshift mutation) and gorilla (due to a stop codon
632 mutation).
633

We combined this analysis with BLAT searches using the human protein or human mRNA sequence to assess the number of aligning loci in other primates; however, this was not conclusive for highly complex loci such as the duplications involving *ANKRD20A* and *CBWD5* genes, where numerous similar genes and pseudogenes are present and the completeness of non-human primate genome assemblies is not certain due to the presence of assembly gaps. In addition, for human-specific candidates that arose by duplication, inspecting the respective genomic locus in the chimpanzee genome browser was useful, since human duplications are visible as additional, overlapping alignment chains.

Paralog-specific and isoform-specific gene expression

To estimate expression differences among cNPC types between (a) given human-specific gene(s) and its/their highly similar ancestral paralog(s) in the human genome, we used the Kallisto probabilistic algorithm, which has been proven to be accurate in assigning reads to specific transcripts, including those originating from highly similar paralog genes in the human genome (Bray et al., 2016).

650 For this analysis, we used reads generated previously by RNA-Seq of human aRG, bRG
651 and N (SRA Access, SRP052294, (Florio et al., 2015)) as input, GRCh38 as genome
652 reference, and Ensembl v89 as genome annotation reference. Transcript abundances
653 were output in Transcripts per Million (TPM) units. To compare expression between
654 human-specific and ancestral paralog genes (Fig. 7), we extracted TPM values for all
655 paralogs in each orthologous group, and summed the TPM values for all protein-coding
656 transcripts (as per Ensembl annotation) for each gene. To compare expression between
657 different splice variants produced by each human-specific gene (Fig S4), we extracted
658 the TPM values specific for each individual splice variant and expressed the data
659 relative to each other.
660
661 Kallisto's transcript abundance measurements represent a probabilistic approximation
662 of actual transcript levels, and thus are an estimate. In order to compare actual paralog
663 gene expression in distinct cNPC types and neurons, we performed a second type of
664 analysis, which did not aim at providing an estimate of absolute transcript abundances,
665 but rather at providing a precise determination of the relative gene expression
666 differences between paralogs. To this end, we aligned mRNA sequences of ancestral

667 and human-specific paralogs in each orthology group, using Clustalw2
668 (http://www.ebi.ac.uk/Tools/msa/clustalw2/), and manually identified the homologous
669 (but not identical) core sequence of each alignment (Suppl. Files S1-S8; see Fig. S3A
670 for illustration of a hypothetical example). The corresponding sequences of each
671 paralog – of same length by design – were used as reference for previously generated
672 RNA-Seq reads from aRG, bRG and N (SRA Access, SRP052294, (Florio et al., 2015))
673 in order to search for paralog-specific mRNA reads. Reads aligning to both, ancestral
674 and human-specific paralogs, were discarded as ambiguous, and only those reads
675 aligning to paralog-specific sites (SNPs or indels), referred to as paralog-specific reads,
676 were used for quantification (Fig S3B). This stringent alignment was carried out using
677 bowtie1 (bowtie -Sp 5 -m 1 -v0).
678
679 It should be noted that, in contrast to the Kallisto-based analysis, the latter type of
680 analysis does not distinguish between reads that originate from protein-coding and
681 non-protein-coding transcripts of a given gene. Therefore, the quantifications shown in
682 Fig. S3B reflect counts of all reads mapping to a given gene, whereas the

683 quantifications shown in Fig. 4 reflect summed counts of protein-coding gene
684 transcripts only.
685
686 **Genomic qPCR**
687 Genomic DNA was obtained from EBV-transformed B cells of human, bonobo and
688 chimpanzee, as described previously (Prüfer et al., 2012). Primers (Table S2) were
689 designed for two different amplicons per orthologous gene group to bind to the same
690 region of the human-specific gene(s) under study, its human paralog(s), and the
691 chimpanzee and bonobo orthologs. Only one mismatch in the primer binding sequence
692 between the reference genomes of the three species was allowed.
693
694 qPCR was performed on human, chimpanzee and bonobo genomic DNA, using either
695 the ABsolute qPCR SYBR Greenmix (Thermo Fisher Scientific) on a Mx3000P qPCR
696 System (Stratagene) or the Fast Start Essential DNA Green Master (Roche) on a
697 Lightcycler 96 (Roche). The relative copy number between the three species was
698 determined by the comparative cycle threshold (Ct) approach (Livak and Schmittgen,
699 2001) as follows. The Ct values for the human, chimpanzee and bonobo genes under

700 study were normalized to the Ct value of the highly conserved single-copy gene *STX12*.
701 The normalized values were then compared between the three species, using bonobo as
702 reference, to determine the relative copy number.
703
704 **In utero electroporation and tissue processing**
705 In utero electroporation was performed on C57BL/6J mice in agreement with German
706 Animal Welfare Legislation, as described previously (Florio et al., 2015). Pregnant
707 dams carrying E13.5 embryos were deeply anesthetized using isoflurane. Embryos were
708 injected into the lateral ventricle with either 1 µg/µl of pCAGGS-NOTCH2NL and 0.5
709 µg/µl of pCAGGS-GFP or 1 µg/µl of empty pCAGGS and 0.5 µg/µl of pCAGGS-GFP
710 in PBS containing 0.1% Fast Green, followed by electroporation (30 V, six 50-msec
711 pulses with 1 sec intervals). Electroporated cerebral cortices were dissected at E15.5
712 and fixed overnight at 4°C in 4% paraformaldehyde in 120 mM phosphate buffer (pH
713 7.4). Fixed cortices were incubated in 20% sucrose for 24 hours at 4°C. Cortices were
714 embedded in Tissue-TEK (O.C.T, Sakura Finetek) and stored at –20°C.
715
716

Immunofluorescence

Cryosections of 20 µm were prepared. Cryosections were first rehydrated in PBS. Antigen retrieval was performed for 1 hour at 70°C in 0.01 M citrate. Cryosections were permeabilized by treatment with 0.1% Triton X-100 in PBS for 30 min. Cryosections were quenched for 30 min in 0.1 M glycine in PBS, blocked in 0.2% gelatin, 300 mM NaCl, and 0.3% Triton X-100 in PBS, and incubated overnight at 4°C with primary antibodies (Ki67, rabbit, Abcam, Ab15580, 1/500; PH3, rat, Abcam, Ab10543, 1/1000; GFP, chicken, Abcam, Ab13970, 1/1000). Appropriate secondary antibodies were incubated for 2 hours at room temperature (Alexa Fluor 488, 594, Molecular Probes, 1:500; DAPI, Sigma, 1/1000). Cryosections were mounted in Mowiol (Merck Biosciences).

In-situ hybridization

Templates were amplified by PCR (see Table S3 for primer sequences) from oligo-dT-primed cDNA prepared from fetal human neocortex total RNA, and RNA probes directed against the mRNA(s) of a given human-specific gene and (if applicable) its paralog(s) in the human genome were synthesized using the DIG RNA labeling Mix

(Roche). The *ARHGAP11B* LNA probe was designed with the Custom LNA mRNA Detection Probe design tool (Exiqon), focusing only on the sequence spanning the *ARHGAP11B* exon5–exon6 boundary, where *ARHGAP11B* is sufficiently different from *ARHGAP11A* (see Fig. S2) (Florio et al., 2016), and searching for hybridization with a predicted RNA melting temperature of 85°C. The LNA probe (5'-AGTCTGGTACACGCCCTTCTTTTCT-3') was synthesized and labelled with digoxigenin at the 5' and 3' ends (Exiqon).

In-situ hybridization was performed on 12–µm cryosections of 13 wpc fetal human neocortex and on COS-7 cells. Prior to the hybridization step, cryosections/cells were sequentially treated with 0.2 M HCl (2x 5 min, room temperature) and then with 6 µg/ml proteinase K in PBS, pH 7.4 (20 min, room temperature). Hybridization was performed overnight at 65°C with either 20 ng/µl of a given RNA probe or 40 nM *ARHGAP11B* LNA probe. TSA Plus DIG detection Kit (Perkin Elmer) was used for signal amplification, and the signal was detected immunohistochemically with mouse anti-digoxigenin HRP antibody (Perkin Elmer) and NBT/BCIP (Roche) as color substrate.

751

752 **Image acquisition**

753 ISH images were acquired on a Zeiss Axio Scan slide scanner, and processed using

754 ImageJ. Fluorescent images of electroporated neocortex were acquired using a Zeiss

755 laser scanning confocal microscope 700 using a 20x objective. Quantifications were

756 performed using Fiji.

Acknowledgements

We are grateful to the Computer Service Facilities of the MPI-CBG and MPI-PKS and to other services and facilities of the MPI-CBG for the outstanding support provided, notably J. Helppi and his team from the Animal Facility and Jan Peychl and his team from the Light Microscopy Facility. We thank Robert Lachmann for providing fetal human tissue, Hella Hartmann of the CRTD of the Technische Universität Dresden for help with the acquisition of the ISH images, and members of the Huttner laboratory for critical discussion. We are grateful to Dr. Tomislav Maricic (MPI for Evolutionary Anthropology) for donating human, chimpanzee and bonobo genomic DNA. M.F. would like to thank Dr. Fenna Krienen (Harvard Medical School) for helpful discussion and critical reading of the manuscript. M.F. was a member of the International Max Planck Research School for Cell, Developmental and Systems Biology and a doctoral student at Technische Universität Dresden. W.B.H. was supported by grants from the Deutsche Forschungsgemeinschaft (DFG) (SFB 655, A2), the European Research Council (250197), and ERA-NET NEURON (MicroKin).

Competing interests

The authors declare that no competing interests exist.

