A Novel Gene Controls a New Structure: PiggyBac Transposable Element-Derived 1, Unique to Mammals, Controls Mammal-Specific Neuronal Paraspeckles

Tamás Raskó,[1] Amit Pande,[†,1] Kathrin Radscheit,[†,1] Annika Zink,[2] Manvendra Singh 🄳,[1] Christian Sommer,[1] Gerda Wachtl,[3,4] Orsolya Kolacsek,[3] Gizem Inak,[2] Attila Szvetnik,[1] Spyros Petrakis,[5] Mario Bunse,[1] Vikas Bansal,[6] Matthias Selbach,[1] Tamás I. Orbán,[3] Alessandro Prigione,[2] Laurence D. Hurst 🄳,*,[7] and Zsuzsanna Izsvák*,[1]

[1]Max Delbrück Center for Molecular Medicine in the Helmholtz Society, Berlin, Germany

[2]Department of General Pediatrics, Neonatology and Pediatric Cardiology, Medical Faculty, Heinrich Heine University, Duesseldorf, Germany

[3]Institute of Enzymology, Research Centre for Natural Sciences, ELKH, Budapest, Hungary

[4]Doctoral School of Biology, Institute of Biology, ELTE Eötvös Loránd University, Budapest, Hungary

[5]Institute of Applied Biosciences/Centre for Research and Technology Hellas, 57001 Thessaloniki, Greece

[6]Biomedical Data Science and Machine Learning Group, German Center for Neurodegenerative Diseases, Tübingen 72076, Germany

[7]Milner Centre for Evolution, Department of Biology and Biochemistry, University of Bath, Bath, UK

[†]These authors contributed equally to this work.

*Corresponding authors: E-mails: zizsvak@mdc-berlin.de; l.d.hurst@bath.ac.uk.

Associate editor: Irina Arkhipova

Abstract

Although new genes can arrive from modes other than duplication, few examples are well characterized. Given high expression in some human brain subregions and a putative link to psychological disorders [e.g., schizophrenia (SCZ)], suggestive of brain functionality, here we characterize *piggyBac* transposable element-derived 1 (*PGBD1*). *PGBD1* is nonmonotreme mammal-specific and under purifying selection, consistent with functionality. The gene body of human *PGBD1* retains much of the original DNA transposon but has additionally captured SCAN and KRAB domains. Despite gene body retention, PGBD1 has lost transposition abilities, thus transposase functionality is absent. PGBD1 no longer recognizes *piggyBac* transposon-like inverted repeats, nonetheless PGBD1 has DNA binding activity. Genome scale analysis identifies enrichment of binding sites in and around genes involved in neuronal development, with association with both histone activating and repressing marks. We focus on one of the repressed genes, the long noncoding RNA *NEAT1*, also dysregulated in SCZ, the core structural RNA of paraspeckles. DNA binding assays confirm specific binding of PGBD1 both in the NEAT1 promoter and in the gene body. Depletion of PGBD1 in neuronal progenitor cells (NPCs) results in increased NEAT1/paraspeckles and differentiation. We conclude that PGBD1 has evolved core regulatory functionality for the maintenance of NPCs. As paraspeckles are a mammal-specific structure, the results presented here show a rare example of the evolution of a novel gene coupled to the evolution of a contemporaneous new structure.

Key words: **PiggyBac transposon, transposase, cerebellum, evolution, novel gene, domestication, SCAN, KRAB, NEAT1, paraspeckle, transcriptional control.**

Introduction

Although duplication is a well understood route to the origin of new genes (Innan and Kondrashov 2010), there is increasing realization that de novo origination or cooption of inactivated transposable elements (TEs) present alternative paths (Kaessmann 2010; McLysaght and Guerzoni 2015; McLysaght and Hurst 2016). Unlike duplicates, de novo and TE-derived genes represent instances of genes derived from sequences that are not derived by ancestry from extant host genes. We thus consider these TE derived genes as both new (i.e., an addition to the prior set of genes) and novel (different from the prior set). By this classification, duplicates are new but not novel.

Unlike spuriously expressed noncoding sequence, TE-derived sequences are ripe for cooption/domestication (Kazazian 2004), as the TEs are most successful at propagating through a genome after horizontal transfer (HT)

Open Access

are those with the ability to function in the context of the host's genome. Indeed, the domestication of TEs often involves cooption of preexisting functionality. For example, *PiggyMac*, sharing sequence similarity to the *piggyBac* superfamily of transposases, utilizes the catalytic activity of the transposase, and performs regulated genome-wide self-splicing (Baudry et al. 2009). This process eliminates DNA to decrease genome complexity in ciliates (Cheng et al. 2010). TE-derived sequences can also provide transcriptional signals (e.g., enhancers, polyA, etc.) (Stocking and Kozak 2008; Lu et al. 2014; Wang et al. 2014) or structural elements, and affect gene expression at both DNA and RNA levels. For example, inverted repeated *Alu* elements (IRAlus) affect nuclear retention of mRNAs of host-encoded genes by generating secondary structure at 3′ UTRs of host genes (Chen and Carmichael 2009; Elbarbary and Maquat 2015; Hu et al. 2015; Torres et al. 2016). These IRAlu mRNAs are specifically recruited to certain nonmembrane-bound nuclear bodies notably paraspeckles. Paraspeckles are mammal-specific dynamic structures in the nucleus (Fox et al. 2002, 2018; Fox and Lamond 2010; An et al. 2019), consisting of a special set of coiled-coil domain (NONO/Paraspeckle, NOPS) containing RNA-binding proteins [e.g., splicing factor proline glutamine-rich protein (SFPQ), NONO, PSPC1], assembled around the long noncoding RNA (lnc)NEAT1 (Nuclear Enriched Abundant Transcript 1 isoform2). Despite such cooptions, in some instances the TE-derived sequences evolve new functionality (e.g., the adaptive immune systems of prokaryotes and vertebrates, CRISPR/Cas, and V(D)J recombination, respectively).

A deficiency in our current knowledge base of novel gene origination (Xie et al. 2019; Lange et al. 2021), is that most insights have been from genome-level studies with few well-characterized examples of both de novo (Chen et al. 1997; Cai et al. 2008; Xie et al. 2019; Lange et al. 2021) and TE domestication (reviewed in Jangam et al. 2017). Although genome scale studies can inform us to whether novel genes, after some period of retention, might be under purifying selection, and in which tissues they might be expressed, they cannot greatly inform us as to what they are doing. This requires case-by-case detailed analysis including genetic manipulations. Broader trends we hope would emerge on accumulation of enough detailed case histories.

In this context, one can take a targeted approach and start by identifying genes and phenotypes of possible interest. For humans, brain functioning is surely one of the most interesting of phenotypes. Although the hominoid brain-specific retroposed GLUD2 is an example of a new gene involved in neuronal activity (Burki and Kaessmann 2004), there are no examples of novel genes that incorporate into the processes that make humans particular, most notably our neuronal functioning, the cerebellum being a focus of selection (Barton and Venditti 2014). In this context, we identified as a gene of interest PGBD1 (Bouallegue et al. 2017), one of five human genes that are related to the

piggyBac transposon (*PB*) (hence the name, *piggyBac* TE-derived 1–5, PGBD1–5). The ancestral *PB*-like transposons were transferred horizontally to vertebrates in multiple waves (Pavelitz et al. 2013), the cabbage looper *PiggyBac* element probably being a good model of this ancestral form (Cary et al. 1989; Ding et al. 2005; Wu et al. 2006; Wilson et al. 2007).

We identified PGBD1 as a gene of interest for several reasons. First, PGBD1 (along with PGBD2) appears to be mammal-specific (Bouallegue et al. 2017). PGBD5 by contrast is seen widely within the vertebrates (Bouallegue et al. 2017). Second, PGBD1 might not just be functional (as evidenced by Ka/Ks < 1, Bouallegue et al. 2017), but may have evolved some core functionality in the brain. Notably, as with PGBD3 and PGBD5 (Fattash et al. 2013; Pavelitz et al. 2013; Henssen, Koche et al. 2017; Henssen, Reed et al. 2017), mutations within PGBD1 may be disease-associated. Genome-wide association studies (GWASs) in independent studies identified single-nucleotide polymorphisms in the first intron of PGBD1 (Stefansson et al. 2009; Yue et al. 2011; Prata et al. 2019), consistent with it possibly being a susceptibility locus for schizophrenia (SCZ). Given that expression is enriched, among other tissues, in the cerebellum and the cerebellar hemisphere of the brain (Schmahmann 2004), this may well be more than coincidence.

Third, *prima facie* evidence suggests that PGBD1 may be unusual in not having evolved new functionality by adopting its prior functionality. Cut-and-paste transposons like *PiggyBac* are characterized by a transposase gene flanked by inverted repeats (IRs). The transposase catalyzes DNA cleavage during the cut and paste process via recognition motifs in these IRs. However, in contrast to PGBD3 and PGBD4, PGBD1 does not have the terminal IR sequences (Sarkar et al. 2003; Newman et al. 2008). Similarly, PGBD1, unlike its contemporary close relative PGBD2, is missing an intact C-terminal cysteine-rich domain (CRD: see fig. 1B) that enables DNA binding as part of the transposition process (Morellet et al. 2018). However, it is questionable whether the CRD domain of PGBD2 binds DNA (Guerineau et al. 2021). Furthermore, unlike other PGBD genes, the transposase-derived ORF of PGBD1 is fused to an upstream exon of a SCAN-domain (Sarkar et al. 2003), and belongs to the SCAN family of transcription factors (supplementary fig. S1B, Supplementary Material online). The SCAN domain can function as a protein interaction domain with other SCAN family members, sometimes including self-association (Williams et al. 1999). PGBD1 thus has been annotated as a protein of unknown function, consisting of a SCAN and a transposase-derived domain (Uniprot) (Sarkar et al. 2003; Bouallegue et al. 2017). Here, then we investigate the evolution and function of PGBD1 and report a serendipitous discovery, namely that PGBD1 regulates a mammalian specific structure, the paraspeckle. This makes it a very rare example of a new gene that regulates a contemporary new structure.
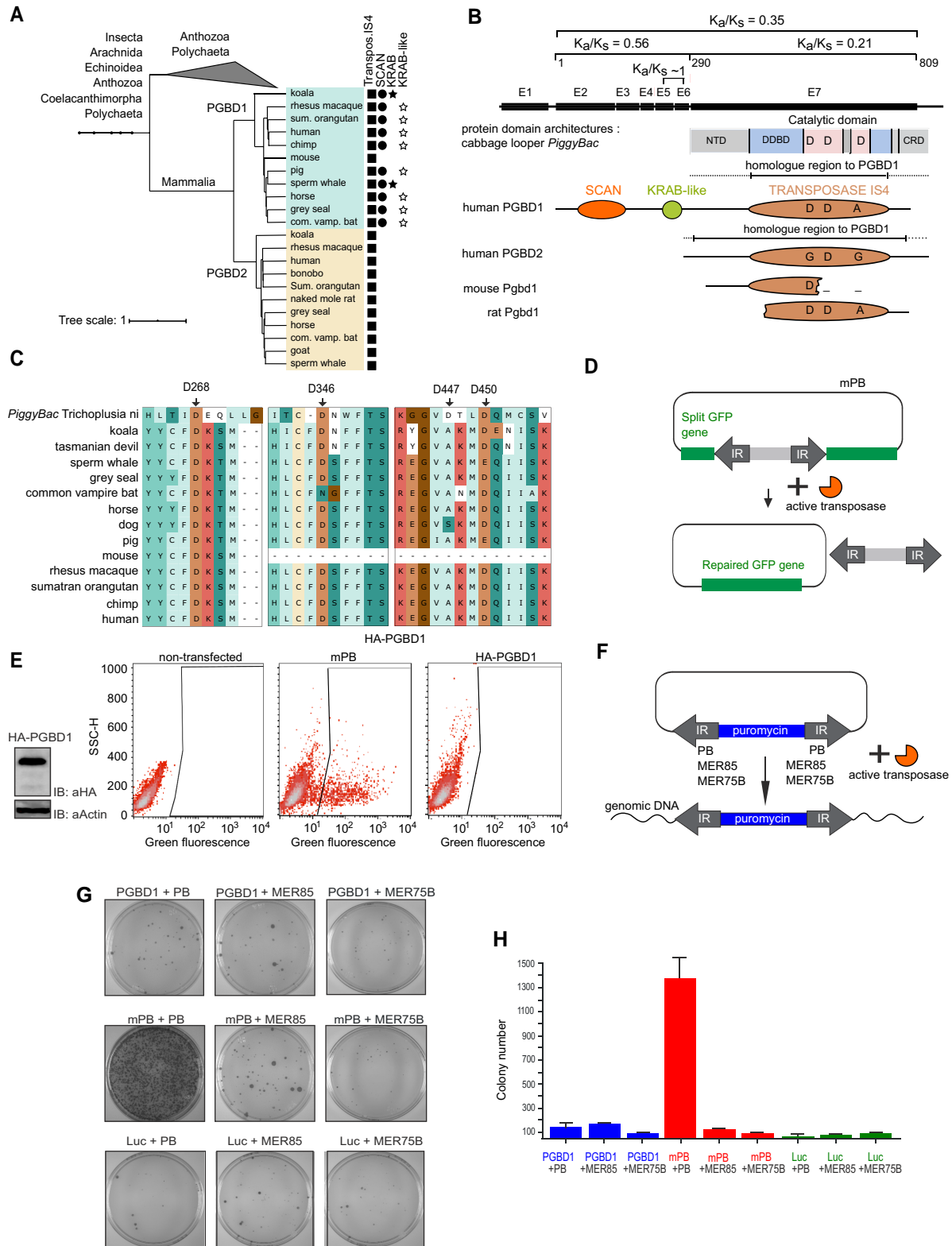
Fig. 1. The domesticated PGBD1 possesses a SCAN-, KRAB- and transposase-derived domains, but has no catalytic activity as a transposase. (A) Phylogenetic tree of PGBD1 and PGBD2. The presence of the transposase-derived, the SCAN, and KRAB domains are shown. The human PGBD1 and PGBD2, with the most closely related sequences (containing transposase IS4) were aligned with *muscle* and a tree was built using *MrBayes*. Protein domains were annotated with hmmerscan and CDD (NCBI). The KRAB domain was annotated with Phyre2. (B) PGBD1 domain structure in comparison to *PiggyBac* of the cabbage looper moth, human PGBD2, rat PGBD1, and mouse PGBD1. The transposase-derived domain (IS4) includes dimerization and DNA binding domains (DDBD) as well as the catalytic domains of PiggyBac (Chen et al. 2020). NTD, N-terminal domain; CRD, C-terminal cysteine rich domain; E1–7 are exons 1–7. The "D"s in the transposase-derived domains represent the catalytic triad DDD (D268, D346, D447). D447 is replaced by (A) in PGBD1. PGBD2 and PGBD1 are highly similar (average pairwise similarity score of ~63% the aligned region which spans 1324 bp exceeds the borders of the annotated transposase IS4 domain, calculated by distance matrix of

3

Results

PGBD1 is a Mammal-Specific Horizontally Transferred Gene

From homology searching, PGBD1 has previously been considered to be mammal-specific (Bouallegue et al. 2017). However, taxonomic presence/absence inferred by homology searching may simply reflect the limits of the methodology. Using a method to determine whether absence of detectable homology is likely to be a failure of homology searching (Weisman et al. 2020), we find that PGBD1 (along with PGBD2) is nonmonotreme mammal-specific, as previously suggested (Bouallegue et al. 2017): probability of homolog detection failure $= 0$ ($E = 0.001$), 99% confidence interval, $a = 1724.6$, $b = 1.18$, $r^2 = 0.98$ (supplementary fig. S1A, Supplementary Material online). As PGBD1/2 show homology to arthropod PGBD sequences (supplementary figs. S1A and S2, Supplementary Material online) HT early in mammalian evolution is the most likely route to origination.