777 **References**

778

779 Aken, B.L., Achuthan, P., Akanni, W., Amode, M.R., Bernsdorff, F., Bhai, J.,
780 Billis, K., Carvalho-Silva, D., Cummins, C., Clapham, P.*, et al.* (2017).
781 Ensembl 2017. Nucleic Acids Res *45*, D635-D642.
782 Antonacci, F., Dennis, M.Y., Huddleston, J., Sudmant, P.H., Steinberg, K.M.,
783 Rosenfeld, J.A., Miroballo, M., Graves, T.A., Vives, L., Malig, M.*, et al.*
784 (2014). Palindromic GOLGA8 core duplicons promote chromosome 15q13.3
785 microdeletion and evolutionary instability. Nat Genet *46*, 1293-1302.
786 Azevedo, F.A.C., Carvalho, L.R.B., Grinberg, L.T., Farfel, J.M., Ferretti,
787 R.E.L., Leite, R.E.P., Jacob, W., Lent, R., and Herculano-Houzel, S. (2009).
788 Equal Numbers of Neuronal and Nonneuronal Cells Make the Human Brain
789 an Isometrically Scaled-Up Primate Brain. J Comp Neurol *513*, 532-541.
790 Bae, B.I., Jayaraman, D., and Walsh, C.A. (2015). Genetic changes shaping
791 the human brain. Dev Cell *32*, 423-434.
792 Bahram, S., Bresnahan, M., Geraghty, D.E., and Spies, T. (1994). A second
793 lineage of mammalian major histocompatibility complex class I genes. Proc
794 Natl Acad Sci U S A *91*, 6259-6263.
795 Bailey, J.A., Gu, Z., Clark, R.A., Reinert, K., Samonte, R.V., Schwartz, S.,
796 Adams, M.D., Myers, E.W., Li, P.W., and Eichler, E.E. (2002). Recent
797 segmental duplications in the human genome. Science *297*, 1003-1007.
798 Betizeau, M., Cortay, V., Patti, D., Pfister, S., Gautier, E., Bellemin-Ménard,
799 A., Afanassieff, M., Huissoud, C., Douglas, R.J., Kennedy, H.*, et al.* (2013).
800 Precursor diversity and complexity of lineage relationships in the outer
801 subventricular zone of the primate. Neuron *80*, 442-457.
802 Borrell, V., and Götz, M. (2014). Role of radial glial cells in cerebral cortex
803 folding. Curr Opin Neurobiol *27*, 39-46.
804 Borrell, V., and Reillo, I. (2012). Emerging roles of neural stem cells in
805 cerebral cortex development and evolution. Dev Neurobiol *72*, 955-971.
806 Bray, N.L., Pimentel, H., Melsted, P., and Pachter, L. (2016). Near-optimal
807 probabilistic RNA-seq quantification. Nat Biotech *34*, 525-527.
808 Brosius, J. (1991). Retroposons - Seeds of Evolution. Science *251*, 753-753.

809 Brunet, M., Guy, F., Pilbeam, D., Lieberman, D.E., Likius, A., Mackaye,
810 H.T., Ponce de Leon, M.S., Zollikofer, C.P.E., and Vignaud, P. (2005). New
811 material of the earliest hominid from the Upper Miocene of Chad. Nature
812 *434*, 752-755.
813 Brunet, M., Guy, F., Pilbeam, D., Mackaye, H.T., Likius, A., Ahounta, D.,
814 Beauvilain, A., Blondel, C., Bocherens, H., Boisserie, J.-R.*, et al.* (2002). A
815 new hominid from the Upper Miocene of Chad, Central Africa. Nature *418*,
816 145-151.
817 Buckner, R.L., and Krienen, F.M. (2013). The evolution of distributed
818 association networks in the human brain. Trends Cogn Sci *17*, 648-665.
819 Charrier, C., Joshi, K., Coutinho-Budd, J., Kim, J.E., Lambert, N., de
820 Marchena, J., Jin, W.L., Vanderhaeghen, P., Ghosh, A., Sassa, T.*, et al.*
821 (2012). Inhibition of SRGAP2 function by its human-specific paralogs
822 induces neoteny during spine maturation. Cell *149*, 923-935.
823 Dehay, C., Kennedy, H., and Kosik, K.S. (2015). The outer subventricular
824 zone and primate-specific cortical complexification. Neuron *85*, 683-694.
825 Dennis, M.Y., Harshman, L., Nelson, B.J., Penn, O., Cantsilieris, S.,
826 Huddleston, J., Antonacci, F., Penewit, K., Denman, L., Raja, A.*, et al.*
827 (2017). The evolution and population diversity of human-specific segmental
828 duplications. *1*, 0069.
829 Dominguez, A., Ramos-Morales, F., Romero, F., Rios, R.M., Dreyfus, F.,
830 Tortolero, M., and Pintor-Toro, J.A. (1998). hpttg, a human homologue of rat
831 pttg, is overexpressed in hematopoietic neoplasms. Evidence for a
832 transcriptional activation function of hPTTG. Oncogene *17*, 2187-2193.
833 Dougherty, M.L., Nuttle, X., Penn, O., Nelson, B.J., Huddleston, J., Baker,
834 C., Harshman, L., Duyzend, M.H., Ventura, M., Antonacci, F.*, et al.* (2017).
835 The birth of a human-specific neural gene by incomplete duplication and
836 gene fusion. Genome Biol *18*, 49.
837 Eichler, E.E., Clark Ra Fau - She, X., and She, X. (2004). An assessment of
838 the sequence gaps: unfinished business in a finished human genome.
839 Fietz, S.A., Lachmann, R., Brandl, H., Kircher, M., Samusik, N., Schroder,
840 R., Lakshmanaperumal, N., Henry, I., Vogt, J., Riehn, A.*, et al.* (2012).
841 Transcriptomes of germinal zones of human and mouse fetal neocortex

suggest a role of extracellular matrix in progenitor self-renewal. Proc Natl Acad Sci U S A *109*, 11836-11841.

Florio, M., Albert, M., Taverna, E., Namba, T., Brandl, H., Lewitus, E., Haffner, C., Sykes, A., Wong, F.K., Peters, J., *et al.* (2015). Human-specific gene ARHGAP11B promotes basal progenitor amplification and neocortex expansion. Science *347*, 1465-1470.

Florio, M., Borrell, V., and Huttner, W.B. (2017). Human-specific genomic signatures of neocortical expansion. Curr Opin Neurobiol *42*, 33-44.

Florio, M., and Huttner, W.B. (2014). Neural progenitors, neurogenesis and the evolution of the neocortex. Development *141*, 2182-2194.

Florio, M., Namba, T., Pääbo, S., Hiller, M., and Huttner, W.B. (2016). A single splice site mutation in human-specific ARHGAP11B causes basal progenitor amplification. Science advances *2*.

Fortna, A., Kim, Y., MacLaren, E., Marshall, K., Hahn, G., Meltesen, L., Brenton, M., Hink, R., Burgers, S., Hernandez-Boussard, T., *et al.* (2004). Lineage-specific gene duplication and loss in human and great ape evolution. PLoS Biol *2*, E207.

Götz, M., and Huttner, W.B. (2005). The cell biology of neurogenesis. Nat Rev Mol Cell Biol *6*, 777-788.

Hirokawa, N., Noda, Y., Tanaka, Y., and Niwa, S. (2009). Kinesin superfamily motor proteins and intracellular transport. Nat Rev Mol Cell Biol *10*, 682-696.

Hurles, M. (2004). Gene duplication: the genomic trade in spare parts. PLoS Biol *2*, E206.

Imayoshi, I., Shimojo, H., Sakamoto, M., Ohtsuka, T., and Kageyama, R. (2013). Genetic visualization of notch signaling in mammalian neurogenesis. Cellular and Molecular Life Sciences *70*, 2045-2057.

Kaas, J.H. (2013). The evolution of brains from early mammals to humans. Wiley interdisciplinary reviews Cognitive science *4*, 33-45.

Kawaguchi, A., Ikawa, T., Kasukawa, T., Ueda, H.R., Kurimoto, K., Saitou, M., and Matsuzaki, F. (2008). Single-cell gene profiling defines differential progenitor subclasses in mammalian neurogenesis. Development *135*, 3113-3124.

875 Kent, W.J., Baertsch, R., Hinrichs, A., Miller, W., and Haussler, D. (2003).
876 Evolution's cauldron: Duplication, deletion, and rearrangement in the mouse
877 and human genomes. Proceedings of the National Academy of Sciences *100*,
878 11484-11489.
879 Kriegstein, A.R., and Götz, M. (2003). Radial glia diversity: a matter of cell
880 fate. Glia *43*, 37-43.
881 Kumar, S., Stecher, G., Suleski, M., and Hedges, S.B. (2017). TimeTree: A
882 Resource for Timelines, Timetrees, and Divergence Times. Mol Biol Evol *34*,
883 1812-1819.
884 Lewitus, E., Kelava, I., Kalinka, A.T., Tomancak, P., and Huttner, W.B.
885 (2014). An adaptive threshold in mammalian neocortical evolution. PLoS
886 Biol *12*, e1002000.
887 Livak, K.J., and Schmittgen, T.D. (2001). Analysis of relative gene
888 expression data using real-time quantitative PCR and the 2(-Delta Delta
889 C(T)) Method. Methods *25*, 402-408.
890 Long, M., Betran, E., Thornton, K., and Wang, W. (2003). The origin of new
891 genes: glimpses from the young and old. Nat Rev Genet *4*, 865-875.
892 Lui, J.H., Hansen, D.V., and Kriegstein, A.R. (2011). Development and
893 evolution of the human neocortex. Cell *146*, 18-36.
894 Lynch, M., and Conery, J.S. (2000). The Evolutionary Fate and
895 Consequences of Duplicate Genes. Science *290*, 1151-1155.
896 Marques, A.C., Dupanloup, I., Vinckenbosch, N., Reymond, A., and
897 Kaessmann, H. (2005). Emergence of young human genes after a burst of
898 retroposition in primates. PLoS Biol *3*, e357.
899 Meyer, M., Kircher, M., Gansauge, M.T., Li, H., Racimo, F., Mallick, S.,
900 Schraiber, J.G., Jay, F., Prufer, K., de Filippo, C.*, et al.* (2012). A
901 high-coverage genome sequence from an archaic Denisovan individual.
902 Science *338*, 222-226.
903 Miller, J.A., Ding, S.L., Sunkin, S.M., Smith, K.A., Ng, L., Szafer, A., Ebbert,
904 A., Riley, Z.L., Royall, J.J., Aiona, K.*, et al.* (2014). Transcriptional landscape
905 of the prenatal human brain. Nature *508*, 199-206.
906 Mitchell, C., and Silver, D.L. (2017). Enhancing our brains: Genomic
907 mechanisms underlying cortical evolution. Semin Cell Dev Biol.

908 Namba, T., and Huttner, W.B. (2017). Neural progenitor cells and their role
909 in the development and evolutionary expansion of the neocortex. Wiley
910 Interdiscip Rev Dev Biol *6*, e256.
911 Otani, T., Marchetto, M.C., Gage, F.H., Simons, B.D., and Livesey, F.J.
912 (2016). 2D and 3D Stem Cell Models of Primate Cortical Development
913 Identify Species-Specific Differences in Progenitor Behavior Contributing to
914 Brain Size. Cell Stem Cell *18*, 467-480.
915 Pierfelice, T.J., Schreck, K.C., Eberhart, C.G., and Gaiano, N. (2008). Notch,
916 Neural Stem Cells, and Brain Tumors. Cold Spring Harb Symp Quant Biol
917 *73*, 367-375.
918 Pollen, A.A., Nowakowski, T.J., Chen, J., Retallack, H., Sandoval-Espinosa,
919 C., Nicholas, C.R., Shuga, J., Liu, S.J., Oldham, M.C., Diaz, A.*, et al.* (2015).
920 Molecular Identity of Human Outer Radial Glia during Cortical
921 Development. Cell *163*, 55-67.
922 Prüfer, K., Munch, K., Hellmann, I., Akagi, K., Miller, J.R., Walenz, B.,
923 Koren, S., Sutton, G., Kodira, C., Winer, R.*, et al.* (2012). The bonobo genome
924 compared with the chimpanzee and human genomes. Nature *486*, 527-531.
925 Prüfer, K., Racimo, F., Patterson, N., Jay, F., Sankararaman, S., Sawyer, S.,
926 Heinze, A., Renaud, G., Sudmant, P.H., de Filippo, C.*, et al.* (2014). The
927 complete genome sequence of a Neanderthal from the Altai Mountains.
928 Nature *505*, 43-49.
929 Rakic, P. (2009). Evolution of the neocortex: a perspective from
930 developmental biology. Nat Rev Neurosci *10*, 724-735.
931 Riley, B., Williamson, M., Collier, D., Wilkie, H., and Makoff, A. (2002). A
932 3-Mb map of a large segmental duplication overlapping the alpha7-nicotinic
933 acetylcholine receptor gene (CHRNA7) at human 15q13-q14. Genomics *79*,
934 197-209.
935 Roach, W.G., Chavez, J.A., Miinea, C.P., and Lienhard, G.E. (2007).
936 Substrate specificity and effect on GLUT4 translocation of the Rab
937 GTPase-activating protein Tbc1d1. Biochem J *403*, 353-358.
938 Sharma, V., Elghafari, A., and Hiller, M. (2016). Coding exon-structure
939 aware realigner (CESAR) utilizes genome alignments for accurate
940 comparative gene annotation. Nucleic Acids Res *44*, e103-e103.

941 Sharma, V., and Hiller, M. (2017). Increased alignment sensitivity improves
942 the usage of genome alignments for comparative gene annotation. Nucleic
943 Acids Res *45*, 8369-8377.
944 Sousa, A.M.M., Meyer, K.A., Santpere, G., Gulden, F.O., and Sestan, N.
945 (2017). Evolution of the Human Nervous System Function, Structure, and
946 Development. Cell *170*, 226-247.
947 Striedter, G.F. (2005). Principles of Brain Evolution (Sinauer Associates
948 Inc.).
949 Sudmant, P.H., Kitzman, J.O., Antonacci, F., Alkan, C., Malig, M., Tsalenko,
950 A., Sampas, N., Bruhn, L., Shendure, J., and Eichler, E.E. (2010). Diversity
951 of human copy number variation and multicopy genes. Science *330*, 641-646.
952 Tyner, C., Barber, G.P., Casper, J., Clawson, H., Diekhans, M., Eisenhart,
953 C., Fischer, C.M., Gibson, D., Gonzalez, J.N., Guruvadoo, L.*, et al.* (2017).
954 The UCSC Genome Browser database: 2017 update. Nucleic Acids Res *45*,
955 D626-D634.
956 Vignaud, P., Duringer P Fau - Mackaye, H.T., Mackaye Ht Fau - Likius, A.,
957 Likius A Fau - Blondel, C., Blondel C Fau - Boisserie, J.-R., Boisserie Jr Fau
958 - De Bonis, L., De Bonis L Fau - Eisenmann, V., Eisenmann V Fau - Etienne,
959 M.-E., Etienne Me Fau - Geraads, D., Geraads D Fau - Guy, F.*, et al.* (2002).
960 Geology and palaeontology of the Upper Miocene Toros-Menalla hominid
961 locality, Chad.
962 Vlotides, G., Eigler, T., and Melmed, S. (2007). Pituitary tumor-transforming
963 gene: physiology and implications for tumorigenesis. Endocrine reviews *28*,
964 165-186.
965 Wilkinson, G., Dennis, D., and Schuurmans, C. (2013). Proneural genes in
966 neocortical development. Neuroscience *253*, 256-273.
967 Zhang, X., Horwitz, G.A., Prezant, T.R., Valentini, A., Nakashima, M.,
968 Bronstein, M.D., and Melmed, S. (1999). Structure, expression, and function
969 of human pituitary tumor-transforming gene (PTTG). Molecular
970 endocrinology (Baltimore, Md) *13*, 156-166.
971 Zhu, C., Zhao, J., Bibikova, M., Leverson, J.D., Bossy-Wetzel, E., Fan, J.-B.,
972 Abraham, R.T., and Jiang, W. (2005). Functional Analysis of Human
973 Microtubule-based Motor Proteins, the Kinesins and Dyneins, in

974 Mitosis/Cytokinesis Using RNA Interference. Molecular Biology of the Cell
975 *16*, 3187-3199.

976

977

978 # Figure Legends

979

980 **Fig. 1. A screen for human cNPC-enriched protein-coding genes and**

981 **determination which of them have orthologs only in primates.**

982 (**A**) Cartoon illustrating the main zones and neural cell types in the fetal human cortical

983 wall that were screened for differential gene expression in the human transcriptome

984 datasets as depicted in (B). Adapted from (Florio et al., 2017). SP, subplate; MZ,

985 marginal zone.

986 (**B**) The indicated four published transcriptome datasets from fetal human neocortical

987 tissue (Fietz et al., 2012; Miller et al., 2014) and cell populations (Florio et al., 2015;

988 Pollen et al., 2015) were screened for protein-coding genes showing higher levels of

989 mRNA expression in the indicated germinal zones and cNPC types than in than in the

990 non-proliferative zones and neurons.

991 (**C**) Venn diagram showing the data-subsets of human protein-coding genes displaying

992 the differential gene expression pattern depicted in (B). Numbers within the diagram

993 indicate genes found in two (pink), three (yellow) or all four (red) data-subsets. Genes

994 found in at least two data-subsets were considered as being cNPC-enriched.

995 (**D**) Selected genes with established biological roles found in two (pink), three (yellow),
996 or all four (red) data-subsets.
997 (**E**) Stepwise analysis leading from the 3,722 human cNPC-enriched protein-coding
998 genes to the identification of 54 primate-specific genes.
999
1000 **Fig. 2. Occurrence of the primate-specific genes in the various primate clades.**
1001 Assignment of the 54 primate-specific genes to a primate clade, based on the primate
1002 genome(s) in which an intact reading frame was found in the present analysis. Clades
1003 are specified on the top left. The color-coding and brackets indicate the species in each
1004 clade analyzed in the present study. Note that the occurrence of the genes in the various
1005 clades does not necessarily apply to every species in the clade.
1006
1007 **Fig. 3. Evolutionary origin of the *PTTG2* gene.**
1008 (**A**) Origin of the *PTTG2* gene by reverse transcription of the *PTTG1* mRNA and
1009 insertion as a retroposon into the *TBC1D1* locus in the ancestor to New-World monkeys,
1010 Old-World monkeys and apes (Simiiformes).

(**B**) Comparison of the PTTG1 and hominoidea PTTG2 polypeptides, and of the prematurely closed open reading frames of non-ape simiiformes *PTTG2*.

Fig. 4. Evolution of the human-specific cNPC-enriched protein-coding genes.

Diagrams depicting the evolutionary origin of the 13 human-specific genes.

(**A**) Duplication of the entire ancestral gene, which applies to nine of the human-specific genes. *NOTCH2NL* is included in this group because it arose by entire duplication of a short *NOTCH2* splice variant; the latter, however, comprises only parts of the *NOTCH2* gene (hence the asterisk). Note that the gene duplication giving rise to *SMN2* occurred after the Neandertal – modern human lineage split, whereas the other eight gene duplications occurred before that split (Dennis et al., 2017).

(**B**) Partial gene duplication (~5 Mya) giving rise to *ARHGAP11B* (Antonacci et al., 2014; Dennis et al., 2017; Riley et al., 2002). Note that a single C–>G substitution in exon 5 (red box), which likely occurred after the gene duplication event but before the Neandertal – modern human lineage split, created a new splice donor site, causing a reading frame shift that resulted in a novel, human-specific 47 amino acid C-terminal sequence (Florio et al., 2015; Florio et al., 2016).

(**C**) Exon duplication and replacement giving rise to human *ZNF492*. Exon 4 of *ZNF98* (blue) is duplicated and inserted into intron 3 of *ZNF492* (orange), rendering the original *ZNF492* exon 4 a pseudoexon.

(**D**) Removal of a stop codon converting the non-coding *FAM182B* of non-human primates into the protein-coding human *FAM182B*. A single T–>G substitution removes the stop codon at the 5' end of exon 3, thereby generating an open reading frame (purple).