PGBD1 has Gained (and Sometimes Lost) Domains

PGBD1 is more complex than a simple HT event as PGBD1, unlike PGBD2, has acquired a SCAN domain. With no evidence for the SCAN domain or homology across the relevant region in any PGBD2 ortholog (fig. 1A and B and supplementary fig. S3, Supplementary Material online), SCAN acquisition happened once by PGBD1 in the common ancestor of eutherians and marsupials shortly after the duplication (or parallel integration) event. Indeed, PGBD1 is found in a genomic domain rich in SCAN domain proteins suggesting gene fusion after integration into this site.
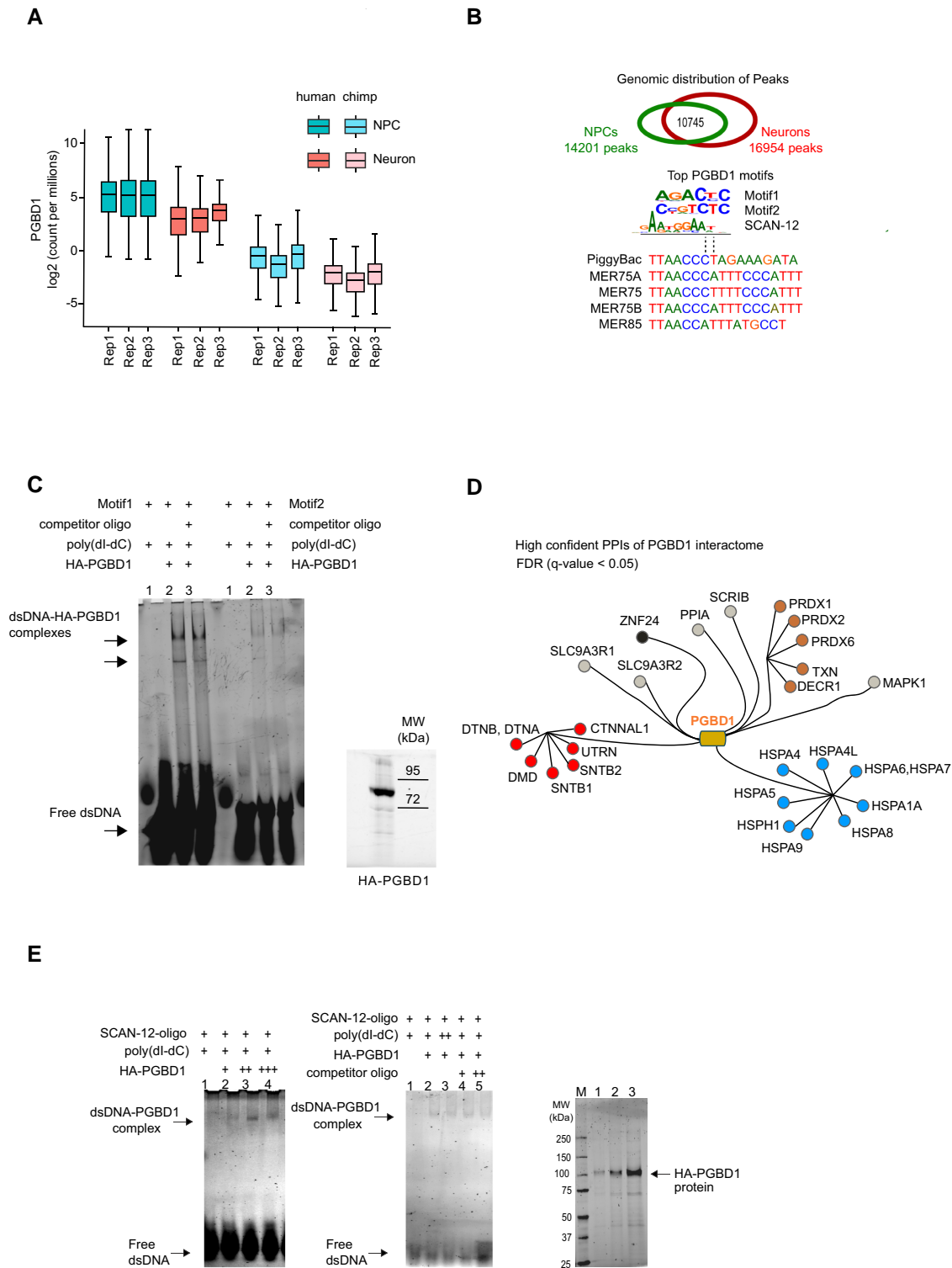
In silico methods also suggest that there have been multiple independent losses of the SCAN domain for example in rodents, cats, gray lemurs, and some marsupials (fig. 1A and B and supplementary figs. S2 and S3, Supplementary Material online). To validate one such loss, we examined rat Pgbd1 in detail. Although we observe a relatively high homology between the transposase domains of the murine and human PGBD1s (87%) (supplementary fig. S4A, Supplementary Material online), we could not identify the SCAN domain by HMMERsearch, nor could we observe homology to the SCAN domain within multiple sequence alignment when employing the annotated sequences of mouse and rat (fig. 1A and supplementary fig. 4B, Supplementary Material online). Loss of the SCAN domain was confirmed by cloning the 5′ end of the rat Pgbd1 gene following reverse transcription of total RNA isolated from a rat cell line (supplementary fig. S4C and D, Supplementary Material online). This identified several STOP codons upstream of the transposase-like domain at the genomic locus (supplementary fig. S4E, Supplementary Material online). The murine loss can be dated to post the common ancestor between murines and fellow rodents, ground squirrels (supplementary fig. S3A, Supplementary Material online). This has the side consequence that standard rodent models cannot be employed to investigate human PGBD1's biology. Here then we employ human cell lines for analysis.

Aside from the SCAN domain we also identified a KRAB-like domain (fig. 1A and B and supplementary fig. S4F, Supplementary Material online). The KRAB domain is also part of the KRAB-ZNF (Zn-finger) family of sequence-specific transcriptional regulators, involved in cell differentiation and development (Nowick et al. 2013) (supplementary fig. S1B, Supplementary Material online) and can mediate repression of transcription (Margolin et al. 1994; Witzgall et al. 1994). Automated protein domain search algorithms could detect the KRAB domain in the PGBD1 protein in only a few species (fig. 1A). Species with the KRAB domain include marsupials (e.g.,

Ugene). Note that the ZN-finger containing CRD domain, required for ITR binding in the piggyBac transposase is missing in PGBD1 (Morellet et al. 2018). The PGBD1 sequences in rodent animal models are truncated, resulting in degenerated copies. The Ka/Kv values of the entire PGBD1 as well as for various subdomains are shown. Note the ~1 value for the KRAB domain [overall $= 0.35$, N-terminal (aa 1–290) $= 0.56$, C-terminal (aa 291–809) $= 0.21$, SCAN (aa 40–142) $= 0.32$, KRAB (aa 211–267) $= 1.02$, DDBD1 (aa 405–541) $= 0.19$, DDBD2 (aa 750–804) $= 0.26$, catalytic domain 1 (aa 541–651) $= 0.14$, and catalytic domain 2 (aa 726–750) $= 0.07$, reference is the human amino acid sequence of PGBD1. (C) Protein sequence alignment of the transposase-derived DDD catalytic domain of PGBD1. The first raw of the alignment shows the corresponding sequence of the piggyBac transposase, identified in Trichoplusia ni (cabbage looper moth). The alignment includes koala and gray seal, from where the KRAB domain was reported (supplementary fig. S4F, Supplementary Material online), and various mammalian species. The conserved amino acids D268/D346/D447 of the conserved DDD catalytic domain and D450 of the piggyBac transposase are arrowed (Sarkar et al. 2003). The numbers refer to their position using the piggyBac amino acid sequence as reference. (D) Transposon excision repair assay detects no activity of the PGBD1. Schematic representation of the reporter assay of PiggyBac excision. The PiggyBac transposon (flanked by inverted terminal repeats, ITRs) splits the coding sequence of the GFP reporter. In the presence of an active transposase, transposon excision occurs, and the readout is the restored GFP reporter signal. (E) Quantitative FACS (fluoresccence-activated single cell sorting) analysis of GFP positive cells generated in the transposon excision repair assay. (Left panel) Western blot analysis of the HA-tagged PGBD1 (HA-PGBD1) protein tested in the excision repair assay. HeLa cells were cotransfected with plasmids harboring HA-tagged PGBD1 along with the reporter construct. Nontransfected HeLa and cells transfected with mPB (mammalian codon optimized piggyBac transposase) along with the reporter served as controls (right panel). Transposition assay detects no activity of the PGBD1 protein. (F) Schematic representation of the colony forming transposition assay to detect stable integration of the puromycin resistance gene marked reporter in HEK293 cells. In case of active transposition, the transposase cuts at the IRs—(terminal IRs), and inserts the reporter-marked transposon into the genome, providing antibiotic resistance for the transfected cells. In addition to the piggyBac ITRs, reporters were also built with PiggyBac-derived miniature IR (MITEs) ITRs of MER75B, MER85 (see also fig. 2B). (G) Puromycin resistant HEK293 colonies are shown as the readout of the assay. The constructs were transfected in various combinations. HEK293 cells transfected with the mPB transposase and the nonrelevant Luciferase expression construct (Luc) along with the reporter served as controls. (H) Quantification of the transposition assay. Colonies were quantified in a 75S model gel imager, using the Quantity One 4.4.0

A

B

Genomic distribution of Peaks

C

D

High confident PPIs of PGBD1 interactome
FDR (q-value < 0.05)

E

Fig. 2. PGBD1 target genes influence human neural progenitor identity. (A) Relative expression levels of PGBD1 in human and chimpanzee NPCs and neurons (GSE83638). Note that the in cross-species comparison the expression of PGBD1 has a higher expression in humans versus chimpanzee, and it is expressed at a higher level in progenitor versus differentiated cells. Variable box widths show the data distribution in all replicates. NPC $n = 3$ and Neuron $n = 3$. FPKM (Fragments Per Kilobase Million) value was normalized using Trimmed Mean of M-values (TMM) and converted to log2 for calculating fold change at P-value <0.05 (Benjamini–Hochberg). (B) ChIP-exo analysis of PGBD1 binding motifs with an overlapping, but specific distribution pattern in NPCs and in differentiated neurons. (Upper panel) Venn diagram shows the number of identified PGBD1 ChIP-exo peaks in hESC-derived NPCs, in hESC-derived neurons and common to both. (Middle panel) The top two PGBD1 motifs (Motif1 and Motif2) derived from overrepresented sequences identified from both NPCs and neurons. SCAN-12 represents the shared consensus DNA-binding motif identified between PGBD1 and 12 SCAN-ZNF-proteins, including ZNF167, ZNF174, ZNF18, ZNF232, ZNF274, ZNF394, ZNF483, ZNF496, ZNF500, ZNF498, ZNF187, ZNF323. (Lower panel) Sequence alignment of the *piggyBac* inverted repeat sequences (ITRs) and ITR-like motifs of the *piggyBac* transposon-derived miniature inverted repeat elements (MITEs) of MER75A, MER75, MER75B, and MER85. Note the lack of similarity between PGBD1 ChIP motifs and the *PiggyBac* or the MITE ITR sequences. (C) Electromobility shift assay (EMSA) confirms

koala) (fig. 1A and supplementary fig. S4F, Supplementary Material online), supporting the hypothesis that KRAB inclusion is, like SCAN inclusion, the ancestral condition but with numerous loss/decay events. With protein structural prediction server Phyre2 (Kelley et al. 2015), we could detect structural similarities between the KRAB domain template and PGBD1 proteins, which we call KRAB-like domains.

Although we find evidence that structurally PGBD1 is somewhat variable, we can also ask whether the protein and its subdomains are largely under purifying selection, and hence likely to be functional to the host. To investigate evolutionary forces, PGBD1 mRNA (CDS) sequences from 11 mammalian and 18 primate species were analyzed with PAML (Yang 1997). We find that PGBD1 is largely under purifying selection, indicative of being a functionally important gene with an overall Ka/Ks ratio ≤ 0.35 (fig. 1B).

For the KRAB domain, however, our analysis of branch-specific ratios revealed several organisms with Ka/Ks > 1 (horse, gray seal, common vampire bat, and primates). To test if the evolution of the KRAB (-like) domain might be better explained by neutral or adaptive forces, we tested the two hypotheses and found that adaptive evolution explains the data better than neutral evolution ($\chi^2 = 7.47$, df = 2, P-value = 2.38e−02). Given the broad evolutionary distances (and associated problems of synonymous site saturation), we repeated the analysis exclusively in primates (Ka/Ks ~0.44 and 1.13, overall protein and KRAB region, respectively). We find that adaptive evolution explains the processes in this region significantly better ($\chi^2 = 7.43$, df = 1, P-value = 6.42e−03) than neutral evolution. In particular, two sites were identified positively selected, beneficial mutations: 227V (prob ~ 0.98) and 252M (prob ~ 0.96, naive empirical bayes, reference: human full-length aa sequence). Although 252M is typically not well conserved in other KRAB domains, and its functional relevance is unknown, position 227 is a well conserved F in other KRAB domains, and is important for binding the TE suppressor TRIM28 (Peng et al. 2009). We conclude that the gene is functional and mostly under purifying selection, but in one sub region in some lineages it is functional and subject to positive selection for reasons unknown.