(**E**) Validation of the human-specific nature of selected human genes by determination of their copy numbers. Human (blue), chimpanzee (orange) and bonobo (yellow) genomic DNA was used as template to perform a qPCR that would generate two distinct amplicons of both, the gene common to all three species (black regular letters) and the human-specific gene(s) under study (red bold letters), as indicated. The relative amounts of amplicons obtained for each of the four gene groups are depicted with the amounts of amplicons obtained with the bonobo genomic DNA as template being set to 1.0. Note that compared to chimpanzee and bonobo genomic DNA, the copy number in human genomic DNA is (i) two-fold higher for *ARHGAP11*, consistent with the presence of the human-specific gene *ARHGAP11B* in addition to the common gene

ARHGAP11A; (ii) four-fold higher for *FAM72*, consistent with the presence of the human-specific genes *FAM72B*, *FAM72C* and *FAM72D* in addition to the common gene *FAM72A*; (iii) three-fold higher for *GTF2H2*, consistent with the presence of the human-specific genes *GTF2H2B* (black bold letters, not among the cNPC-enriched genes identified in this study) and *GTF2H2C* in addition to the common gene *GTF2H2A*; and (iv) two-fold higher for *SMN*, consistent with the presence of the human-specific gene *SMN2* in addition to the common gene *SMN1*.

Fig. 5. In-situ hybridization analysis of the mRNA levels of the human-specific cNPC-enriched protein-coding genes in the various zones of the fetal neocortical wall.

Coronal sections of human fetal neocortex (13wpc) were subjected to ISH using probes either specific for the mRNA of the human-specific gene under study (B, C, F) or recognizing the mRNAs of both the human-specific gene(s) and the paralog gene common to other primates as well (A, D, E, G, H, I, J). The five patterns of preferential mRNA expression in the various zones of the fetal neocortical wall (see labeling on the left and red dashed lines) are indicated above the images. Green, yellow and orange

boxes indicate areas of the VZ, SVZ and CP, respectively, that are shown at higher magnification in the respective images labeled ''', '' and ' on the right. Scale bars, 100 µm.

Fig. 6. In-situ hybridization analysis of the mRNA levels of 3 selected primate-specific genes in the various zones of the fetal human neocortical wall.

Coronal sections of human fetal neocortex (13wpc) were subjected to ISH using probes recognizing the mRNAs of either both, the human cNPC-enriched protein-coding gene under study that has orthologs specific to non-human primates (**A**, *PTTG2*; **C**, *KIF4B*) and the paralog gene common to non-primates as well (**A**, *PTTG1*; **C**, *KIF4A*), or the human cNPC-enriched protein-coding gene *MICA* that has orthologs specific to non-human primates and a paralog specific to primates, *MICB* (**B**). The pattern of preferential mRNA expression of the 3 genes under study in the various zones of the fetal neocortical wall (see labeling on the left and red dashed lines) is indicated above the images. Green, yellow and orange boxes indicate areas of the VZ, SVZ and CP, respectively, that are shown at higher magnification in the respective images labeled ''', '' and ' on the right. Scale bars, 100 µm.

Fig. 7. Comparison of the mRNA expression of 11 human-specific cNPC-enriched protein-coding genes with their ancestral paralogs in isolated cell populations enriched in aRG, bRG and neurons from fetal human neocortex.

A previously published genome-wide transcriptome dataset obtained by RNA-Seq of cell populations isolated from fetal human neocortex, that is, aRG (orange) and bRG (yellow) in S-G2-M and a fraction enriched in neurons but also containing bRG in G1 (N, purple) (Florio et al., 2015), was analyzed for the abundance of mRNA-Seq reads assigned to either the indicated human-specific gene(s) under study (blue background) or the corresponding ancestral paralog (white background), using the Kallisto algorithm. (**A**) Min-max box-and-whiskers plots showing mRNA levels (expressed in Transcripts Per Million, TPM); red lines indicate the median.

(**B**) Stacked bar plots showing the cumulative mRNA expression levels in the indicated cell types (sum of the median TPM values shown in (A)).

Fig. S1. Forced expression of *NOTCH2NL* in mouse embryonic neocortex increases cycling basal progenitors.

The neocortex of E13.5 mouse embryos was in utero co-electroporated with a plasmid encoding GFP together with either an empty vector (Control) or a *NOTCH2NL* expression plasmid (NOTCH2NL), all under constitutive promoters, followed by analysis 48 hours later.

(**A**) GFP (green) and Ki67 (magenta) double immunofluorescence combined with DAPI staining (grey) of control (left) and *NOTCH2NL*-electroporated (right) neocortex.

(**B**) Quantification of the percentage of targeted, i.e. GFP+, cells that are Ki67+ in the VZ, SVZ and IZ upon control (white bars) and *NOTCH2NL* (black bars) electroporation.

(**C**) GFP (green) and phosphohistone H3 (PH3) double immunofluorescence of control (left) and *NOTCH2NL*-electroporated (right) neocortex.

(**D**, **E**) Quantification of the number of ventricular (**D**) and abventricular (**E**) targeted (GFP+) cells in mitosis (PH3+) in a 200 µm-wide microscopic field upon control (white bars) and *NOTCH2NL* (black bars) electroporation.

(**A**, **C**) Images are a single 2-µm optical sections. Scale bars, 50 µm.

(**B**, **D**, **E**) Data are the mean of 10 embryos each, with 2-4 cryosections (**B**, 100 µm-wide microscopic field) per embryo counted and averaged. Error bars indicate SEM; *, $P < 0.05$; Student's *t*-test in **B** (VZ: $P = 0.914$, df = 18, $t = 0.108$; SVZ: $P = 0.012$, df = 18, $t = 2.731$; IZ: $P = 0.020$, df = 18, $t = 2.501$), **D** ($P = 0.718$, df = 18, $t = 0.3656$) and **E** ($P = 0.011$, df = 18 $t = 2.841$).

Fig. S2. *ARHGAP11B*-specific ISH probe.

(**A**) Nucleotide sequences at the exon 5 (purple background) – exon 6 (orange background) junction of the *ARHGAP11B* (top) and *ARHGAP11A* (bottom) mRNAs (note that U is depicted as T). The *ARHGAP11B* LNA ISH probe shown is are complementary to the nucleotides shown in red. The 55 nucleotides shown in green are unique to the 3'-end of the *ARHGAP11A* exon 5 and interfere with the binding of the LNA ISH probe to the *ARHGAP11A* mRNA, rendering the probe *ARHGAP11B*-specific.

(**B**) Images of COS-7 cells that were either untransfected, or transfected with either an *ARHGAP11A*- or *ARHGAP11B*-expressing construct and stained with the *ARHGAP11B* LNA ISH probe. Note that an ISH signal is detected only in *ARHGAP11B*-transfected

COS-7 cells, confirming the specificity of the LNA ISH probe for *ARHGAP11B*. Scale bar, 50 µm.

Fig. S3. Comparison of the paralog-specific mRNA expression between 11 human-specific cNPC-enriched genes and their respective ancestral paralog in aRG, bRG and neuron-enriched cell populations from fetal human neocortex.

(**A**) Diagram outlining the strategy used to ascertain paralog-specific mRNA expression in a given cell type of interest. mRNA sequences of an ancestral vs. a human-specific paralog (paralog A vs. B in the example shown) were aligned, and the homologous, yet distinct, core sequences of each alignment were extracted. The corresponding sequences of each paralog were used as a mapping reference for RNA-Seq reads from aRG, bRG and neuron-enriched cell populations from fetal human neocortex (Florio et al., 2015). Only reads aligning to "unique mappers", i.e. paralog-specific sites (SNPs or indels), were used for the analysis shown in (B). In the example shown, paralog-specific reads specific for paralog A or paralog B, as defined by the paralog-specific base (vertical yellow line) are colored in purple and orange, respectively.

(**B**) Bar plots showing the total numbers of paralog-specific RNA-Seq reads (identified as described in (A)) found in aRG vs. bRG vs. neuron-enriched (N) cell populations from fetal human neocortex (Florio et al., 2015). Grey bars indicate human-specific genes; black bars indicate their respective ancestral paralog. Data are the mean of four individual samples isolated from two human specimens; errors bars, SD.

Fig. S4. Cell-type specificity of mRNA expression of splice variants encoded by 10 human-specific cNPC-enriched genes.

Heatmaps showing TPM expression levels (see color keys) of all protein-coding splice variants encoded by the indicated human-specific cNPC-enriched genes in aRG, bRG and neuron-enriched (N) cell populations from fetal human neocortex (Florio et al., 2015). See Table S4 for mRNA expression data for each cell type and splice variant, including non-coding transcripts. Human-specific genes are grouped based on orthology, and splice variants (indicated by NCBI transcript IDs) encoded by the respective cNPC-enriched human-specific gene(s) are grouped together. Note the specific expression of ENST00000428041, a splice variant of *ARHGAP11B* uniquely expressed

in bRG. Splice variant-specific mRNA expression was assessed using the Kallisto algorithm.

Tables

Table 1 Primate-specific genes

Table S1 cNPC-enriched genes

Table S2 qPCR primer

Table S3 Primer for ISH probes

Table S4 mRNA expression data of splice variants

Supplementary files

File S1 ANKRD20A alignment

File S2 ARHGAP11 alignment

File S3 CBWD alignment

File S4 DHRS4 alignment

File S5 FAM72 alignment

File S6 GTF2H2 alignment

File S7 NOTCH2 alignment

File S8 ZNF98 alignment

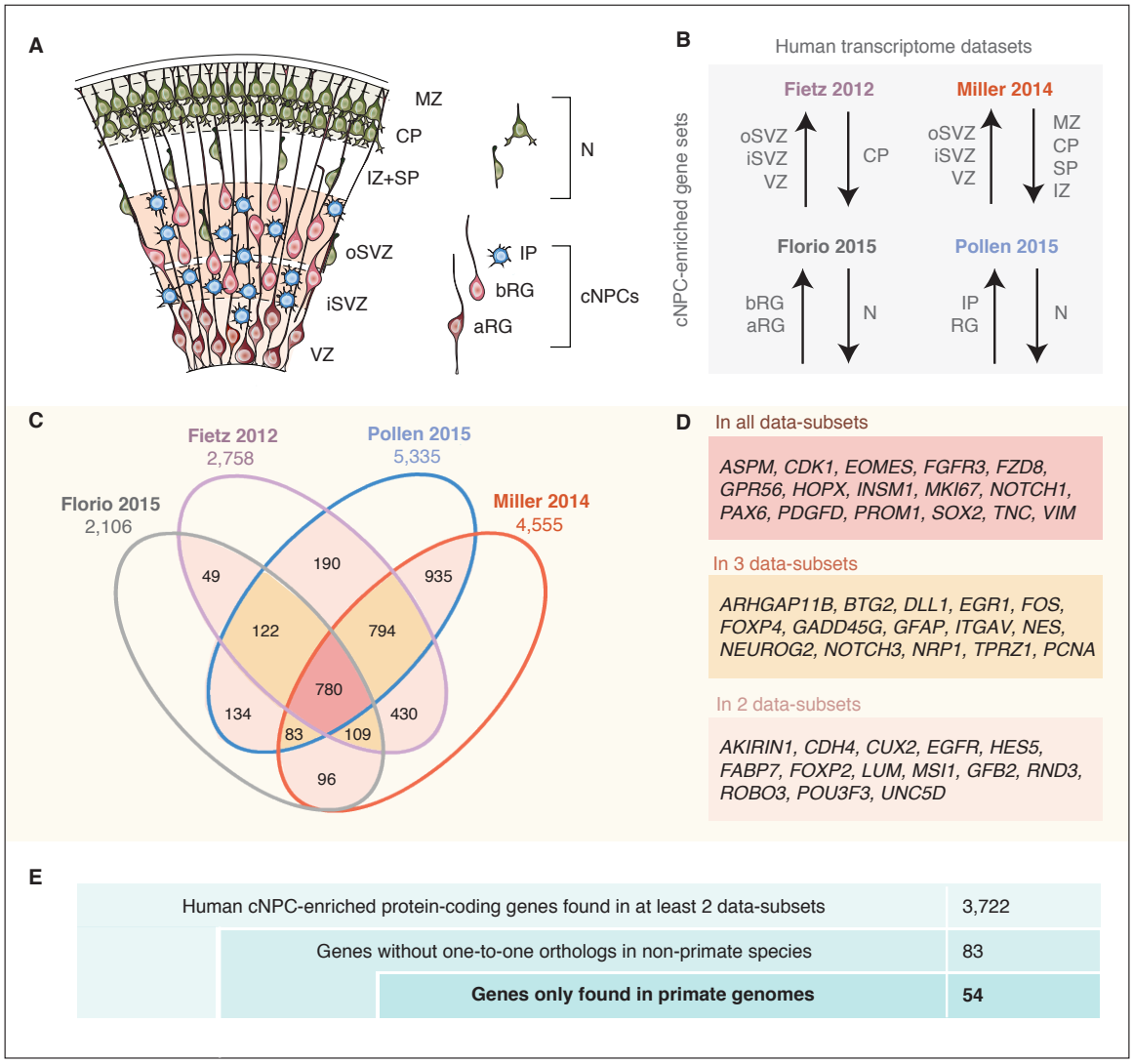

Figure 1

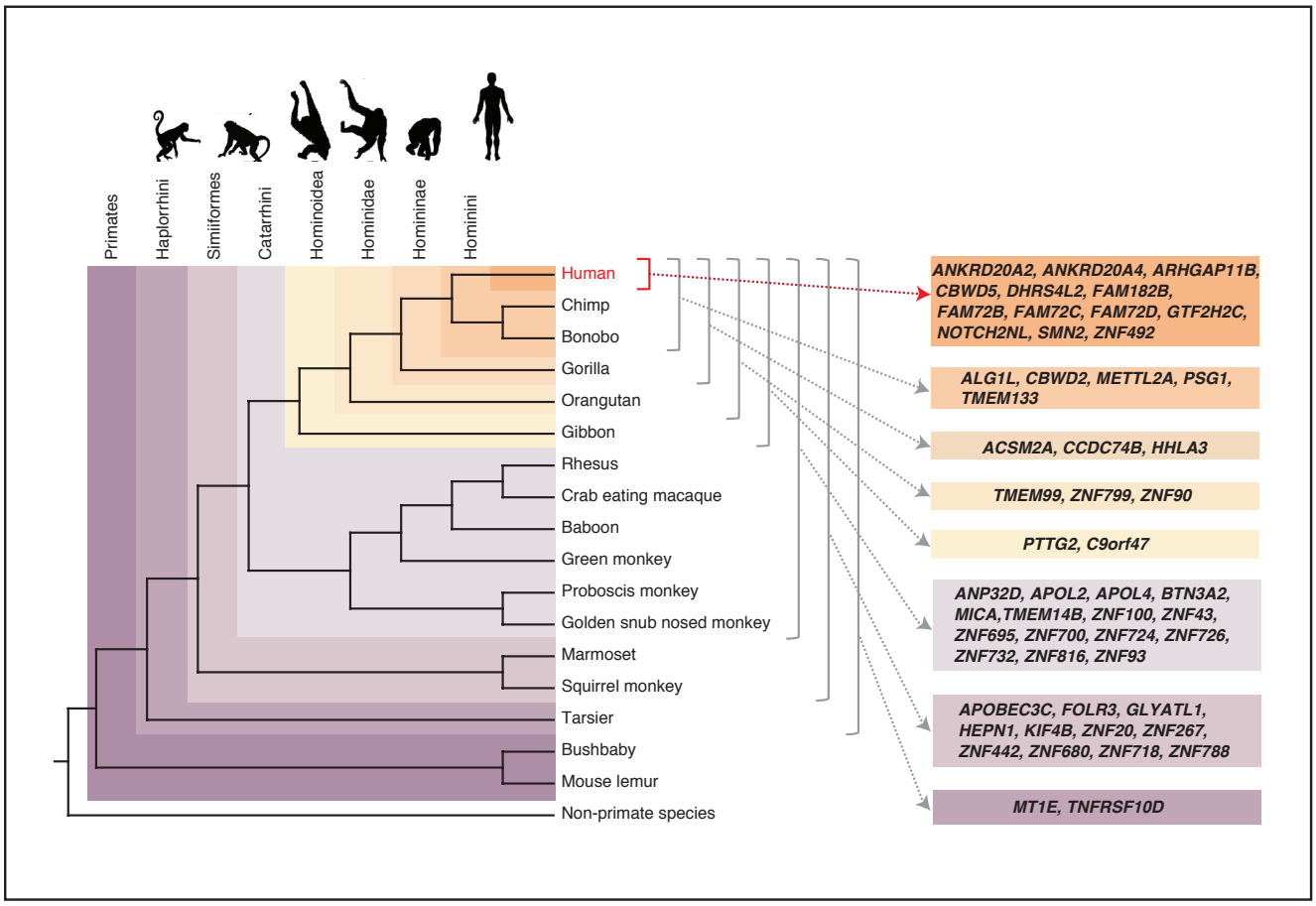