PGBD1 has Lost Transposase Activity

TEs have transposase activity key to their successful genomic colonization. This same functionality can be coopted on domestication (Baudry et al. 2009; Gray et al. 2012). However, in PGBD1, we observe a replacement by alanine (A) at the third aspartic acid (D) within the DDD motif of the catalytic domain (D447A) responsible for the transposition reaction (fig. 1C). As mutation at D447 abolishes catalytic activity (Sarkar et al. 2003; Keith et al. 2008), reduced functionality is to be expected. This loss is also seen in koala and Tasmanian devil (fig. 1C), indicating an early loss event. To test for transposase activity, we first used a tissue culture-based excision assay (Stoilov et al. 2008) that restores the open reading frame of the green fluorescence protein reporter, following the excision of the piggyBac (PB) transposon (fig. 1D). As PGBD1 has no obvious IRs flanking the transposase-derived sequence, we utilized the piggyBac IRs. Using the mPB transposase (Cadinanos and Bradley 2007) as a positive control, we detected no transposon excision activity in the presence of the human PGBD1 (fig. 1E). As the assay requires precise excision of the transposon, we also performed a less restrictive transposition assay (fig. 1F). For both assays, in addition to the piggyBac IRs, we generated additional reporter constructs where we used IR sequences flanking the piggyBac-derived genes of PGBD3 and PGBD4 that have been also amplified as Miniature Inverted-repeat Transposable Elements of MER75B or MER85, respectively, in the human genome (Sarkar et al. 2003) (Pace and Feschotte 2007; Pace et al. 2008) (Newman et al. 2008) (fig. 1G and H and supplementary fig. S5A, Supplementary Material online). However, no detectable catalytic activity of the PGBD1 was observed using any of the reporters, whereas the positive controls worked as expected.

To further determine whether PGBD1 might bind piggyBac transposon-related recognition sequences we sought to perform genome-wide ChIP-exo analysis. To identify relevant cell types for the analysis, we first checked expression profiles of PGBD1/2 proteins. Although the expression of PGBD1 has an enrichment in neural tissues, PGBD2 is expressed in a broader range of tissues (supplementary fig. S5B and C, Supplementary Material

---

that PGBD1 directly binds to the NEAT1 gene regulatory region and gene body. (Left panel) EMSA detects stable complexes (two upper bands) formed between HA-PGBD1 (HA-tagged, purified) and Motif1 and Motif2 oligonucleotides in the presence of nonspecific competitor (polydI-dC). The stability of the complexes is challenged by the equimolar presence of competitor oligonucleotides. Note that both of the upper bands represent specific complexes, and the upper band is likely an oligomeric complex. (Right panel) Purified HA-PGBD1 protein (5 µl) on 6% Stain Free BioRad SDS-PAGE. Experimental description of gel-shift and HA-PGBD1 purification is in Supplementary Material online. (D) High confidence protein interaction partners of PGBD1 [Log2FC(H/L ratio) L-HA-PGBD1 < −2,0 and Log2FC(H/L ratio) H-HA-PGBD1 > 2,0] identified by MS-SILAC (MDC-PGBD1 PPIs). PPI, protein–protein interactor. PPIs of similar function are marked by the same color. In clockwise: chaperones, oxidation/reduction status modifiers, member of the dystrophin-associated proteins (DAGs), SCAN-ZNFs, unclassified. (E) Electromobility shift assay (EMSA) confirms that PGBD1 directly binds to the SCAN-12 consensus DNA motif. (Left panel) EMSA detects stable complexes formed between HA-PGBD1 (HA-tagged and purified) and SCAN-12 oligonucleotides in the presence of nonspecific competitor (polydI-dC). (Left and middle panels) Note that the stability of the complexes, formed between HA-PGBD1 and SCAN-12 oligonucleotides, can be challenged only by the presence of competitor oligonucleotides (moderate reduction; middle panel, lanes 4 and 5). (Right panel) Purified HA-PGBD1 protein on 6% Stain Free BioRad SDS-PAGE (M: protein marker; 1, 5 µl; 2, 10 µl; 3, 20 µl purified HA-PGBD1). Experimental description of gel-shift and HA-PGBD1 purification is in Supplementary Material online.
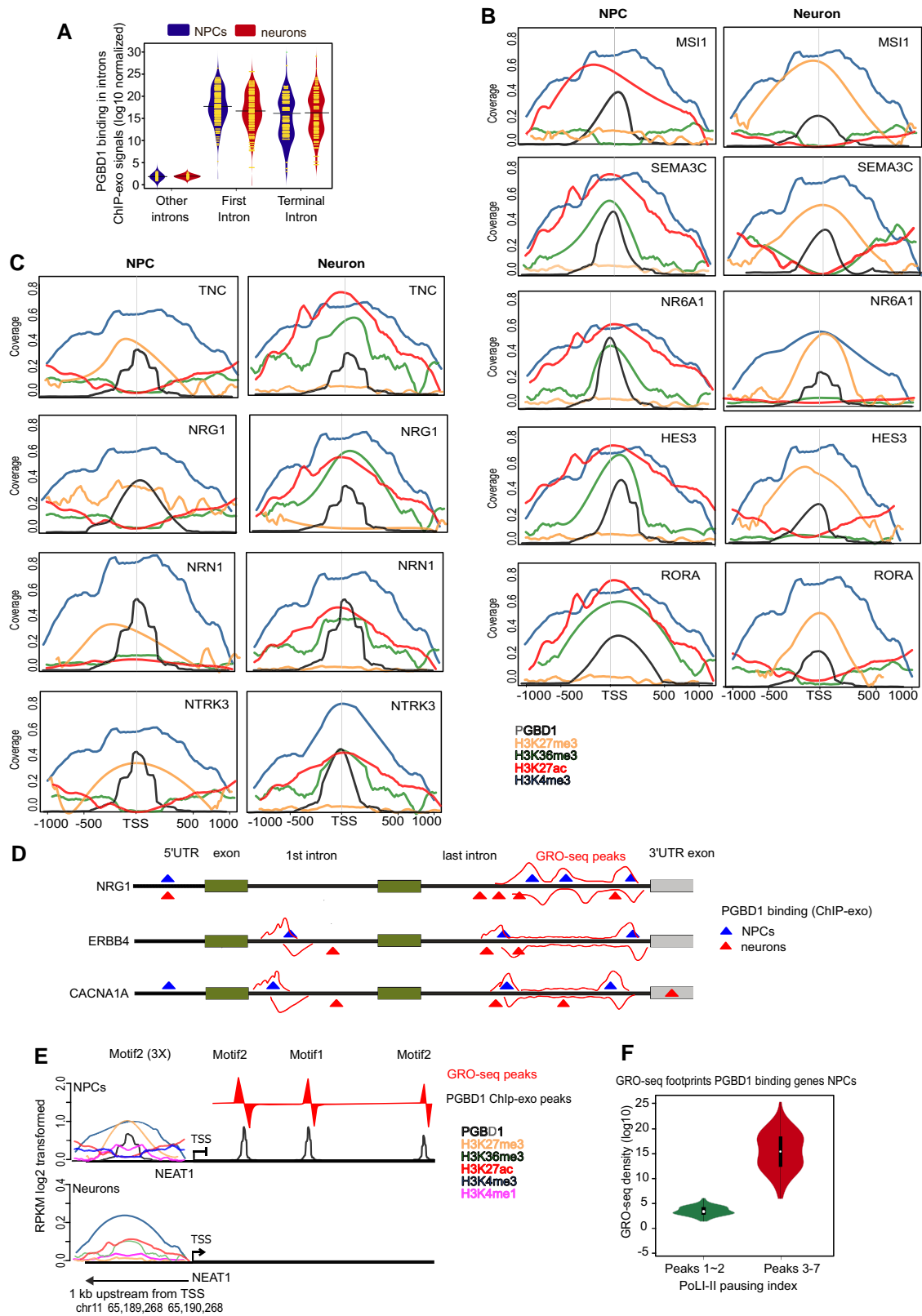
Fig. 3. PGBD1 target genes influence human neural progenitor identity. (*A*) Distribution of PGBD1 ChiP-exo binding signals in introns. Note that the ChIP-exo signals (log10 normalized, introns were length-normalized by plotting empirical densities of PGBD1 signals in bins of equal read count) were seen specifically in the first and terminal introns in both NPC and neurons. (*B* and *C*) Binding of PGBD1 to the promoter region of selected target genes influence neural progenitor identity. Analysis of regulatory genomics of the promoter region (1 kb upstream from TSS). Panels show the coverage (reads per million mapped reads) of PGBD1 (ChIP-exo) and epigenetic histone marks of a selected set of gene promoters in NPCs and in neurons. The selected genes are known markers of cell proliferation maintenance (*B*) or neuronal differentiation (*C*). (*D*) Representative examples of differential PGBD1 binding in NPCs and neurons. Differentially bound PGBD1 peaks between NPCs and neurons at the NRG1, ERBB4, and CACNA1A genes (not to scale). Only first/last exons and 5′ upstream/downstream genic regions are shown. Red wavy

7

online). The two highest expression levels of PGBD1 are detected in the cerebellar hemisphere (samples size = 215, median 18.59 TMP) and cerebellum (sample size $n = 241$, median 17.49 TPM) (GTEx). In single cell data analyses (proteinatlas.org), the top enhanced cell types are neuronal cells (excitatory/inhibitory neurons), followed by glial cells (oligodendrocytes, oligodendrocyte precursor cells, and astrocytes).

To further narrow down the neuronal activity profile, we examined PGBD1 expression levels in neuronal progenitor cells (NPCs) and in differentiated neurons of human and chimpanzee (GSE83638). PGBD1 has a higher expression in humans versus chimpanzee, and it is expressed at a higher level in progenitor versus differentiated cells (fig. 2A). The observed expression pattern of PGBD1 in differentiating human embryonic stem cells (hESCs), NPCs versus neurons is consistent with regulatory genomic analyses over the PGBD1 genomic locus (supplementary fig. S5D, Supplementary Material online). Notably, upon differentiating hESCs to neurons, PGBD1 is expressed highest (both RNA and protein) in hESCs, followed by NPCs and its differentiated derivatives (supplementary fig. S5E and F, Supplementary Material online).

Given the above enrichment patterns, we performed a ChIP-exo assay in both NPCs and neurons, differentiated from hESC_H1 (Reinhardt et al. 2013; Lorenz et al. 2017). The individual binding peaks show distinct NPC- and neuron-specific binding sites with an overlap (fig. 2B), suggesting that PGBD1 has both a shared and cell type-specific binding pattern in NPCs and neurons. If PGBD1 has lost its ability to recognize IRs, we expect an absence of IR-related sequences in motif candidates associated with ChIP-exo peaks. We thus determined PGBD1 ChIP-exo peaks based on read distribution against control. The IR sequences do not feature even as less significant hits or AUC (area under the curve) value(s). Conversely, the assay identified two top ChIP-exo motifs (Motif-1 and Motif-2) of PGBD1 indicating its sequence specific binding capacity (fig. 2B). We found no similarity between these PGBD1 ChIP-exo motifs and the *piggyBac* transposase recognition sequences, located in the IRs of the transposon or the IRs, flanking the PGBD3 and PGBD4 genes (also PB-like MER75B and MER85) IRs (fig. 2B). Both motifs were confirmed using electrophoretic mobility shift assay (EMSA) with fluorescently labeled oligonucleotides, containing either Motif-1 and Motif-2 and purified human haemagglutinin peptide (HA)-tagged PGBD1 protein (fig. 2C).

We conclude that there is no evidence that PGBD1 binds *piggyBac* transposase-related recognition sequences in the human genome, consistent with its loss of the CRD (Bouallegue et al. 2017), this domain being key to terminal IR recognition (Chen et al. 2020).

PGBD1 has Gained SCAN-12 DNA Binding Capacity

Via their multimerization domain, the proteins of the SCAN-domain family might interact with each other (Schumacher et al. 2000). Stable isotope labeling by amino acids in cell culture (SILAC)-based quantitative affinity purification mass spectrometry (q-AP-MS) of overexpressed HA-tagged PGBD1 and nontagged-PGBD1 overexpressing control cells quantified 1,103 proteins in both a forward and a reverse (label swap) experiment (supplementary table S7, Supplementary Material online). From these, we identify 19 high confidence interacting partners of PGDB1 (defined as false discovery rate, FDR < 0.05) that also have expression ≥10 TPM in NPCs (fig. 2D and supplementary fig. S6, Supplementary Material online).

As expected of a SCAN domain protein, PGBD1 binds at least one member of the SCAN family, ZNF24 (Zn-finger protein 24) (fig. 2D). The gene is implicated in maintaining the neural progenitor fate (Khalfallah et al. 2009) and is expressed at >70 TPM in NPCs. In addition to ZNF24 (alias ZCAN3), further protein interaction partners of the SCAN family (e.g., SCAND1, ZKSCAN1,3,4,8, 20; ZSCAN1,12, 18,20,22,25,32) are reported by BioGRID/STRING (supplementary table S1, Supplementary Material online) (Itokawa et al. 2009; Emerson and Thomas 2011; Huang et al. 2019).

Is this possible partnership with SCAN domain proteins reflected in DNA binding? Data mining of reported binding motifs of the SCAN-ZNF transcription factors predicts a 9 bp sequence, shared by 12 ZNF proteins (SCAN-12) and PGBD1 (fig. 2B and supplementary table S1, Supplementary Material online), suggesting that several family members can bind the same consensus, potentially modulating binding. To confirm that PGBD1 can bind the consensus, we used EMSA, with oligonucleotides, containing the consensus SCAN-12-motif. EMSA supported a stable DNA substrate–PGBD1 association (fig. 2E). To determine whether PGBD1 binds the consensus SCAN-12 motif alone or only in cooperation with other SCAN family members we added antiPGBD1 antibody to the reaction mixture. The observed supershifted complex suggests that PGBD1 is able to specifically bind the SCAN-12-motif alone (fig. 2E, middle right).

Forty percent of PGBD1 peaks correspond to this 9 bp sequence. This motif is also unrelated to the classical IR motif, consistent with no IR binding. This motif is also

lines represent the schematic GRO-seq peaks in NPCs. (E) PGBD1 binding suppresses NEAT1 expression in NPCs. PGBD1 ChIP-exo binding peaks (black) and regulatory genomic analysis of the NEAT1 (1 Kb upstream from the TSS) in NPCs and in neurons. Note that PGBD1 specifically binds NEAT1 in NPCs, but not in differentiated neurons. The NEAT1 promoter is in a repressed state in NPCs, supported by GRO-seq analysis (red), whereas activated in differentiated neurons. (F) GRO-Seq plot demonstrating global pausing indices of transcriptionally active RNA polymerase II (Pol-II) overlapping PGBD1 binding sites with peaks ($N = 1–2$) compared with ($N = 3–7$). Note that the pausing index in genes having multiple PGBD1 peaks was higher ~4.1, when compared with genes having less peaks ~1.2.

not related to the two top hit motifs from unbiased analysis. That it is not a top hit may reflect the fact that it is also degenerate (which statistically biases against enrichment scores).

PGBD1 Binds in and Around Genes

The genome-wide distribution of binding peaks suggests gain of new binding activity that could be regulatory. Consistent with the latter, in both NPCs and neurons, we observed binding sites mapping predominantly in or around (±1 kb) protein-coding genes (~84%) (supplementary fig. S7A, Supplementary Material online) ($Z = 4.2$), rather than noncoding regions (~16%). Generally, in genic regions we commonly observed multiple binding peaks on targeted genes that were mappable to upstream regulatory regions [1 kb from transcriptional start site (TSS)] and introns. Curiously, PGBD1 ChIP-exo signals were exclusively distributed in either the first and last introns (NPCs, 51 and 49%; neurons, 50% each), 5′ and 3′ UTRs, frequently associated with regulatory features (Bradnam and Korf 2008; Park et al. 2014). No peaks were mappable to introns located more centrally in the gene bodies (fig. 3A). Furthermore, intersecting PGBD1 Chip-exo peaks, genomic regions >2 kb upstream TSS of RefSeq annotated genes and H3K4me1 histone mark signals revealed ~2000 PGBD1 binding peaks (68% NPCs, 72% neurons) overlapping with regulatory regions genome-wide (supplementary fig. S7B, Supplementary Material online).

PGBD1 Targeted Genes are Associated with Neuronal Development

In both NPCs and neurons, among the most significant gene ontology (GO) terms of targeted protein-coding genes we observed *neuron development, neuron differentiation and neurogenesis* (supplementary fig. S7C and table S2, Supplementary Material online). Restricting analysis to those cases where the PGBD1 binding is 5′ of the gene body (hence more likely to be regulatory and less affected by gene body size artifacts), reveals nervous system development (NPC) and neuron development to be among the most enriched categories (supplementary table S3, Supplementary Material online). Since the ChIP-exo peaks shared the binding motifs, but had specificities in NPCs and neurons (fig. 2B), we hypothesized that the sites would have differential accessibility. To test this hypothesis, we first analyzed the common target genes in both cell types around their TSSs (1 kb upstream) for various histone marks of active and repressive activities in promoter regions (e.g., H3K4me3, H3K36me3, H3K27me3, H3K27ac) and chromatin accessibility (ATAC-seq) (supplementary fig. S7D–F, Supplementary Material online). This integrative analysis revealed a set of common target genes, where the PGBD1 peaks overlapped with either activating or repressive histone marks (supplementary fig. S7B, D, and E, Supplementary Material online).

To determine what the repressed and activated biological processes are, we selected a set of common target genes from the most significant GO categories e.g., *nervous system development* (NPCs: $n = 887$) and *neuronal differentiation* (NPCs: $n = 5450$). This revealed that the PGBD1-targeted gene set which overlaps with activating histone marks (H3K4me3, H3K27ac, H3K36me3) includes several genes associated with the maintenance of the progenitor state (e.g., MSI1, SEMA3C, NR6A1, HES3, RORA) (fig. 3B), whereas the repressed genes (histone mark, H3K27me3) are generally involved in activating neuron differentiation (e.g., TNC, NRG1, NRN1, NTRK3) (fig. 3C and supplementary table S2, Supplementary Material online). This is consistent with PGBD1 being involved in keeping the regulatory regions of genes responsible for neural progenitor maintenance in an active state, whereas simultaneously repressing those that could initiate the differentiation process. We further test this via knockdown (KD) analyses reported below.

In the common gene set of PGBD1 targets in NPCs and neurons, we observe both shared and differential binding peaks (supplementary table S2, Supplementary Material online), supporting the observation that PGBD1 binding might exert differential, cell type specific modulation of gene activity. In addition to the genes of distinct accessibility at their upstream regulatory regions (fig. 3B and C), among the targets with a differential intronic binding pattern, we identify NRG1 (neuregulin 1) and its associated cell surface receptor, ERRB4 (Erb-B2 Tyrosin kinase 4) or calcium voltage-gated channel subunit alpha1 A (CACNA1A), all required for normal development of the embryonic central nervous system (Brinkmann et al. 2008; Gauthier et al. 2013; Mei and Nave 2014; Humbertclaude et al. 2020) (fig. 3D).

Although the identified NPC-specific targets did not show up in any GO category significantly, we observed several lncRNAs among the repressed nonprotein-coding genes, including NEAT1, MIR100HG (fig. 3E and supplementary table S2, Supplementary Material online).

PGBD1 Modulates Transcriptional Pausing

The ChIP-exo analysis indicates that PGBD1 typically binds at multiple positions at its targets. Aside the 5′ regulatory region, PGBD1 binds the gene bodies at multiple positions. Could PGBD1 binding over the gene body affect transcriptional rate? Might PGBD1 intragene body binding more generally modulate transcriptional pausing?

To test this hypothesis, we reanalyzed a genome-wide GRO-seq datasets in NPCs (Wang et al. 2020) that measures nascent RNA, with bidirectional internal signals indicative of transcriptional pausing. Protein-coding target genes bound by PGBD1 were divided into two groups based on the number of the observed ChIP-exo peaks (1–2 vs. 3–7). In order to quantify transcriptional pausing for regions of interest, we computed the quotient of binned expression per analyzed feature (PGBD1 intersecting GRO-seq peaks) and the entire gene body. Our findings
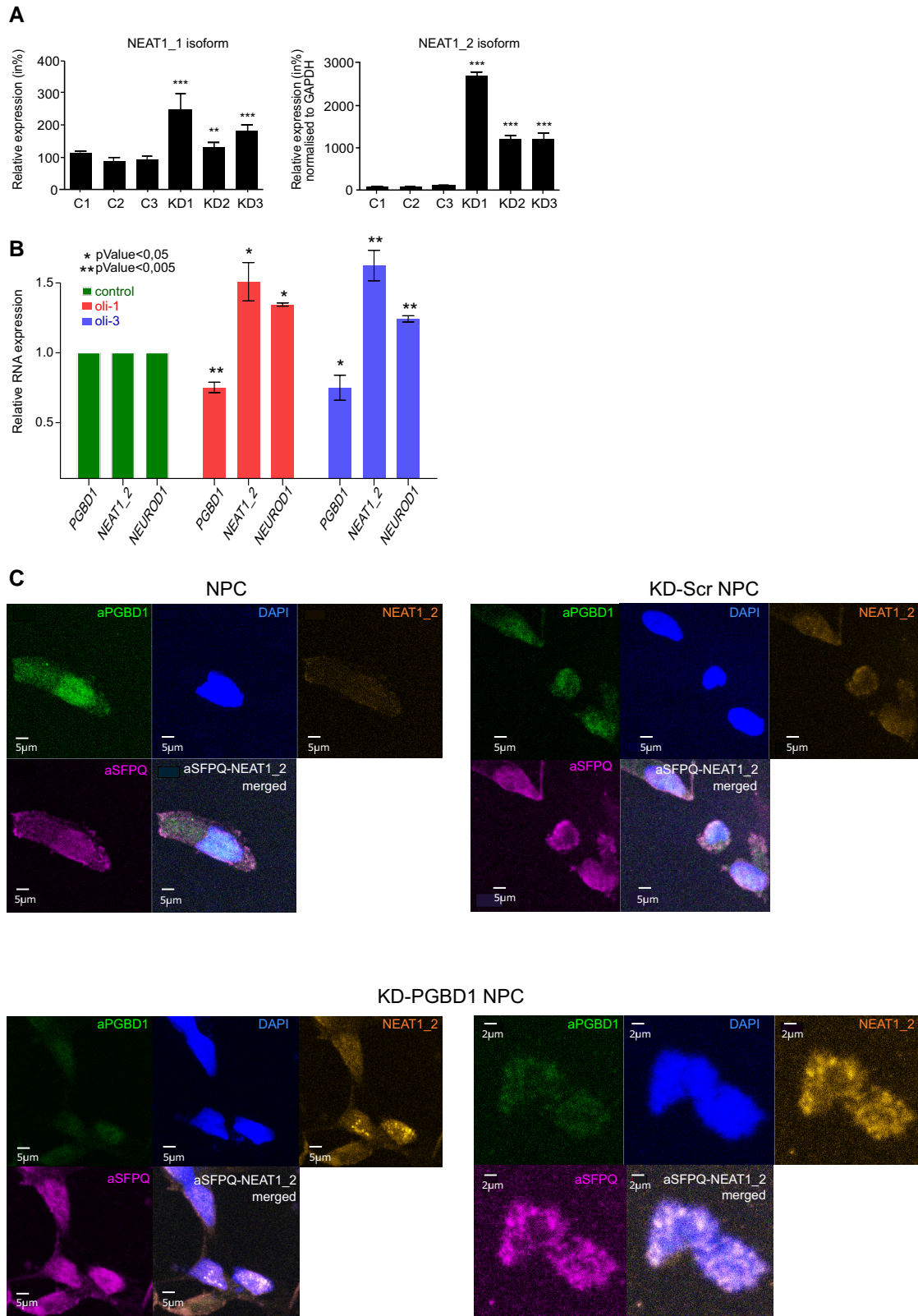
Fig. 4. PGBD1 controls mammal-specific neuronal paraspeckles. (*A*) qPCR confirms the knock-down effects of PGBD1 on NEAT1 transcription in NPCs (miRNA/SB100X RNAi approach). See also Supplementary Material online. Data shown are representative of three independent experiments with biological triplicates per experiment (normalized to GAPDH) *P*-values: *$P < 0.05$, **$P < 0.005$, ***$P < 0.0005$. Error bars indicate *s.d.* Note that while both isoforms are up-regulated upon KD, the values are a magnitude higher for the NEAT1_2 isoform. The same samples were subjected to RNA-sequencing (fig. 5). (*B*) Quantification of the transcription level of PGBD1, NEAT1_2, and NEUROD1 by qPCR in PGBD1 depleted NPCs using the dCAS9-CRISPR-KRAB-MeCP2 method (proof of concept, see also Supplementary Material online). The data are normalized to GAPDH. Graph shows the results of three independent experiments in triplicates. *P*-values: *$P < 0.05$, **$P < 0.005$. (*C*) Depletion of PGBD1 induces paraspeckle formation of in NPCs. Representative fluorescent microscopy images using antibodies against PGBD1, SFPQ paraspeckle

(density of GRO-seq reads) (fig. 3F) suggest that regions with multiple PGBD1 binding sites, including the NEAT1 locus, are prone to transcriptional pausing (fig. 3D and E).

PGBD1 Suppresses Paraspeckle Formation by Repressing NEAT1_2

Of the PGBD1 binding sites NEAT1 especially attracted our attention for several reasons. First, NEAT1 is associated with SCZ (Katsel et al. 2019) and may there share a role with PGBD1. Second, NEAT1 is the core structural RNA of mammal-specific paraspeckles (Clemson et al. 2009). The observation of a coupling between PGBD1 and NEAT1 suggested the possibility of a new gene coincident with a new structure. Third, ZNF24, a PGBD1 interactant (fig. 2D), is also associated with NEAT1 levels and paraspeckles (Fong et al. 2013; Yamazaki and Hirose 2015). Finally, as NEAT1, a hallmark of differentiated cells, is not observed in embryonic stem cells or in neural progenitors (Chen and Carmichael 2009; Mercer et al. 2010), PGBD1 binding at the NEAT1 locus (fig. 3E) might be associated with its suppression and correlated with prevention of NPC differentiation. We thus focus to characterize in more detail the PGBD1/NEAT1 interaction.

First, we sought to verify that PGBD1 does indeed bind NEAT1. Analysis of the PGBD1-ChIP-exo identified both Motif1 and Motif2, validated by EMSA (fig. 2B and C), multiple times over the NEAT1 locus (fig. 3E). We conclude that PGBD1 binds both to the promoter and gene body of NEAT1 (fig. 3E). In contrast, no significant binding signal was detectable over the NEAT1 locus in differentiated neurons (fig. 3E) suggesting that PGBD1 modulates NEAT1 transcription specifically in NPCs.

Given the specific DNA binding and the transcriptional pausing of PGBD1 at the NEAT1 locus (fig. 3E), we hypothesize a role of PGBD1 in regulating the transcription of NEAT1. To validate this, we determined NEAT1 transcription in NPCs, depleted for PGDB1 expression. As the knockout (KO) strategy interfered with cell renewal (supplementary fig. S8A, Supplementary Material online), preventing stable maintenance of a colony, we used a miRNA KD approach to deplete PGBD1 (supplementary fig. S8B, Supplementary Material online), this being compared with NPCs treated with scrambled miRNA (KD1–3 vs. C1–3). This method combines the RNAi and Sleeping Beauty mediated transposition (Mates et al. 2009; Bunse et al. 2014), and is suitable to generate stable KD clones in NPCs. Although by knocking down PGBD1 in NPCs elevates the abundance of the transcript levels of both isoforms (NEAT1_1/_2), the reduced PGBD1 primarily

affects the level of NEAT1_2 (∼2-fold and ∼17-fold elevation, respectively), validated by quantitative polymerase chain reaction (qPCR) (fig. 4A). Notably, the longer isoform NEAT1_2 is the structural component of the paraspeckles (Clemson et al. 2009) and has an essential role in their formation, the tRNA-like triple helix at its 3'-end being required to stabilize paraspeckles. Similar results are observed in a neuroblastoma cell line (SHEP cells) (supplementary fig. S8C–E, Supplementary Material online). To specifically inactivate the promoter of PGBD1, and further exclude the possibility of off-target effects, we also apply the CRISPR-KRAB-MeCP2 repressor approach (supplementary fig. S9 and table S4, Supplementary Material online) (Yeo et al. 2018). This similarly reveals up-regulation of NEAT1_2 on PGBD1 down-regulation (fig. 4B and supplementary fig. S9A–C, Supplementary Material online). Together these data show that PGBD1 plays a key role in repressing NEAT1_2 and thus paraspeckle formation.

In order to further scrutinize this, we used confocal microscopy in WT and PGBD1-depleted NPCs. To visualize NEAT1_2, we performed a FISH assay by using a specific probe for the NEAT1_2 transcript (the one associated with paraspeckles). We combined the FISH with immunohistological staining against PGBD1 and SFPQ as a paraspeckle marker. Our expectation under a model in which PGBD1 suppresses NEAT1_2 is that in the presence of PGBD1 (which should be intranuclear and diffuse), we should not observe SFPQ/NEAT1_2 colocalized foci, the SFPQ should be intranuclear and diffuse and NEAT1_2 largely absent. On PGBD1 depletion, SFPQ and NEAT1_2 should now colocalize in intranuclear foci (paraspeckles).