Figure 2

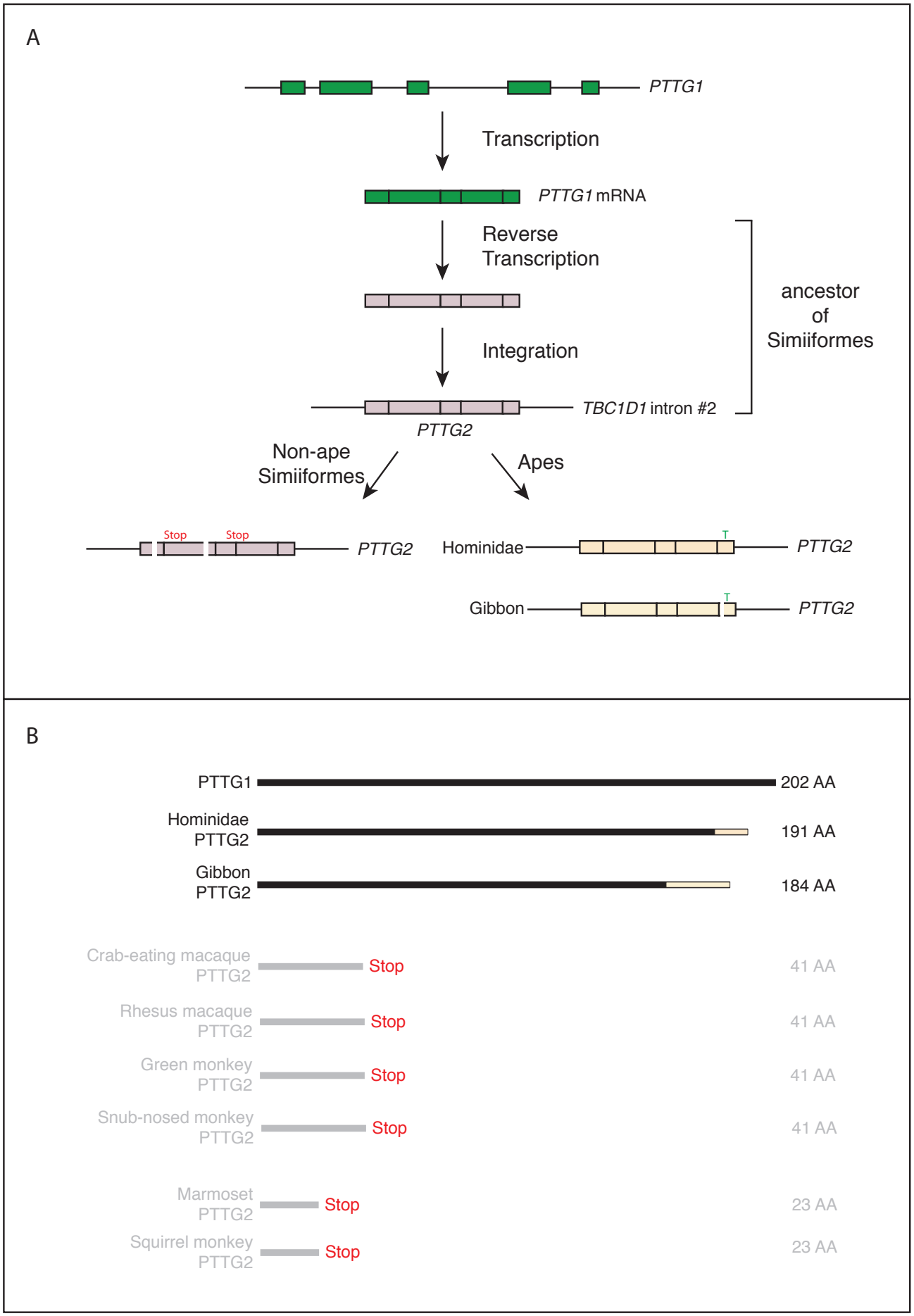

Figure 3

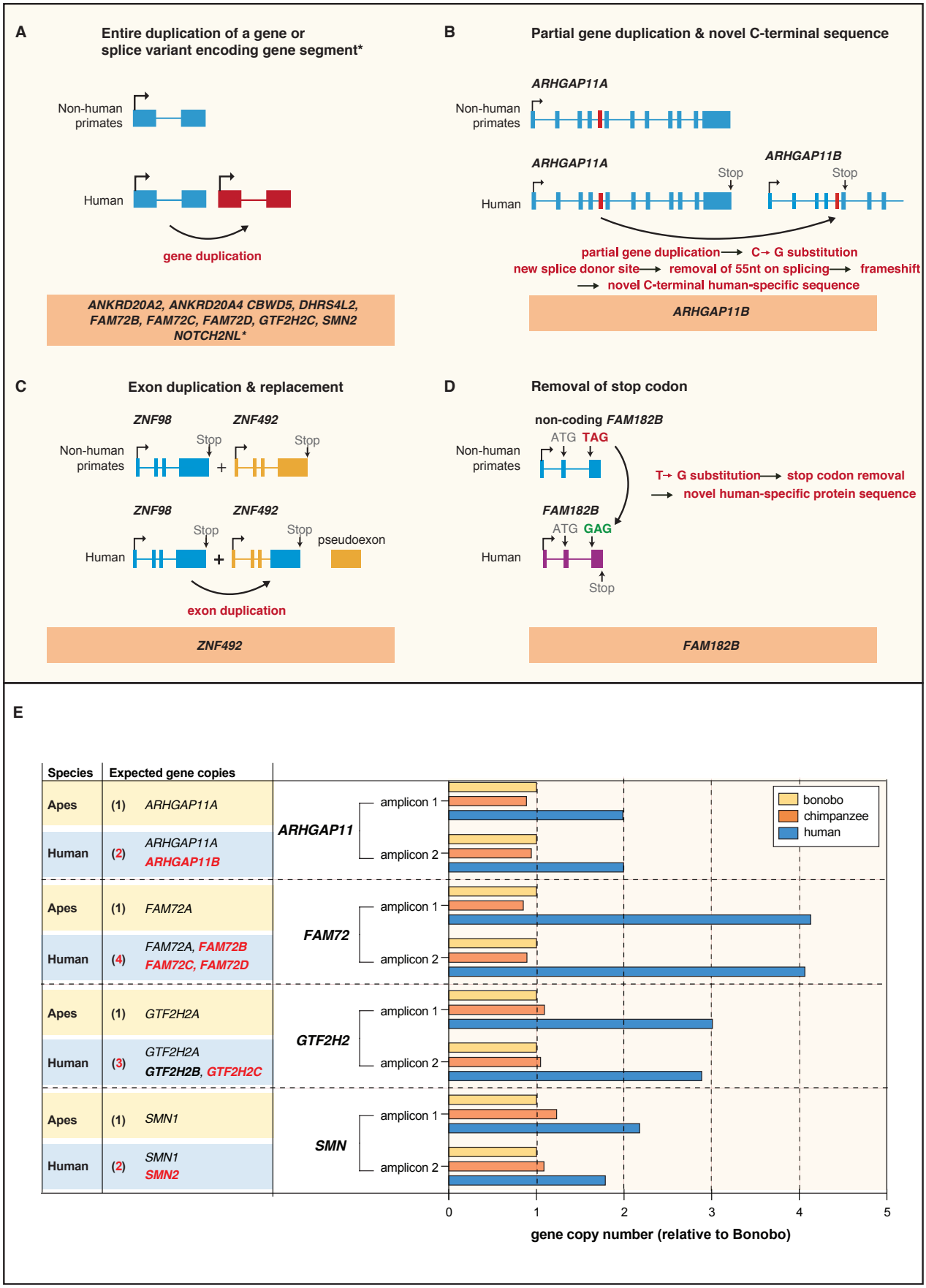

Figure 4

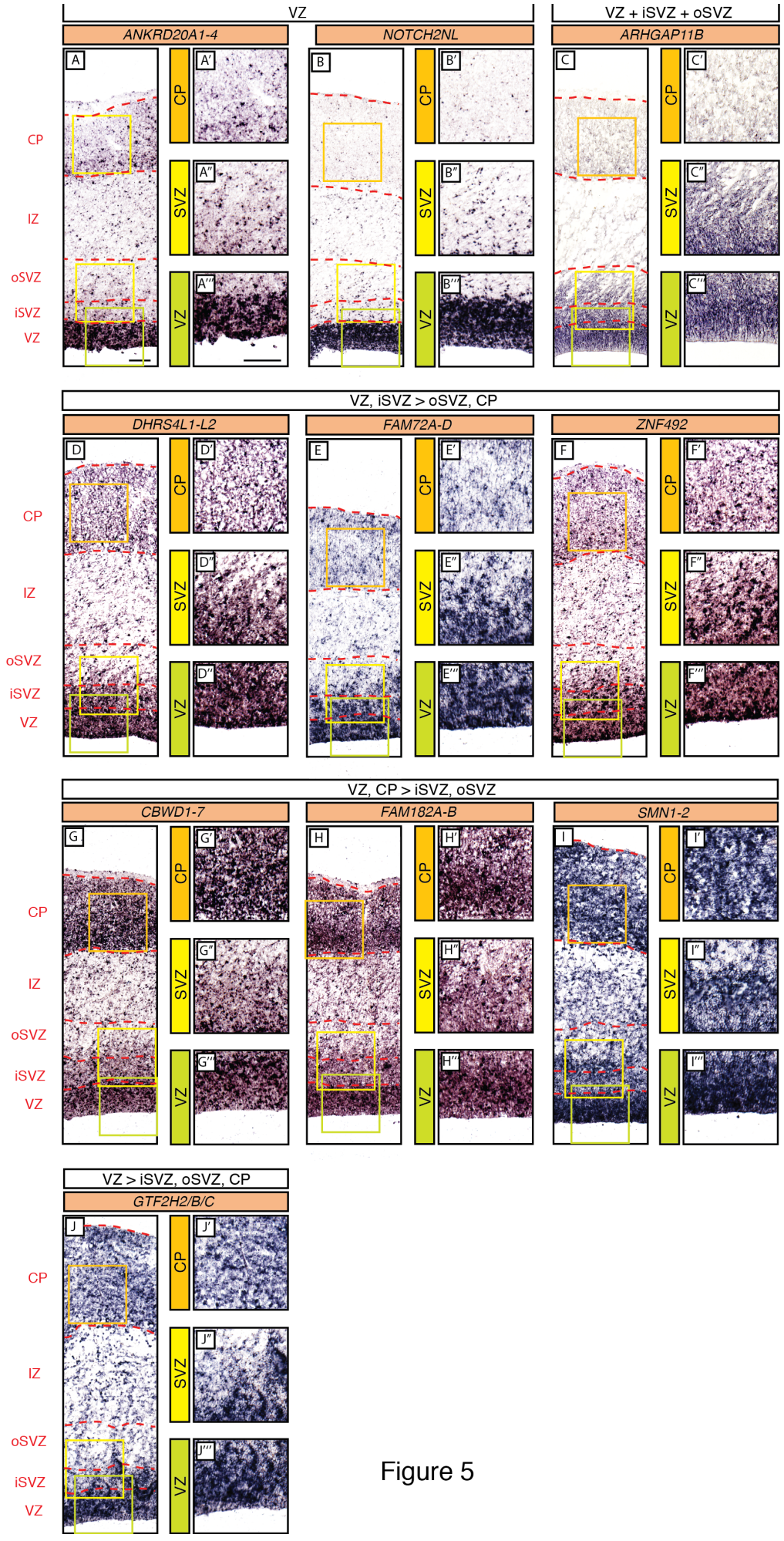

Figure 5

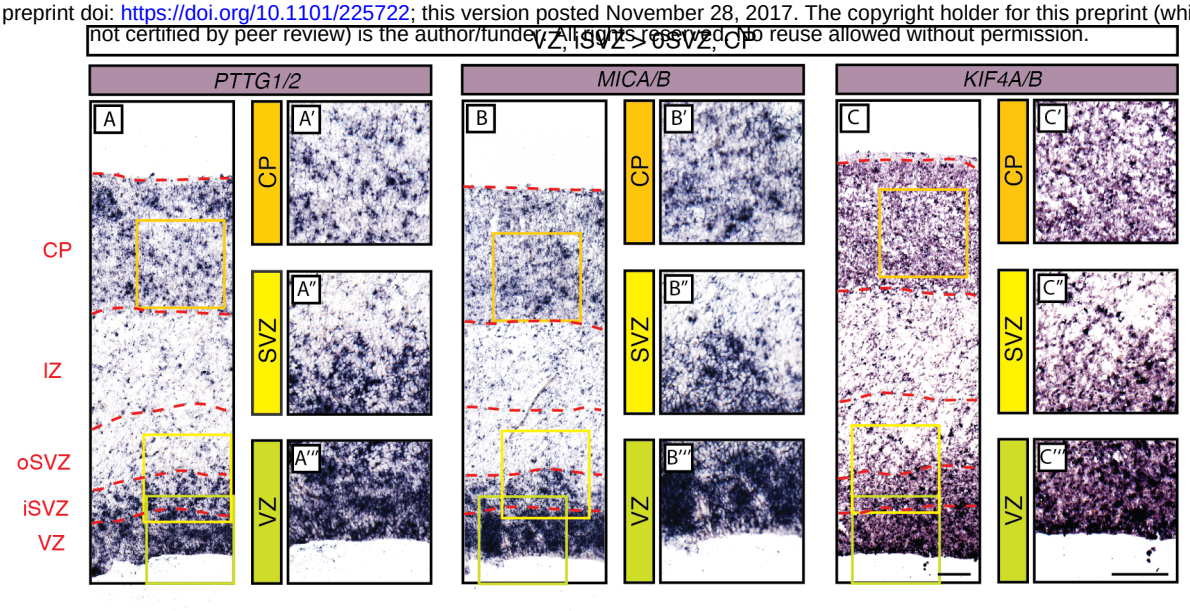

Figure 6