In agreement with expectations, in WT-NPCs and in control cells treated with a scrambled RNAi KD, no obvious signs of either elevated level of NEAT1_2 transcripts or SFPQ-marked paraspeckles were seen (fig. 4C). Both PGBD1 and SFPQ are predominantly nuclear and diffusely distributed as predicted. These observations are consistent with previous reports that paraspeckles are not detectable in NPCs (Mercer et al. 2010), and appear upon differentiation (Bond and Fox 2009; Modic et al. 2019). In PGBD1-depleted NPCs, by sharp contrast, we detect a robust nuclear NEAT1_2 signal, indicating intensive NEAT1_2 transcription. The NEAT1_2 signal accumulated in multiple SFPQ-marked nuclear bodies, colocalized with SFPQ immunostaining, which is no longer diffuse (fig. 4C). These observations indicate that a decreased level of PGBD1 is associated with extensive paraspeckle formation (fig. 4C), consistent with a key suppressor role of PGBD1 in the biogenesis of paraspeckles in NPCs (Mercer et al. 2010).

proteins, combined with FISH visualization of NEAT1_2 RNA and DAPI staining. (Upper panels) No significant level of NEAT1_2 RNA-FISH signal is detectable in (left panel) untreated NPCs or (right panel) in scrambled-miRNA transfected KD-Scr NPC controls. Note the colocalization of the aPGBD1 and aSFPQ signals mostly in the nuclei (merged image). (Lower panels) Representative images of paraspeckle formation in KD-PGBD1 NPCs (at two different magnifications). In sharp contrast to controls, upon PGBD1 depletion, a robust NEAT1_2 RNA-FISH signal appears that accumulates in speckles (yellow-brown). The NEAT1_2 RNA-FISH signal colocalization with the aSFPQ immunostaining (merged image) defines the nuclear structures paraspeckles.
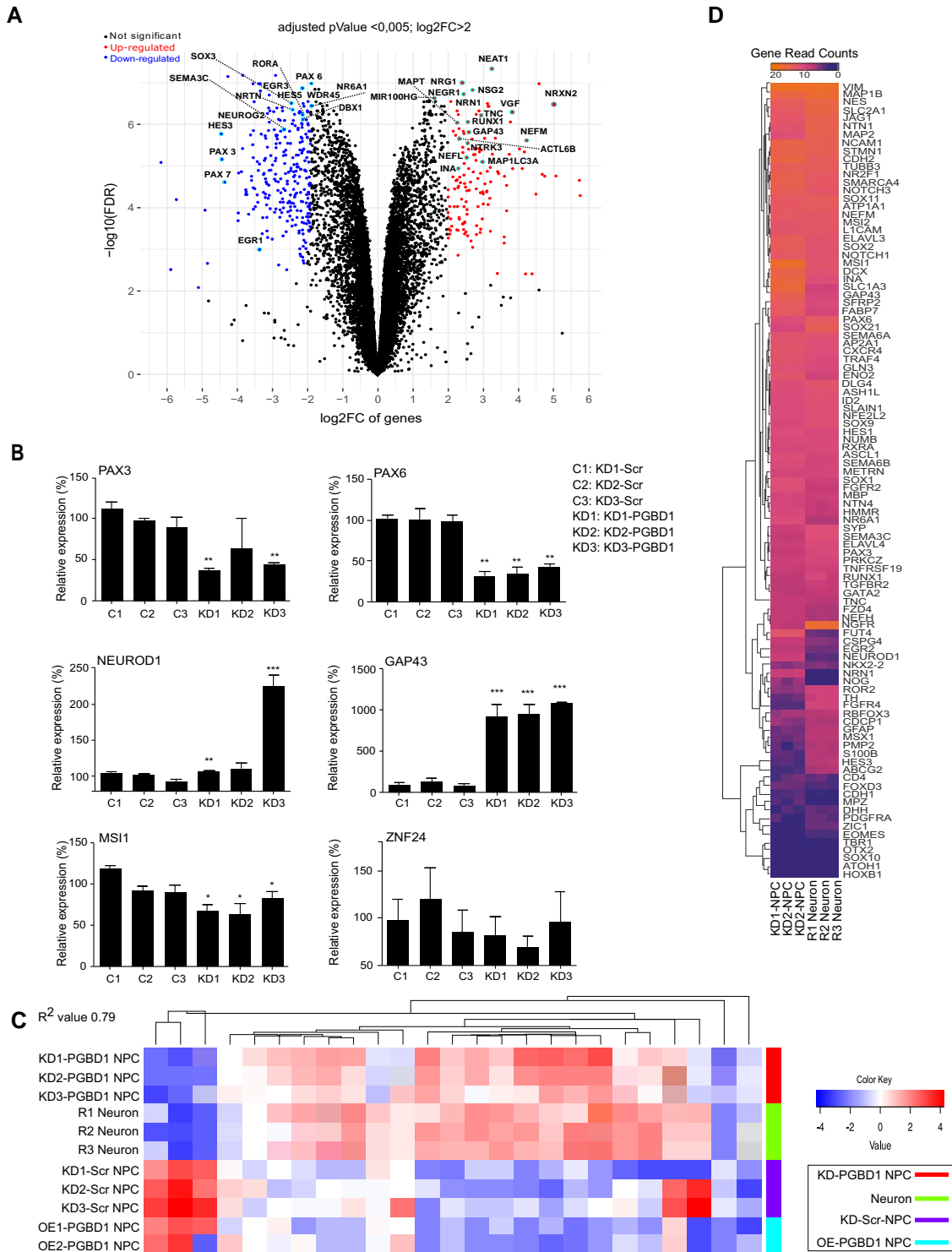
Fig. 5. Depletion of PGBD1 compromises the progenitor state of NPCs. (A) Volcano plot shows the 762 DEGs in the transcriptome of KD PGBD1 neural progenitors, KD-PGBD1_NPC (log2 fold >2 change (log2FC), P-value <0.05. Depletion of PGBD1 resulted in the up-regulation of 475 genes, whereas 287 genes were down-regulated. The highlighted up-regulated genes are mostly associated with neural differentiation, whereas several of the down-regulated genes have relevant functions in maintaining the self-renewal of NPCs (for the list of DEGs, see Table S5). (B) qPCR confirms the effects of PGBD1 KD on transcription of selected neuronal marker genes in NPCs. Data shown are representative of three independent experiments with biological triplicates per experiment. The relative expressions are normalized to GAPDH. P-values: *P < 0.05, **P < 0.005, ***P < 0.0005. Error bars indicate s.d. (C) The transcriptomes of the PGBD1-depleted NPCs (KD-PGBD1) and differentiated neurons are highly similar ($R^2 = 0.79$). Global comparative analysis of the transcriptomes of PGBD1-depleted NPCs, treated miRNA scramble control NPCs and differentiated neurons (Reinhardt et al. 2013). Overexpression (OE) of PGBD1 in NPCs does not generate robust transcriptome changes. The transcriptome of the OE-PGBD1 NPCs is highly similar to the scrambled control NPCs, smNPC-miRNA620 ($R^2 = 0.8$). (D) Depletion of PGBD1 compromises the NPC identity and results in neuronal differentiation. HeatMap demonstrates the comparison of a gene expression of a selected neuronal lineage marker set ($n = 99$) (https://www.rndsystems.com/research-area/neural-stem-cell-and-differentiation-markers) in PGBD1-depleted neural progenitor cells (KD1–3-NPCs) and differentiated neurons in three replicates (R1–3).

A

B
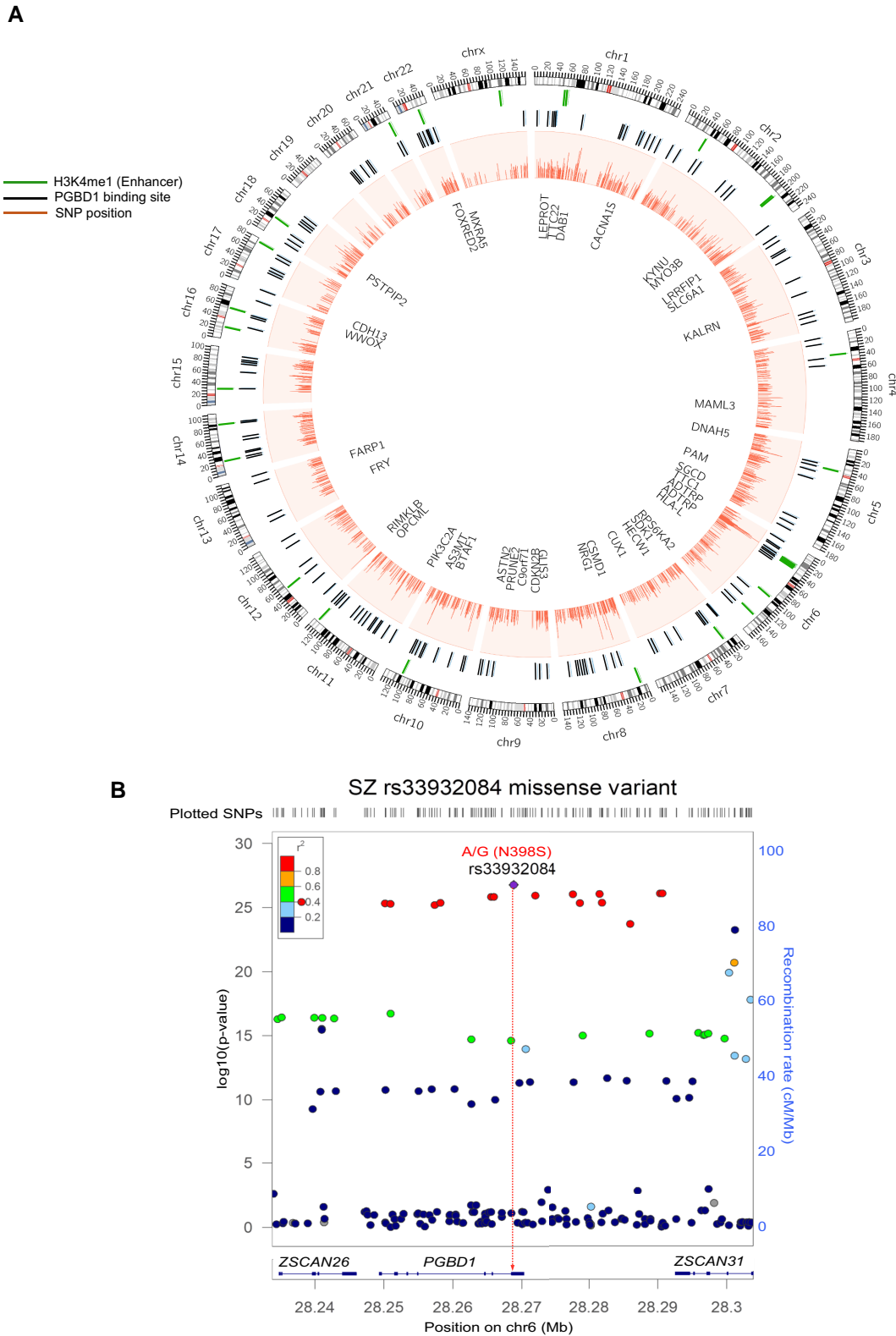
Fig. 6. The KD-PGBD1_NPC model mimics certain aspects of schizophrenia. (A) PGBD1 binding is enriched at the enhancer regions of a subset of schizophrenia (SCZ) susceptibility genes. SCZ-associated SNPs ($N = 1251$), PGBD1 ChiP-exo peaks ($N \sim 2000$), mapping >2 kb upstream of TSS of genes were overlaid with the H3K4me1 histone marks (indicative of active enhancers). CIRCOS shows the schizophrenia susceptibility genes indicated with PGBD1 binding in their enhancer region. (B) SCZ associated SNPs mapping on the chr6 in and around the PGBD1 locus. SCZ patients ($N = 443,581$) and a replication cohort (1169 controls; 1067 cases) (Bansal et al. 2018). Y-axes show the strength of association ($\log_{10}$- -P-value). Color code: LD r^2. Rs33932084 was identified in last exon of PGBD1, resulting in a missense mutation (N398S).

13

We also repeated the visualization experiment in a stable, PGBD1-depleted neuroblastoma cell line (SHEP) with a similar result (supplementary fig. S9D, Supplementary Material online). These data indicate that the release of the NEAT1_2 transcription from PGBD1-mediated suppression is a requirement for paraspeckle formation in neuronal cells.

PGBD1 Regulates Neuronal Cell Differentiation

The above KD experiment revealed a further peculiarity, namely that the NPCs tended to differentiate on PGBD1 KD. This may be mediated through the NEAT1 axis but there is no reason to suppose that this is the unique axis. To decipher, and identify target gene activation/repression, we performed transcriptome analysis on PGBD1-depleted NPCs. Indeed, a comparison of the NPC and KD-PGBD1_NPC transcriptomes (Log2-fold change, L2FC) identified 762 differentially expressed genes (DEGs) (fig. 5A; supplementary table S5, Supplementary Material online), some of them also validated by qPCR (fig. 5B). The overall GO analysis of the DEGs in KD-PGBD1_NPC suggests that the function of PGBD1 is associated with *gene regulation of cell differentiation, nervous system development*, and *neurogenesis* (supplementary table S4, Supplementary Material online). Importantly, from the 762 significant DEGs, we found 212 genes (over 1/3), with significant PGBD1 ChIP-exo signals consistent with differential expression being owing to PGBD1 DNA binding (supplementary table S6, Supplementary Material online). In this overlapping dataset, the peaks for PGBD1 binding are located in the upstream regulatory (1 kb upstream from TSS) regions of 38 DEGs, again consistent with PGBD1 playing a transcriptional regulatory role of these targets (supplementary table S6, Supplementary Material online). Most of the 212 genes have intronic binding and we identified 15 genes with a 3′ UTR PGBD1 interaction. For several of the neuron marker genes (HES3, NRG1, RORA, SEMA3C, NTRK3), in addition to promoter binding of the genes, intronic PGBD1 binding sites were identified.

Analysis of the 38 genes revealed the down-regulation of SEMA3C, NR6A1, HES3, RORA, NRTN, and DBX1 genes in KD-PGBD1, all implicated in maintaining the proliferative status of NPCs (Hu et al. 2010; Zhang and Jiao 2015; Vinci et al. 2016). Among the up-regulated genes, by contrast, are genes associated with neuronal differentiation (e.g., TNC, NRG1, NRN1, ACTL6B, and NTRK3) (fig. 5A). Aside from the 38 protein-coding genes, and in addition to lncNEAT1_2, we identify another essential lncRNA MIR100HG (fig. 5A and supplementary table S6, Supplementary Material online), also implicated in neuronal differentiation regulation (e.g., encoding for the miRNA cluster, including LET7a-2) (Bevilacqua et al. 2015; Cui et al. 2019).

Collectively, our transcriptome analysis, in conjunction with the integrative chromatin status determination, suggests that the depletion of PGBD1 compromised the identity of NPCs, and triggered the cells to activate their differentiation program. To test this, we compared the transcriptome of control (scrambled miRNA) NPCs, KD-PGBD1_NPCs, and in vitro differentiated neurons from hESCs (4 weeks) (Reinhardt et al. 2013). The comparison revealed a strong correlation ($R^2 = 0.79$) between the transcriptomes of PGBD1-depleted NPCs and differentiated neurons, and anticorrelation between NPCs to both, supporting a high similarity between the transcriptomes of KD-PGBD1 and differentiated neurons (fig. 5C). In addition, analysing the transcriptional changes of 99 key neuronal markers supports the hypothesis that depleting PGBD1 drives cells toward a differentiated phenotype (fig. 5D). Thus, PGBD1 depletion activates NPC differentiation, arguing for an essential role of PGBD1 in the maintenance of the progenitor state of neuronal cells. In contrast to the robust transcriptional changes generated by PGBD1 depletion, overexpression (OE) of PGBD1 results in no dramatic changes in NPCs (fig. 5C), and the OE-PGBD1 transcriptome stays close to the scramble control ($R^2 = 0.8$). The later result also further implies that experimental manipulation of NPCs is not itself a trigger to differentiation (as supported by both the scrambled RNAi and CRISPR-KRAB-MeCP2 control studies (fig. 4A and B).

As PGBD1 coordinates with other SCAN domain proteins, we additionally sought to ask whether there might also be a transcriptional coupling. The null expectation is that there would not and we had no mechanistic reason to expect that there would be. To address this, we also determined the transcriptional status of other SCAN domain family members in PGBD1-depleted cells. Among the significant DEGs, we found only the brain/cerebellum-specific SCAN-KRAB domain ZNF483 with moderate expression alteration (L2FC-1.79), whereas other family members were not affected transcriptionally (not shown). This observation is in line with the assumption that the family members are more likely to modulate each other's activity via protein–protein interaction (via the SCAN domain).

Discussion

These results suggest that PGBD1 is a mammal-specific gene that in NPCs suppresses production of a mammal-specific structure, paraspeckles, via suppression of its core long noncoding RNA lncNEAT1 (Clemson et al. 2009). PGBD1 thus is not just the first example of novel gene integrating into human neuronal functioning, it is also a case of a new gene regulating a new structure. This appears to be part of a broader function of PGBD1 as a regulator maintaining NPC status and blocking differentiation. It adds to the list of domesticated TEs, for example, HERVH (Wang et al. 2014) that likely have a role in self-renewal regulation. The fact that depletion of this gene compromises self-renewal of the neural progenitors in human suggests that this is an unusual case in which a new gene has evolved a cell-level core function. Such

circumstances are intrinsically paradoxical, as we must query how organisms survived before the evolution of the new core gene. In this instance, the resolution appears to be, at least in part, that the new gene is associated with control of a new core process, paraspeckle formation.

Note that we do not wish to claim that PGBD1 is the sole regulator of paraspeckles. PGBD1-mediated regulation of paraspeckles might be specific to NPC/neural cells, where PGBD1 is dominantly expressed, whereas paraspeckle regulation by other means [e.g., CARM1 (coactivator-associated arginine methyl-transferase1) (Torres et al. 2017)] might be more typical in other cell types. On the other hand, PGBD1 is specific for regulating (lnc)NEAT1/paraspeckle biogenesis, other nuclear bodies, namely nuclear speckle and CS body assembled around structural lncRNAs of MALAT1(NEAT2) (Tripathi et al. 2010) and GOMAFU (Ishizuka et al. 2014), respectively, are not affected upon PGBD1 KD.

NEAT1 and paraspeckle regulation mediated by PGBD1 repression adds to the limited inventory of new genes associated with new structures, be these macroscopic or microscopic (Dupressoir et al. 2012; Santos et al. 2017). The closest related examples are the TE-derived syncytins involved in the formation of syncytiotrophoblast in some mammals (Dupressoir et al. 2012). In contrast to PGBD1, however, the evolution of syncytins was mediated by co-option of extant capacity. PGBD1's current biological activity has little resemblance to the ancestral activity, not least owing to the loss of transposase catalytic abilities and gain of function (e.g., by recruitment of SCAN/KRAB). The loss of catalytic ability is in line with the assumption that the DDD catalytic domain of the *piggyBac* transposase is not conserved among the domesticated *piggyBac*-derived PGBD sequences (Sarkar et al. 2003; Newman et al. 2008; Pavelitz et al. 2013) and hence probably not key to their domestication (Bouallegue et al. 2017). This may contrast with PGBD5 that has been suggested to have a residual DNA transposase-like activity, capable of mobilizing a synthetic DNA transposon in human cells (Henssen et al. 2015; Ivics 2016; Henssen, Koche et al. 2017), although this has been recently challenged (Beckermann et al. 2021). Given that a common mode of TE recruitment is one in which prior activity is coopted, the loss of potential nuclease activity and the gain of new binding activity is at first sight surprising. However, recent analysis has suggested that chimeric TE—KRAB-SCAN genes are a common mode of TE domestication (Cosby et al. 2021). PGBD1 presents a paradigmatic example of such a process.

Although the effects of PGBD1 on suppression of NEAT1 seem relatively clear, beyond transcriptional pausing, the precise mechanism by which this happens we have not considered. Curiously, PGBD1 typically binds several positions in a target gene, and possibly affects transcription in multiple ways. For example, it might act as a physical barrier to progression of RNA PolII. This could make sense of the intronic deposition of PGBD1. Alternatively, it might physically obstruct the promoter preventing

other transcription factors from binding or enable recruitment of suppressor complexes (N.B. we have no evidence that its TRIM28 would be one of these). As to the block to differentiation, our results suggest many possible routes via the 38 DEGs, some or all of which might be causative. This downstream analysis we leave to future study. The nature of the interaction with ZNF24 is also worthy of further study. Although KDs of both alter NEAT1 levels, the effects are opposite: ZNF24 KD reduces NEAT1 and paraspeckles, whereas PGBD1 KD increases both. ZNF24 likewise is involved in regulating neural progenitor fate (Khalfallah et al. 2009). Given the physical interaction between the two this suggests the possibility of the control of PGBD1 activity by sequestration/titration by ZNF24.

Is PGBD1 Also a Neuronal Stress Response Gene?

Above we have focused on PGBD1's activity in binding DNA, potentially in collaboration with other SCAN domain family members (i.e., through heterodimers). Notably, despite of the KRAB domain, neither TRIM28, a KRAB interactor (Tycko et al. 2020), nor any of the members of the TRIM28-associated gene regulation complex, are within the set of significant interactors. The protein interactome of PGBD1, however, is by no means restricted to SCAN domain proteins and suggests three particular clusters, which in turn suggest multiple modes by which PGBD1 might regulate neurogenesis as well as other forms of activity, most notably stress response.

First, the dystrophin associated protein (DAP) complex comprises at least ten proteins (Gao and McNally 2015), from which, we identify in the PGBD1 interactome Dystrophin (DMD), Uthrophin (UTRN), Syntrophin beta 1 and 2 (SNTB1, SNTB2), Catenin alpha like 1 (CTNNAL1), and Dystrobrevin (DTNA, DTNB). Although, dystrophin is primarily expressed in the skeletal muscle, DTNB is a member of the brain-specific dystrophin complex (Blake et al. 1998; Loh et al. 1998), regulating distinct aspects of neurogenesis (reviewed in Waite et al. 2012). We hypothesize that the presence of multiple members in the interactome suggests that PGBD1 could have a significant modulatory effect on the processes, controlled by the brain-specific DAP complex.

In addition, two further gene set clusters are evident in the PGBD1 interactome (fig. 2D). Using *GO*-based classification, DMD, SNTB1, SNTB2, and UTRN belong to the *GO glycoprotein complex*. The second cluster has *GO chaperon cofactor dependent protein refolding*, r*eactome regulation of Hsf1 mediated heat shock response*, and *cellular response to heat stress* as the most significant categories (supplementary fig. S6C, Supplementary Material online). Indeed, there are significant interactions with multiple stress-responsive chaperones, involved in protein quality control (e.g., HSPA4, HSPA4L, HSPA5, 6/7, 8, 9, HSPH1, HSPA1A) (Kampinga et al. 2009; Kampinga and Bergink 2016) (fig. 2D and supplementary fig. S6A and C, Supplementary Material online), indicating that at the protein level, PGBD1 may also be associated with the stress

response system. This role in stress response appears to also be supported by PGBD1's activity via DNA binding as well. There are 191 stress response target genes from the GO categories of *response to stress* in neurons, but not NPCs. PGBD1 binds 23% (43/191) of these genes at their upstream/regulatory region (supplementary table S3, Supplementary Material online), supporting a distinct role of PGBD1 in NPCs and neurons. A role in stress response and neural homeostasis we suggest to be worthy of follow-up.

Is PGBD1 a SCZ Gene?

PGBD1 is a highly transcribed gene in the cerebellum (supplementary fig. S5B, Supplementary Material online), dysfunction of which is connected with neuronal disorders (reviewed in Picard et al. 2008; Yeganeh-Doost et al. 2011; Chen et al. 2013; Parker et al. 2014; Sathyanesan et al. 2019). Nonetheless it is not routinely identified as a gene associated with neurological diseases. This is surprising as NEAT1 is a negative regulator of neuronal excitability and axonal maintenance (An et al. 2018), the hallmarks of neurodegenerative disorders. Dysregulated NEAT1 expression is reported from amyotrophic lateral sclerosis, Huntington disease, Alzheimer's disease, Parkinson's disease (reviewed in Lein et al. 2007; An et al. 2018 and SCZ, Katsel et al. 2019).

Although NEAT1_2 is the paraspeckle structural RNA, both NEAT1_1 and _2 RNA isoforms bind DNA (West et al. 2014). Might there be an intersection between lncNEAT and PGBD1 binding and is there any evidence that the genes regulated by both are especially enriched as SCZ candidate genes? Both PGBD1 and lncNEAT1 are expressed in differentiated neurons, and both bind the target genes around transcriptional start and termination sites (TSS and TTS) (fig. 3A) (West et al. 2014). We compared their binding targets using our PGBD1 ChiP-exo and published human NEAT1-ChIRP-seq data from neurons (Katsel et al. 2019). Of the 75 commonly targeted genes several are involved in *neurogenesis* ($N = 21$) and *neuronal development* ($N = 19$) (supplementary table S8, Supplementary Material online). This suggests that in differentiated neurons PGBD1 and (lnc)NEAT1/paraspeckles may regulate similar biological processes (Mercer et al. 2010; Lellahi et al. 2018; Modic et al. 2019). Notably, of these 75 common targets 68 are SCZ susceptibility genes (Jaccard index = 0.67, $P < 0.01$). Seven are known as markers for oligodendrocytes (7/75) (Lein et al. 2007), involvement in the development of which is a hallmark of SCZ (Hoffmann et al. 2019; Schmitt et al. 2019; Gouvea-Junqueira et al. 2020). However, the direction of effects is contradictory: NEAT1 appears to be downregulated in SCZ (Li et al. 2018; Katsel et al. 2019), yet PGBD1 is also. As cells increase paraspeckle abundance as part of a general stress response (McCluggage and Fox 2021), this overlap may additionally reflect a correlated response to stress, mediated both via DNA interactions and protein–protein interactions (An et al. 2018).

As evidenced above, an association with SCZ seems to be especially noteworthy. Among the reported SCZ susceptibility genes, we find several PGBD1 targets, including ERBB4, NRG1, ATXN3, SNAP91, SRPK2, or CACNA1A (fig. 3B and D and supplementary table S2, Supplementary Material online) (Zhuchenko et al. 1997; Stefansson et al. 2002; Silberberg et al. 2006; Brinkmann et al. 2008; Buonanno et al. 2008; Gauthier et al. 2013; Mei and Nave 2014; Schizophrenia Working Group of the Psychiatric Genomics 2014; Takata et al. 2017; Humbertclaude et al. 2020). More generally, given that SNPs in enhancer regions might contribute to dysfunctional gene regulation of their targets (Huo et al. 2019; van Arensbergen et al. 2019), one can ask whether an association exists between the enhancer-promoter gene regulatory network of PGBD1-bound genomic sites and SCZ susceptibility. We analyzed the promoter interactions of PGBD1-bound, H3K4me1-enriched putative enhancer regions (~2000) (fig. 6A) in the HACER database (Human Active Enhancers to interpret Regulatory variants), and intersected them with SNPs, identified by SCZ GWAS (Schizophrenia Working Group of the Psychiatric Genomics 2014). This revealed that around 53% of the target genes interacting with PGBD1-bound enhancer regions (1982 and 2101 in NPCs and neurons, respectively) are also associated with SCZ susceptibility, defined by GWAS risk SNPs studies (Wang et al. 2019) (fig. 6A). Possibly of significance, several of them are involved in oligodendrocyte development.

There is then a possible mechanistic coupling between PGBD1 and SCZ susceptibility genes. How can we square this with inconclusive GWAS data? PGBD1-associated SNPs (rs2142731, rs1150772, rs3800324, rs13211507) have been reported by independent GWASs in SCZ patient cohorts (Yue et al. 2011; Zhang et al. 2013; Schizophrenia Working Group of the Psychiatric Genomics 2014; Schrode et al. 2019). The rs2142731 SNP-SCZ association was not, however, confirmed in all ethnic groups (Stefansson et al. 2009; Kitazawa et al. 2012; Ma et al. 2013). One reason the GWAS evidence base is weak, however, is that in some surveys PGBD1 has been ignored due to its proximity to the MHC region (chr6) (Schizophrenia Working Group of the Psychiatric Genomics 2014). Using a 25 Mb window around the MHC region (that now includes the PGBD1 locus), the (re)analysis of a large-scale SNP/SCZ GWAS data (Schizophrenia Working Group of the Psychiatric Genomics 2014) collected in multi-ethnic populations, supports that the genomic region between ZSCAN26 and ZSCAN31 on chr6 has several highly significant associations ($R^2 = 0.8$) (fig. 6B). Among the highly significant ($R^2 = 0.8$) SNP–SCZs, is rs33932084 (chr6:28,268,824) a potential regulatory SNP (rSNP) (Liang et al. 2019). This (now reported https://www.ebi.ac.uk/gwas/variants/rs33932084) missense mutation generating variant (N398S), maps to the transposase-derived domain of PGBD1 (fig. 6B). Missense variants associated with SCZ are rare and notable. Although the above data

are preliminary and suggestive at best, we suggest that follow on analysis of this SNP is warranted.

Materials and Methods

PGBD1 and SCAN Domain Evolution

To identify all SCAN family members, the human genome hg19 was downloaded from the UCSC genome browser (Kent et al. 2002; Speir et al. 2016) and translated with EMBOSS (http://emboss.sourceforge.net/) (Rice et al. 2000) in all six reading frames. The SCAN domain motif was extracted from the pfam database (http://pfam.xfam.org/) (Finn et al. 2016) and the motif search was performed with the HMMER software (http://hmmer.org/) (Eddy 1998) on all potential ORFs (including alternative start codons). All SCAN domain hits with a higher score than 25 were considered as significant. Each hit was classified as protein coding when it could be matched (BLASTP) to an entry of the UniprotKB/Swiss-Prot (The UniProt Consortium 2017) database and the other hits were located with TBLASTN, both from the NCBI (https://blast.ncbi.nlm.nih.gov/Blast.cgi). The results were compared with KEGG (Kanehisa et al. 2016), NCBI, Uniprot, and BioMart (Smedley et al. 2015) databases. Proteins which match the domain alignment only partly (<58aa) are not shown (ZFP69B). All other domains were assigned using PFAM and SMART (http://smart.embl-heidelberg.de/) (Letunic and Bork 2018) with the additional option for pfam domains. The figure (supplementary fig. S1B, Supplementary Material online) shows only the longest transcript of each gene.