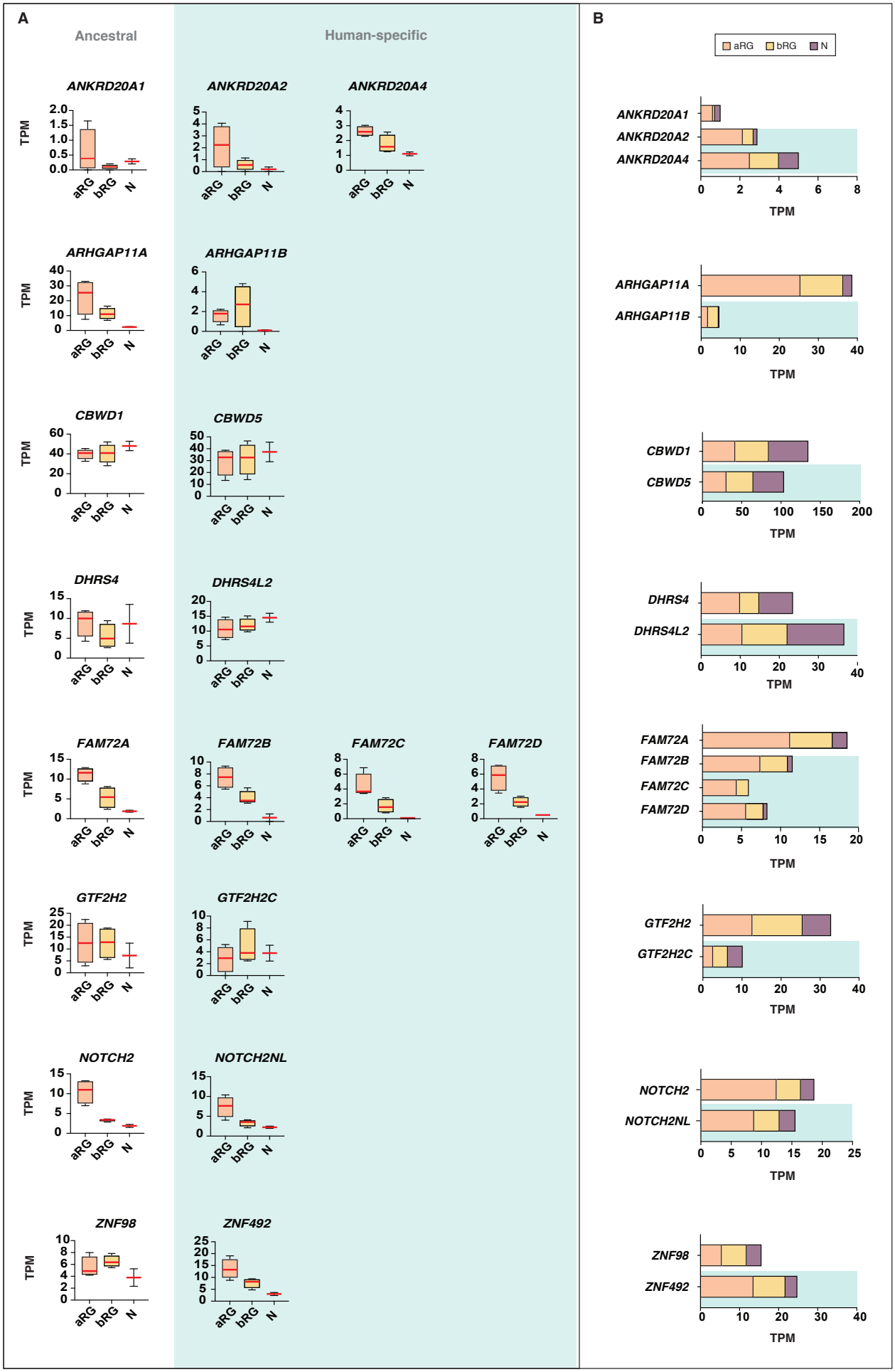

Figure 7

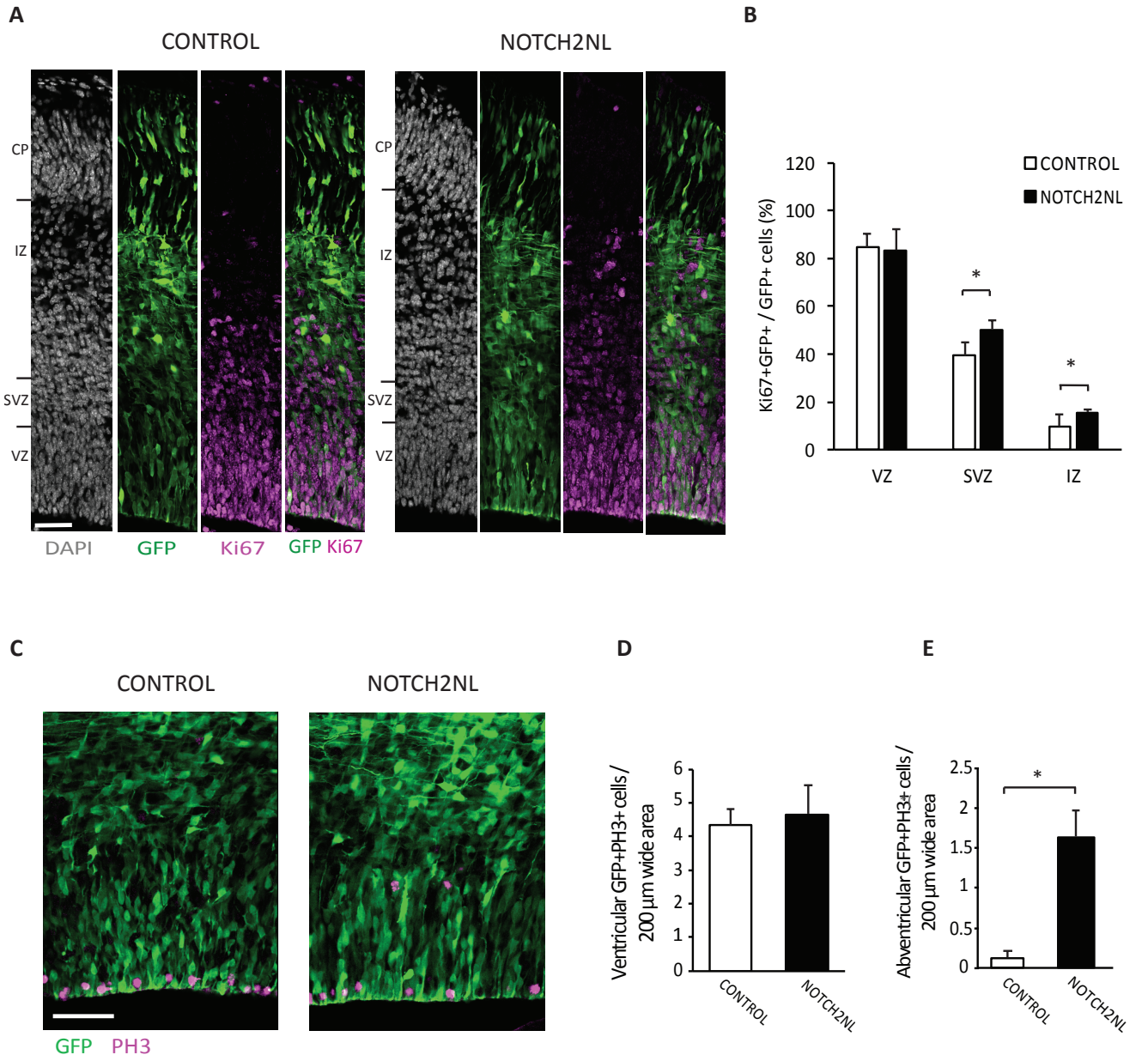

Figure S1

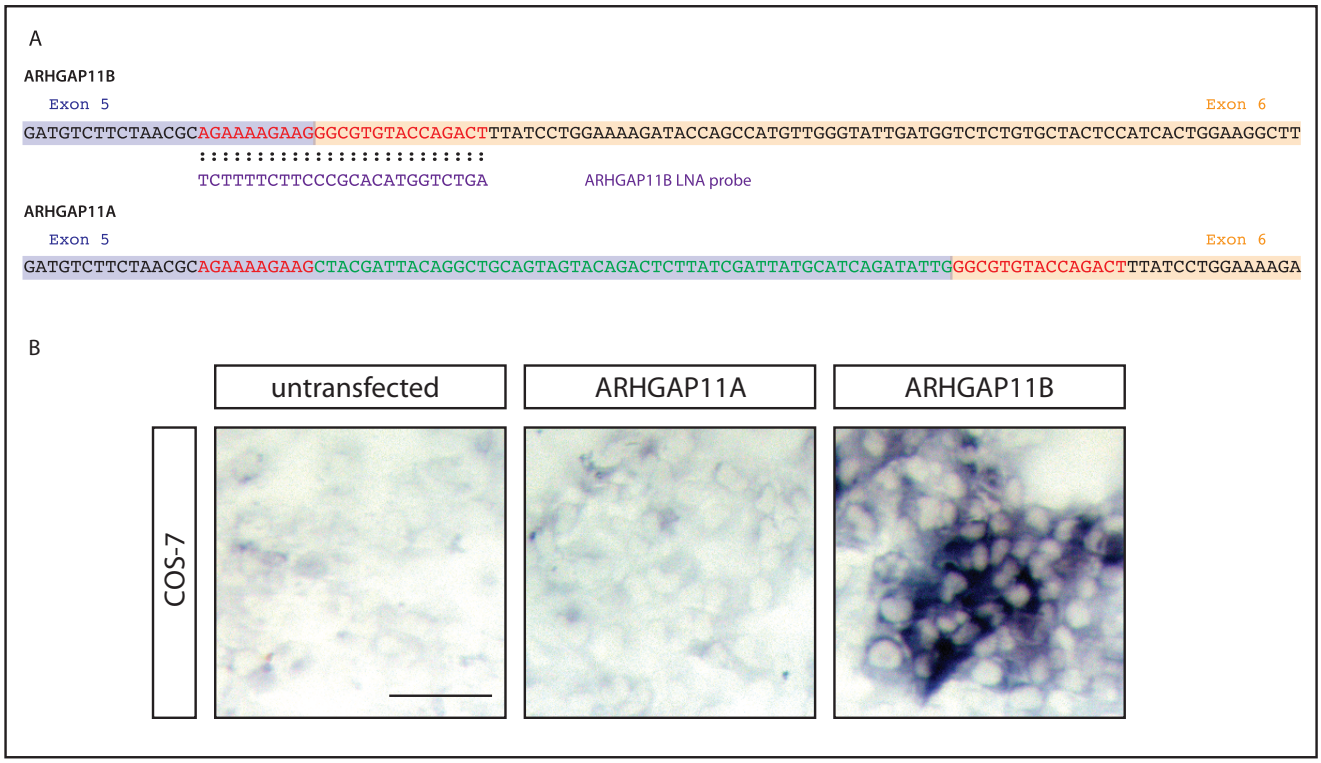

Figure S2

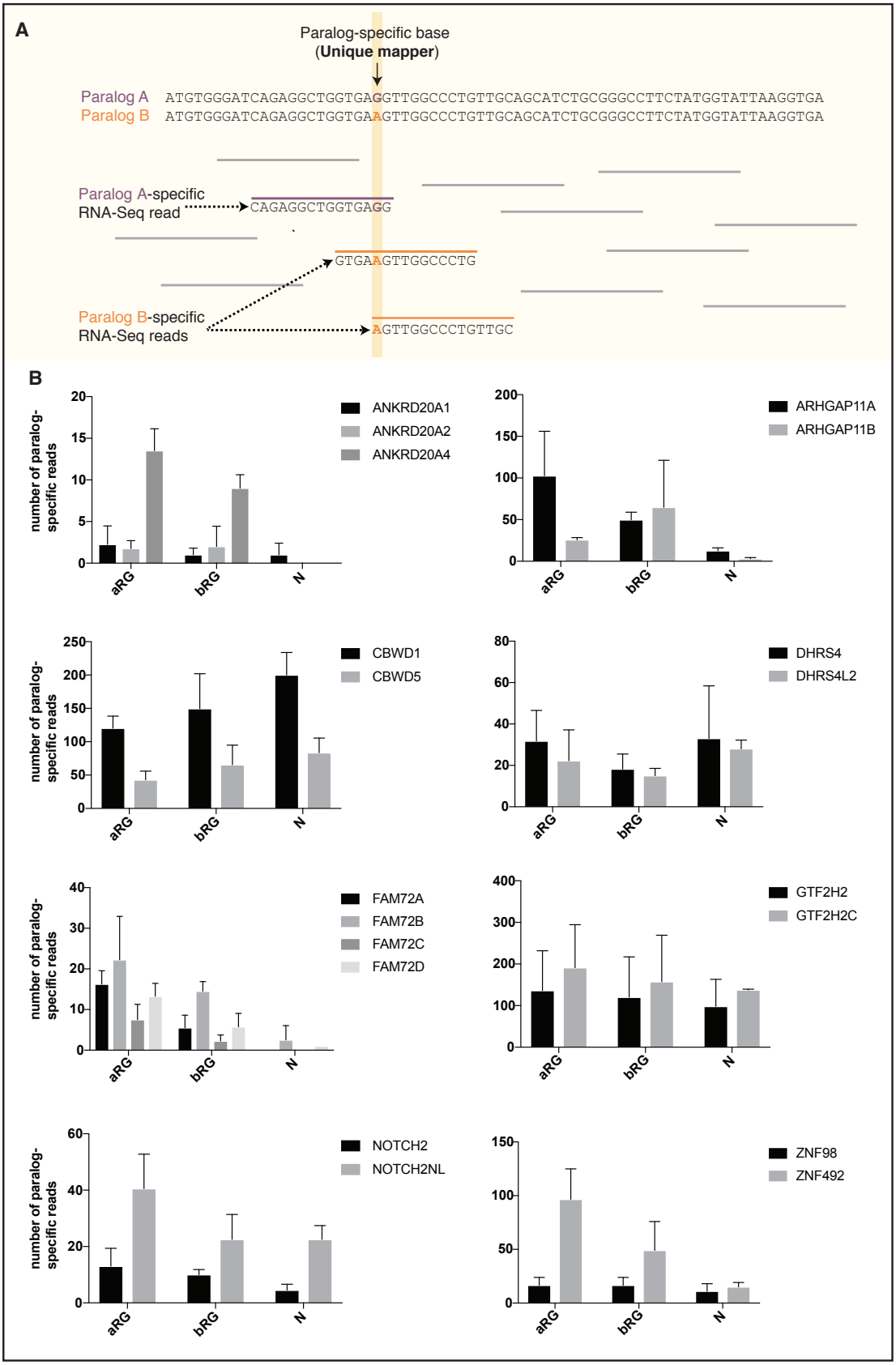

Figure S3

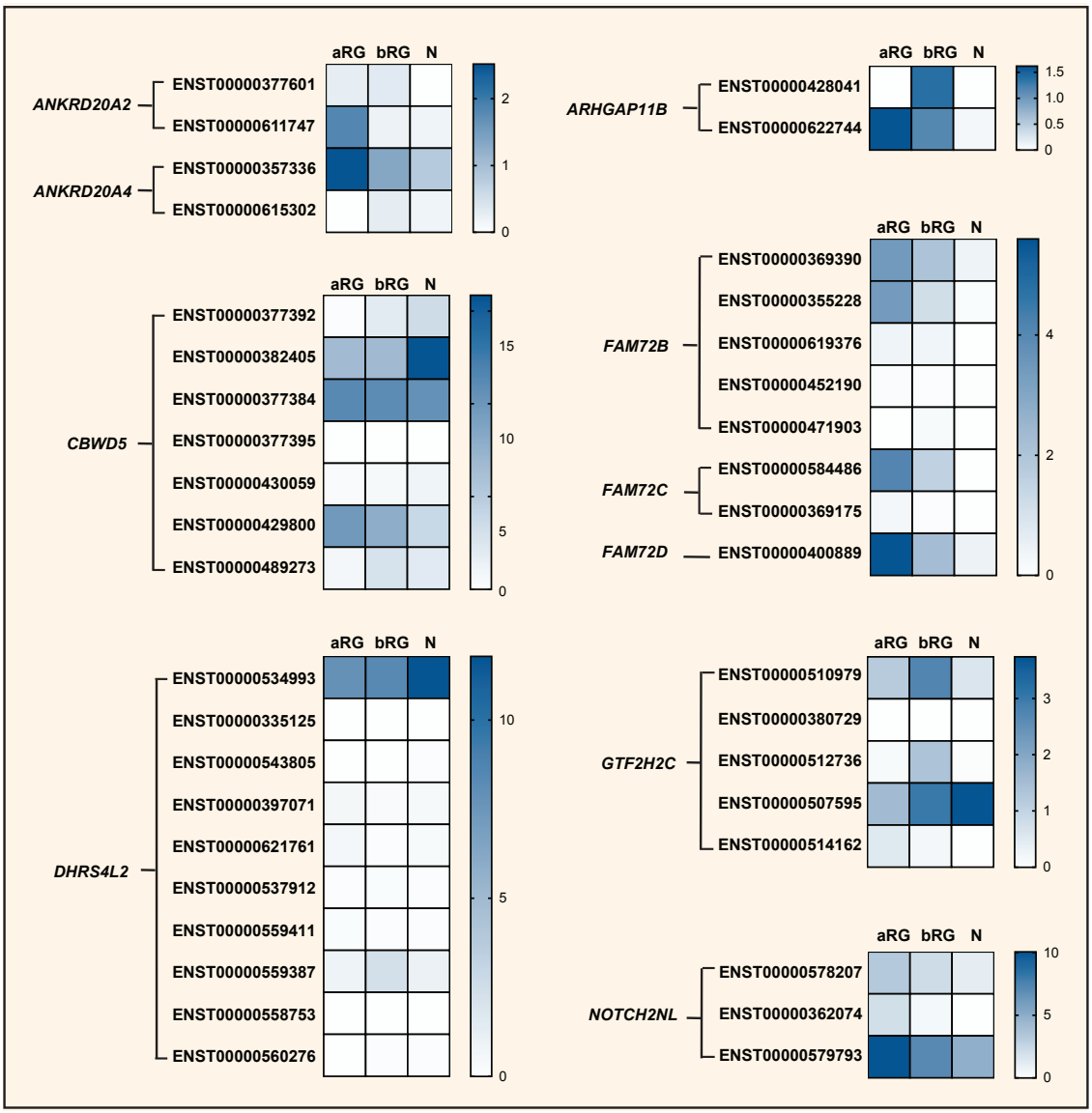

Figure S4