To identify genomes which contain PGBD1 sequences we performed BLAST and BLASTN searches (NCBI online platform) and found sequence similarities in almost all Eu- and Metatherian species, including *Phascolarctos cinereus* (koala), *Sarcophilus harrisii* (Tasmanian devil), and *Dasypus novemcinctus* (armadillo) but not in phylogenetically older species. To validate whether these first transcripts encoded both PGBD1 associated protein domains, available RNA-seq data of these species were downloaded, and mapped against their reference genome, using STAR. Alignments of PGBD1 amino acid sequences were performed with MEGA7 (http://www.megasoftware.net/) using MUSCLE (Edgar 2004) algorithm with default settings. PGBD1 amino acid sequences were retrieved from the NCBI database.

Phylogenetic Tree of PGBD1 and 2

All sequences(~12k) containing the *pfam* domain Transposase IS4 have been downloaded from *interpro* Uniprot DB (Blum et al. 2021) and aligned with *mafft* (default settings) (Katoh et al. 2002). An initial tree has been calculated with the UPGMA algorithm (default settings) from which a subtree has been manually picked. The subtree includes the cluster of PGBD1 and 2 plus some closely clustering sequences. Identical sequences (CD-HIT 100% identical) and sequences shorter than 250 bp have been

removed. The PGBD1 and 2 sequences (XP_020822236.1 and XP_020822393.1) from Koala have been added manually. The picked transcripts were realigned using muscle (default settings) and a phylogeny tree was built using *MrBayes* (Ronquist and Huelsenbeck 2003) (settings: mixed rate model, single chain and average standard deviation of split frequencies <0.05). The tree was visualized with *iTOL* (Letunic and Bork 2019). Protein domains have been annotated with *hmmerscan* from the pfam db. For visualization reasons another tree of representative PGBD1 and 2 plus invertebrate sequences has been built using *MrBayes* (settings: mixed rate model, single chain, and average standard deviation of split frequencies <0.05). Protein domains were annotated with hmmerscan and CDD (NCBI). The KRAB domain was annotated with Phyre2.

K_A/K_S Ratio Determination

PGBD1 mRNA (CDS) sequences from 11 mammalian non-primate and 18 primate species were manually selected and downloaded from NCBI. The 11 mammalian sequences were picked to represent a heterogeneous group of species. A multiple sequence alignment was performed with MUSCLE (for translated amino acids, default parameters) in UGENE for mammalian and primate specific analysis. The following taxonomy trees were used (it was manually modified to an unrooted tree) and retrieved from the NCBI: mammalian:

(KOALA,(MOUSE,(PONAB,(HUMAN,PANTR),
   MACMU)),(HORSE,(PIG,PHYMC),CALUR,DESRO));
primates: (PROCO,TARSY,(SAIBB,AOTNA,((PONAB,
   (HUMAN,(PATNR,PANPA),GORGO)),((9PRIM,
   COLAB, RHIBE),(CHLSB,MANLE,THEGE,(MACNE,
   MACMU,MACFA),CERAT)))));

The K_A/K_S ratios were calculated for different regions using PAML (version 4.9) (Yang 2007) (M0) [overall, N-terminal (aa 1–290), C-terminal (aa 291–809), SCAN (aa 40–142), KRAB (aa 211–267), DDBD1 (aa 405–541), and DDBD2 (aa 750–804), reference is the human protein sequence of PGBD1].

Neutral and Adaptive Evolution

Adaptive evolution of primates in the mammalian tree was tested in PAML with M1a versus M2a (primates were foreground, all others background), with a χ^2 test (df = 2). Foregrounds (in PAML) are manually marked branches, which are tested against a background (unmarked). Adaptive evolution of the KRAB(-like) region in primates was tested with M1 versus M2 with χ^2 test (df = 1).

Dating Horizontal Gene Transfer

Ensembl synteny browser could not allocate a syntenic region between monotremes and human around the PGBD1 locus. Thus rather than synteny data we employ a recent method that aims to infer whether absence of a candidate gene in a given taxon is evidence for homology search

failure or reflects true absence. AbSENSE (Weisman et al. 2020) was run to test the possibility that PGBD1 was not detected in monotremes and reptiles due to failure of homology detection.

Evolutionary distances of nine species pairs (human–rhesus macaque, human–Ma's night monkey, human–goat, human–camel, human–koala, human–platypus, human–American alligator, human–green anole and African clawed frog) have been calculated as described in the Weisman et al. (2020): orthologs have been retrieved from BUSCO curated vertebrate dataset. A total of 73 genes were common to all selected species. Isoforms have been selected according to their IsoSel score (Philippon et al. 2017). Sequences were aligned with MUSCLE (default) on a gene by gene basis and concatenated to one alignment. The evolutionary distances were calculated with protdist (PHYLIP, default). The focal species was human. Bitscores were calculated with BLATSP (NCBI). Significance testing was performed as proscribed (Weisman et al. 2020).

PGBD1 Conservation in Rodent Model Organisms
The PGBD1 exon architecture and conservation track were retrieved from the UCSC genome browser (hg19). Multiple sequence alignment of mammalian PGBD1 sequences were used to detect the conservation of the catalytic domain (mouse: XP_030103153.1, rat: XP_017456282.1).

Transposon Excision and Transposition Assays
To detect transposon excision events from the donor plasmids, plasmid DNA was isolated from the transfected cells 48 h posttransfection, using standard phenol/chloroform extraction method, followed by ethanol precipitation. To detect those plasmids that were recircularized following transposon excision, followed by cellular DNA repair, the extracted plasmids were subjected to a nested PCR, whereas a *colony forming assay* was used to detect stable integration (Supplementary Material online).

Generation of NPCs and Neurons from hESCs
The work with human embryonic stem cells (hESC_H1) was performed under the license approved by the Robert Koch Institute (A. Prigione # AZ: 3.04.02/0077-E01). The protocols of NPC derivation (Lorenz et al. 2017) and the generation of midbrain dopaminergic neurons (Reinhardt et al. 2013).

Depletion of PGBD1
To deplete PGBD1, we first used a CRISPR/Cas9-mediated KO approach, however, no stable, proliferative KO line could be generated in either hESCs or NPCs (supplementary fig. S8A, Supplementary Material online), suggesting an essential function of PGBD1 in cell survival. Alternatively, we applied knock-down (KD) RNA depletion strategies using (1) dCas9-CRISPR-KRAB-MeCP2 (Yeo et al. 2018) and (2) miRNA/SB100X (RNAi) methods (Bunse et al. 2014). Although (1) was used as a proof of concept,

(2) combined with the *Sleeping Beauty* transposon system (Mates et al. 2009), was suitable to generate stable KD NPC clones that were subjected to further analyses. PGBD1 KD and OE cell lines were also generated in human SHEP neuroblastoma cells.

RNA-Sequencing
For transcriptome analysis, the KD-PGBD1 NPC and the control KD-Scr (scramble) samples (in three independent replicates) were generated using the miRNA/SB100X (RNAi) technology (Supplementary Material online). The KD-PGBD1 stable clones had ~65% depletion of PGBD1 on protein level (supplementary fig. S8B, Supplementary Material online).

To generate OE HA-PGBD1 samples, NPCs were transfected by using pTR-HA-PGBD1 plasmid.

Bioinformatic Analyses
RNA-sequencing, ChIP-seq for histone tail modifications (peak calling), ATAC-seq in genomic PGBD1 peak regions, GRO-Seq data analysis are detailed in Supplementary Material online.

Stable Isotope Labeling by Amino Acids in Cell Culture
Two populations of HEK293 cells were cultivated in cell culture for 3 weeks. One population of cells was fed with growth medium containing normal amino acids ("light cell population"). The second population of cells was cultured in growth medium, containing amino acids labeled with stable heavy isotopes ($^{13}C_6$-$^{15}N_4$ L-arginine; $^{13}C_6$-$^{15}N_2$ L-lysine) ("heavy cell population"). In our experimental approach, untagged PGBD1 and HA-tagged PGBD1 were overexpressed in both conditions, thus no systematic bias will be introduced by OE. The overexpressing plasmids encoding the HA-tagged or untagged PGBD1 were transfected into the two cell populations of HEK293 cells, respectively (Supplementary Material online). The significance of protein–protein interaction was calculated as described in (Cox and Mann 2008) in Supplementary Material online.

Immunofluorescence Microscopy
The cells were seeded on coverslips in 12-well cell culture plates (100,000 cells/well). Forty-eight hours after transfection, cells were fixed with 4% paraformaldehyde (Sigma) supplemented with Hoechst 33,342 (1:1,250, Invitrogen) in PBS for 15 min, and permeabilized with 0.1% Triton X-100 in PBS for 2 min. Coverslips were incubated with primary antibodies for overnight at 4°C, then washed three times with PBS, followed by an incubation using secondary antibodies for 60 min. After an additional washing step, the samples were mounted using ProLong® Gold antifade reagent (Invitrogen). The images were taken using a Leica LSM710 point-scanning single photon confocal microscope. For RNA-FISH, Stellaris RNA-FISH probes

labeled with Quasar 570 Dye for NEAT1_2 (SMF-2037-1) (1:100, Biosearch Technologies) were used and subsequently subjected to immunofluorescence staining (Supplementary Material online).

ChiP-Exonuclease (exo) Assay

The ChiP-exonuclease assay protocol was performed as in Serandour's method (Serandour et al. 2013). The libraries were quantified by using the KAPA library quantification kit for Illumina sequencing platforms (KAPA Biosystems, KK4824) and sequenced on HiSeq. For details and for peak calling see (Supplementary Material online).

Electrophoretic Mobility Shift ASSAY

Approximately 11×10^6 HEK293 cells were transfected with 12 μg pTRHA-PGBD1 plasmids encoding HA-PGBD1 fusion protein. Two days posttransfection, the HA-PGBD1 protein was purified by using EZview™ Red Anti-HA Agarose beads or HA tagged Protein PURIFICATION KIT-BioZol (MBL-3320). Binding reactions (Supplementary Material online) were performed in 25 μl volumes on ice for 20 min. Protein–DNA complexes were separated by electrophoresis in 6% nondenaturing polyacrylamide gels at 4°C. Electrophoresis was performed at constant voltage of 200 V for 3 h. The fluorescent signal was detected by using a BioRad ChemiDoc™ MP Imaging System.

Quantification and Statistical Analysis

All data are shown as mean and standard deviation (s.d.) of multiple replicates/experiments. Analysis of all experimental data was done with GraphPad Prism 5 (San Diego, CA, USA). P-values were calculated with two-sided, unpaired t-test following the tests. P-values <0.05 were considered significant.

Supplementary Material

Supplementary data are available at *Molecular Biology and Evolution* online.

Acknowledgments

Author Contributions

The work was conceptualized by Z.I., T.R., and L.D.H. The original draft was written by Z.I. The manuscript was edited by L.D.H. Experiments were performed and the methodologies were worked out by T.R., A.Z., C.S., G.W., O.K., G.I., M.B. Bioinformatic analysis was performed by A.Pa., K.R., M.S., L.D.H. The paper was reviewed by T.R., A.P., K.R., A.S., S.P., M.S., T.I.O., A.Pr.

Conflict of interest The authors declare no competing interests.

References

Schizophrenia Working Group of the Psychiatric Genomics C. 2014. Biological insights from 108 schizophrenia-associated genetic loci. *Nature*. **511**:421–427.

An H, Williams NG, Shelkovnikova TA. 2018. NEAT1 and paraspeckles in neurodegenerative diseases: a missing lnc found? *Noncoding RNA Res*. **3**:243–252.

Bansal V, Mitjans M, Burik CAP, Linner RK, Okbay A, Rietveld CA, Begemann M, Bonn S, Ripke S, de Vlaming R, et al. 2018. Genome-wide association study results for educational attainment aid in identifying genetic heterogeneity of schizophrenia. *Nat Commun*. **9**:3078.

Blake DJ, Nawrotzki R, Loh NY, Gorecki DC, Davies KE. 1998. Beta-dystrobrevin, a member of the dystrophin-related protein family. *Proc Natl Acad Sci U S A*. **95**:241–246.

Blum M, Chang HY, Chuguransky S, Grego T, Kandasaamy S, Mitchell A, Nuka G, Paysan-Lafosse T, Qureshi M, Raj S, et al. 2021. The InterPro protein families and domains database: 20 years on. *Nucleic Acids Res*. **49**:D344–D354.

Brinkmann BG, Agarwal A, Sereda MW, Garratt AN, Muller T, Wende H, Stassart RM, Nawaz S, Humml C, Velanac V, et al. 2008. Neuregulin-1/ErbB signaling serves distinct functions in myelination of the peripheral and central nervous system. *Neuron*. **59**:581–595.

Bunse M, Bendle GM, Linnemann C, Bies L, Schulz S, Schumacher TN, Uckert W. 2014. RNAi-mediated TCR knockdown prevents autoimmunity in mice caused by mixed TCR dimers following TCR gene transfer. *Mol Ther*. **22**:1983–1991.

Buonanno A, Kwon OB, Yan L, Gonzalez C, Longart M, Hoffman D, Vullhorst D. 2008. Neuregulins and neuronal plasticity: possible relevance in schizophrenia. *Novartis Found Symp*. **289**:165–177; discussion 177-169, 193-165.

Cadinanos J, Bradley A. 2007. Generation of an inducible and optimized piggyBac transposon system. *Nucleic Acids Res*. **35**:e87.

Cary LC, Goebel M, Corsaro BG, Wang HG, Rosen E, Fraser MJ. 1989. Transposon mutagenesis of baculoviruses: analysis of *Trichoplusia ni* transposon IFP2 insertions within the FP-locus of nuclear polyhedrosis viruses. *Virology* **172**:156–169.

Chen Q, Luo W, Veach RA, Hickman AB, Wilson MH, Dyda F. 2020. Structural basis of seamless excision and specific targeting by piggyBac transposase. *Nat Commun*. **11**:3446.

Chen YL, Tu PC, Lee YC, Chen YS, Li CT, Su TP. 2013. Resting-state fMRI mapping of cerebellar functional dysconnections involving multiple large-scale networks in patients with schizophrenia. *Schizophr Res*. **149**:26–34.

Cheng C-Y, Vogt A, Mochizuki K, Yao M-C. 2010. A domesticated piggyBac transposase plays key roles in heterochromatin dynamics and DNA cleavage during programmed DNA deletion in *Tetrahymena thermophila*. *Mol Biol Cell*. **21**:1753–1762.

Ding S, Wu X, Li G, Han M, Zhuang Y, Xu T. 2005. Efficient transposition of the piggyBac (PB) transposon in mammalian cells and mice. *Cell* **122**:473–483.

Fong KW, Li Y, Wang W, Ma W, Li K, Qi RZ, Liu D, Songyang Z, Chen J. 2013. Whole-genome screening identifies proteins localized to distinct nuclear bodies. *J Cell Biol*. **203**:149–164.

Gao QQ, McNally EM. 2015. The dystrophin complex: structure, function, and implications for therapy. *Compr Physiol*. **5**: 1223–1239.

Gauthier MK, Kosciuczyk K, Tapley L, Karimi-Abdolrezaee S. 2013. Dysregulation of the neuregulin-1-ErbB network modulates endogenous oligodendrocyte differentiation and preservation after spinal cord injury. *Eur J Neurosci*. **38**:2693–2715.

Gouvea-Junqueira D, Falvella ACB, Antunes A, Seabra G, Brandao-Teles C, Martins-de-Souza D, Crunfli F. 2020. Novel treatment strategies targeting myelin and oligodendrocyte dysfunction in schizophrenia. *Front Psychiatry*. **11**:379.

Guerineau M, Bessa L, Moriau S, Lescop E, Bontems F, Mathy N, Guittet E, Bischerour J, Betermier M, Morellet N. 2021. The unusual structure of the PiggyMac cysteine-rich domain reveals zinc finger diversity in PiggyBac-related transposases. *Mob DNA*. **12**:12.

Hoffmann A, Ziller M, Spengler D. 2019. Progress in iPSC-based modeling of psychiatric disorders. *Int J Mol Sci*. **20**:4896.

Humbertclaude V, Riant F, Krams B, Zimmermann V, Nagot N, Annequin D, Echenne B, Tournier-Lasserve E, Roubertie A, Episodic Syndrome C. 2020. Cognitive impairment in children with CACNA1A mutations. *Dev Med Child Neurol*. **62**:330–337.

Huo Y, Li S, Liu J, Li X, Luo XJ. 2019. Functional genomics reveal gene regulatory mechanisms underlying schizophrenia risk. *Nat Commun*. **10**:670.

Ishizuka A, Hasegawa Y, Ishida K, Yanaka K, Nakagawa S. 2014. Formation of nuclear bodies by the lncRNA Gomafu-associating proteins Celf3 and SF1. *Genes Cells*. **19**: 704–721.

Kampinga HH, Bergink S. 2016. Heat shock proteins as potential targets for protective strategies in neurodegeneration. *Lancet Neurol*. **15**:748–759.

Kampinga HH, Hageman J, Vos MJ, Kubota H, Tanguay RM, Bruford EA, Cheetham ME, Chen B, Hightower LE. 2009. Guidelines for the nomenclature of the human heat shock proteins. *Cell Stress Chaperones*. **14**:105–111.

Katoh K, Misawa K, Kuma K, Miyata T. 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res*. **30**:3059–3066.

Katsel P, Roussos P, Fam P, Khan S, Tan W, Hirose T, Nakagawa S, Pletnikov MV, Haroutunian V. 2019. The expression of long noncoding RNA NEAT1 is reduced in schizophrenia and modulates oligodendrocytes transcription. *NPJ Schizophr*. **5**:3.

Kelley LA, Mezulis S, Yates CM, Wass MN, Sternberg MJ. 2015. The Phyre2 web portal for protein modeling, prediction and analysis. *Nat Protoc*. **10**:845–858.

Khalfallah O, Ravassard P, Lagache CS, Fligny C, Serre A, Bayard E, Faucon-Biguet N, Mallet J, Meloni R, Nardelli J. 2009. Zinc finger protein 191 (ZNF191/Zfp191) is necessary to maintain neural cells as cycling progenitors. *Stem Cells*. **27**:1643–1653.

Kitazawa M, Ohnuma T, Takebayashi Y, Shibata N, Baba H, Ohi K, Yasuda Y, Nakamura Y, Aleksic B, Yoshimi A, et al. 2012. No associations found between the genes situated at 6p22.1, HIST1H2BJ, PRSS16, and PGBD1 in Japanese patients diagnosed with schizophrenia. *Am J Med Genet B Neuropsychiatr Genet*. **159B**:456–464.

Lein ES, Hawrylycz MJ, Ao N, Ayres M, Bensinger A, Bernard A, Boe AF, Boguski MS, Brockway KS, Byrnes EJ, et al. 2007. Genome-wide atlas of gene expression in the adult mouse brain. *Nature*. **445**:168–176.

Lellahi SM, Rosenlund IA, Hedberg A, Kiaer LT, Mikkola I, Knutsen E, Perander M. 2018. The long noncoding RNA NEAT1 and nuclear paraspeckles are up-regulated by the transcription factor HSF1 in the heat shock response. *J Biol Chem*. **293**:18965–18976.

Letunic I, Bork P. 2019. Interactive Tree Of Life (iTOL) v4: recent updates and new developments. *Nucleic Acids Res*. **47**:W256–W259.

Li J, Zhu L, Guan F, Yan Z, Liu D, Han W, Chen T. 2018. Relationship between schizophrenia and changes in the expression of the long non-coding RNAs Meg3, Miat, Neat 1, and Neat2. *J Psychiatr Res*. **106**:22–30.

Liang X, Wang S, Liu L, Du Y, Cheng B, Wen Y, Zhao Y, Ding M, Cheng S, Ma M, et al. 2019. Integrating genome-wide association study with regulatory SNP annotation information identified candidate genes and pathways for schizophrenia. *Aging (Albany NY)*. **11**:3704–3715.

Loh NY, Ambrose HJ, Guay-Woodford LM, DasGupta S, Nawrotzki RA, Blake DJ, Davies KE. 1998. Genomic organization and refined mapping of the mouse beta-dystrobrevin gene. *Mamm Genome*. **9**:857–862.

Lorenz C, Lesimple P, Bukowiecki R, Zink A, Inak G, Mlody B, Singh M, Semtner M, Mah N, Auré K, et al. 2017. Human iPSC-derived neural progenitors are an effective drug discovery model for neurological mtDNA disorders. *Cell Stem Cell*. **20**:659–674.e659.

Ma L, Tang J, Wang D, Zhang W, Liu W, Liu X-H, Gong W, Yao Y-G, Chen X. 2013. Evaluating risk loci for schizophrenia distilled from genome-wide association studies in Han Chinese from Central China. *Mol Psychiatry*. **18**:638–639.

Margolin JF, Friedman JR, Meyer WK, Vissing H, Thiesen HJ, Rauscher FJ, 3rd. 1994. Kruppel-associated boxes are potent transcriptional repression domains. *Proc Natl Acad Sci U S A*. **91**:4509–4513.

Mates L, Chuah MK, Belay E, Jerchow B, Manoj N, Acosta-Sanchez A, Grzela DP, Schmitt A, Becker K, Matrai J, et al. 2009. Molecular evolution of a novel hyperactive Sleeping Beauty transposase enables robust stable gene transfer in vertebrates. *Nat Genet*. **41**: 753–761.

McCluggage F, Fox AH. 2021. Paraspeckle nuclear condensates: global sensors of cell stress? *Bioessays*. **43**:e2000245.

Mei L, Nave KA. 2014. Neuregulin-ERBB signaling in the nervous system and neuropsychiatric diseases. *Neuron*. **83**:27–49.

Mercer TR, Qureshi IA, Gokhan S, Dinger ME, Li G, Mattick JS, Mehler MF. 2010. Long noncoding RNAs in neuronal-glial fate specification and oligodendrocyte lineage maturation. *BMC Neurosci*. **11**:14.

Modic M, Grosch M, Rot G, Schirge S, Lepko T, Yamazaki T, Lee FCY, Rusha E, Shaposhnikov D, Palo M, et al. 2019. Cross-regulation between TDP-43 and paraspeckles promotes pluripotency-differentiation transition. *Mol Cell*. **74**:951–965.e913.

Morellet N, Li X, Wieninger SA, Taylor JL, Bischerour J, Moriau S, Lescop E, Bardiaux B, Mathy N, Assrir N, et al. 2018. Sequence-specific DNA binding activity of the cross-brace zinc finger motif of the piggyBac transposase. *Nucleic Acids Res*. **46**: 2660–2677.

Newman JC, Bailey AD, Fan HY, Pavelitz T, Weiner AM. 2008. An abundant evolutionarily conserved CSB-PiggyBac fusion protein expressed in Cockayne syndrome. *PLoS Genet*. **4**:e1000031.

Parker KL, Narayanan NS, Andreasen NC. 2014. The therapeutic potential of the cerebellum in schizophrenia. *Front Syst Neurosci*. **8**: 163.

Picard H, Amado I, Mouchet-Mages S, Olie JP, Krebs MO. 2008. The role of the cerebellum in schizophrenia: an update of clinical, cognitive, and functional evidences. *Schizophr Bull*. **34**: 155–172.

Reinhardt P, Glatza M, Hemmer K, Tsytsyura Y, Thiel CS, Hoing S, Moritz S, Parga JA, Wagner L, Bruder JM, et al. 2013. Derivation and expansion using only small molecules of human neural progenitors for neurodegenerative disease modeling. *PLoS One*. **8**: e59252.

Ronquist F, Huelsenbeck JP. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics*. **19**:1572–1574.

Sarkar A, Sim C, Hong YS, Hogan JR, Fraser MJ, Robertson HM, Collins FH. 2003. Molecular evolutionary analysis of the widespread piggyBac transposon family and related "domesticated" sequences. *Mol Genet Genomics*. **270**:173–180.

Sathyanesan A, Zhou J, Scafidi J, Heck DH, Sillitoe RV, Gallo V. 2019. Emerging connections between cerebellar development, behaviour and complex brain disorders. *Nat Rev Neurosci*. **20**:298–313.

Schmitt A, Simons M, Cantuti-Castelvetri L, Falkai P. 2019. A new role for oligodendrocytes and myelination in schizophrenia and affective disorders? *Eur Arch Psychiatry Clin Neurosci*. **269**: 371–372.

Schrode N, Ho SM, Yamamuro K, Dobbyn A, Huckins L, Matos MR, Cheng E, Deans PJM, Flaherty E, Barretto N, et al. 2019. Synergistic effects of common schizophrenia risk variants. *Nat Genet*. **51**: 1475–1485.

Silberberg G, Darvasi A, Pinkas-Kramarski R, Navon R. 2006. The involvement of ErbB4 with schizophrenia: association and expression studies. *Am J Med Genet B Neuropsychiatr Genet*. **141B**: 142–148.

Stefansson H, Ophoff RA, Steinberg S, Andreassen OA, Cichon S, Rujescu D, Werge T, Pietilainen OP, Mors O, Mortensen PB, et al. 2009. Common variants conferring risk of schizophrenia. *Nature*. **460**:744–747.

Stefansson H, Sigurdsson E, Steinthorsdottir V, Bjornsdottir S, Sigmundsson T, Ghosh S, Brynjolfsson J, Gunnarsdottir S, Ivarsson O, Chou TT, et al. 2002. Neuregulin 1 and susceptibility to schizophrenia. *Am J Hum Genet*. **71**:877–892.

Takata A, Matsumoto N, Kato T. 2017. Genome-wide identification of splicing QTLs in the human brain and their enrichment among schizophrenia-associated loci. *Nat Commun*. **8**:14519.

Tripathi V, Ellis JD, Shen Z, Song DY, Pan Q, Watt AT, Freier SM, Bennett CF, Sharma A, Bubulya PA, et al. 2010. The nuclear-retained noncoding RNA MALAT1 regulates alternative splicing by modulating SR splicing factor phosphorylation. *Mol Cell*. **39**: 925–938.

Tycko J, DelRosso N, Hess GT, Aradhana BA, Mukund A, Van MV, Ego BK, Yao D, Spees K, et al. 2020. High-throughput discovery and characterization of human transcriptional effectors. *Cell*. **183**:2020–2035.e2016.

van Arensbergen J, Pagie L, FitzPatrick VD, de Haas M, Baltissen MP, Comoglio F, van der Weide RH, Teunissen H, Vosa U, Franke L, et al. 2019. High-throughput identification of human SNPs affecting regulatory element activity. *Nat Genet*. **51**:1160–1169.

Waite A, Brown SC, Blake DJ. 2012. The dystrophin–glycoprotein complex in brain development and disease. *Trends Neurosci*. **35**:487–496.

Wang J, Dai X, Berry LD, Cogan JD, Liu Q, Shyr Y. 2019. HACER: an atlas of human active enhancers to interpret regulatory variants. *Nucleic Acids Res*. **47**:D106–D112.

Wang M, Wei PC, Lim CK, Gallina IS, Marshall S, Marchetto MC, Alt FW, Gage FH. 2020. Increased neural progenitor proliferation in a hiPSC model of autism induces replication stress-associated genome instability. *Cell Stem Cell*. **26**:221–233.e226.

West JA, Davis CP, Sunwoo H, Simon MD, Sadreyev RI, Wang PI, Tolstorukov MY, Kingston RE. 2014. The long noncoding RNAs NEAT1 and MALAT1 bind active chromatin sites. *Mol Cell*. **55**: 791–802.

Williams AJ, Blacklow SC, Collins T. 1999. The zinc finger-associated SCAN box is a conserved oligomerization domain. *Mol Cell Biol*. **19**:8526–8535.

Wilson MH, Coates CJ, George AL, Jr. 2007. PiggyBac transposon-mediated gene transfer in human cells. *Mol Ther*. **15**:139–145.

Witzgall R, O'Leary E, Leaf A, Onaldi D, Bonventre JV. 1994. The Kruppel-associated box-A (KRAB-A) domain of zinc finger proteins mediates transcriptional repression. *Proc Natl Acad Sci U S A*. **91**:4514–4518.

Wu SC, Meir YJ, Coates CJ, Handler AM, Pelczar P, Moisyadi S, Kaminski JM. 2006. piggyBac is a flexible and highly active transposon as compared to sleeping beauty, Tol2, and Mos1 in mammalian cells. *Proc Natl Acad Sci U S A*. **103**: 15008–15013.

Yamazaki T, Hirose T. 2015. The building process of the functional paraspeckle with long non-coding RNAs. *Front Biosci (Elite Ed)*. **7**:1–41.

Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol*. **24**:1586–1591.

Yeganeh-Doost P, Gruber O, Falkai P, Schmitt A. 2011. The role of the cerebellum in schizophrenia: from cognition to molecular pathways. *Clinics (Sao Paulo)*. **66**(Suppl 1):71–77.

Yeo NC, Chavez A, Lance-Byrne A, Chan Y, Menn D, Milanova D, Kuo CC, Guo X, Sharma S, Tung A, et al. 2018. An enhanced CRISPR repressor for targeted mammalian gene regulation. *Nat Methods*. **15**:611–616.

Yue WH, Wang HF, Sun LD, Tang FL, Liu ZH, Zhang HX, Li WQ, Zhang YL, Zhang Y, Ma CC, et al. 2011. Genome-wide association study identifies a susceptibility locus for schizophrenia in Han Chinese at 11p11.2. *Nat Genet*. **43**:1228–1231.

Zhang W, Muck-Hausl M, Wang J, Sun C, Gebbing M, Miskey C, Ivics Z, Izsvak Z, Ehrhardt A. 2013. Integration profile and safety of an adenovirus hybrid-vector utilizing hyperactive sleeping beauty transposase for somatic integration. *PLoS One*. **8**: e75344.

Zhuchenko O, Bailey J, Bonnen P, Ashizawa T, Stockton DW, Amos C, Dobyns WB, Subramony SH, Zoghbi HY, Lee CC. 1997. Autosomal dominant cerebellar ataxia (SCA6) associated with small polyglutamine expansions in the alpha 1A-voltage-dependent calcium channel. *Nat Genet*. **15**: 62–69.