

Supplementary Tables

Table S1. PGBD1 protein interacting partners belong to the SCAN family of transcription factors

Table S2. ChiP-Exo analysis of PGBD1 binding in NPCs and neurons

Table S3. Gene ontology analyses of the PGBD1 binding target genes

Table S4. qPCR analysis of dCAS9 CRISPR-KRAB-McCP2

Table S5. RNA-seq analysis of PGBD1 depleted (KD) NPCs

Table S6. Intersection analysis of ChiP-Exo targets and DEGs upon PGBD1 depletion

Table S7. MS-SILAC of PGBD1 interactome

Table S8. Intersection analysis of PGBD1 ChiP-Exo and NEAT1 CHIRP targets

Supplementary Figure legends

Fig. S1. Schematic representation of the domain structure of the human SCAN-domain protein family

(A) abSENSE method was used to calculate the probability that homologs of PGBD1 and PGBD2 would fail to be detected by a homology search (using BLAST method) in platypus, alligator, lizard and frog. From the rate parameters of the two genes where orthologs are described (bitscore: y axis), given the evolutionary distances between these species (x axis), we can infer the probability that an ortholog is truly absent in other species, given their evolutionary distance from the focal species (human), rather than simply not findable. For all of the species without an identified ortholog we can reject the hypothesis that the ortholog would not be expected to be detected by homology search even if present. Thus, absence of an orthology cannot be ascribed to failure of homology search.

(B) Schematic representation of the domain structure of the human SCAN-domain protein family of transcription factors. From each protein the longest isoform is presented, including the protein domain annotation from SMART and PFAM. Filled boxes indicate domains, which were annotated in both databases, continuous lines indicate domains, which are annotated in SMART, but not in PFAM and dashed lines indicate domains, which are annotated in PFAM but not in SMART. Note that in PGBD1 the ZNF_C2H2 domain is replaced by the transposase-derived domain.

Fig. S2. Phylogenetic analysis of PGBD1 and PGBD2

Phylogenetic tree of Pgbd1 and Pgbd2. Both Pgbd1 and Pgbd2 are specific to the Mammalian lineage, not including Monotemes (Therian). They are closely related but differ in their protein domain structure. All sequences (~12k) containing the pfam domain Transposase IS4 have been downloaded from interpro Uniprot DB and aligned with mafft. An initial tree has been calculated with the UPGMA algorithm from which a subtree has been manually picked. The subtree includes the cluster of Pgbd1 and Pgbd2 plus some closely clustering sequences. Identical sequences and sequences shorter than 250 bp have been removed. The Pgbd1 and Pgbd2 sequences from Koala have been added manually. The picked transcripts were realigned using muscle and a phylogeny tree was build using MrBayes. Annotated protein domains originate from the pfam db. While Pgbd2 carries exclusively the IS4 transposase-like domain, Pgbd1 has an additional SCAN and occasionally KRAB domains at the N-terminal. Average pairwise similarity score of ~ 63% of the aligned region which spans 1324 bp exceeding the borders of the annotated transposase IS4 domain, calculated by distance matrix of Ugene).

Fig. S3. EggNOG phylogeny analyses for PGBD1 and PGBD2

A) EggNOG phylogeny tree for PGBD1, PGBD2 and the closest relatives showing SMART domain predictions. Note that the KRAB domain is not predicted in *Homo sapiens*.

B) EggNOG phylogeny analysis for PGBD1, PGBD2 and the closest relatives.

Fig. S4. The plasticity of PGBD1 evolution

(A) The transposase-derived domains of human (hPGBD1) and rat (rPgbd1) sequences are highly similar (87% identity). Amino acid sequence alignment of the human and rat transposase-derived domains of Pgbd1.

(B) Protein sequence alignment of the SCAN domain template c3lhrA_ (identified by Phyre2) and PGBD1 from various mammalian and marsupial species.

(C) The structure of rPgbd1 is not conserved in the rat genome. Exon architecture of the human and rat Pgbd1 genes. The human hPGBD1 consists of 7 exons. The N-terminal (SCAN and KRAB) domains (exons 1-6 of hPGBD1) are not detectable in the rat. Arrows represent the positions of the

PCR primers that were used to analyse the N-terminus of the rPgbd1. Sequences of the forward and reverse primer that were used in the final analysis are shown.

(D) PCR amplified DNA fragments were cloned into pJET1.2 vector (Fermentas) then the individually purified plasmids were digested by BglIII restriction endonuclease then further analysed by agarose gel electrophoresis. DNA fragment of the lanes 1 and 6 PCR products were analysed by Sanger sequencing.

(E) The predicted amino acid sequences of two PCR products amplified from the rat genome (yellow) identify several STOP codons upstream of the transposase-derived domain (gray).

(F) Protein sequence alignment of the KRAB domain template d1v65a_ (identified by Phyre2) and PGBD1 from various mammalian and marsupial species.

Fig. S5. The enrichment of the human PGBD1 and PGBD2 expression

(A) PCR-based *transposon excision repair assay* detects no activity of the PGBD1. (Upper panel) Schematic view of the PCR-based 2-step transposon excision assay in HEK-293 cells. Panel shows the position of the primers on the DNA plasmids that were used in the assay. In case of excision, the 2nd PCR step amplifies an approximately a 357 bp product. Luciferase (Luc) is used as a control protein. (Lower panel) Agarose gel electrophoresis image of the PCR products from the excision assay. Only the insect codon optimised *piggyBac* transposase (mPB) has detectable activity, indicated by the 374 bp PCR product. To monitor the efficiency of plasmid isolation, PCR on the plasmid backbone (on the ampicillin resistance gene sequence) was also carried out resulted in a 340 bp long PCR product.

(B) Tissue specific gene expression profile of the human PGBD1 protein based on www.gtexportal.org. Note the enriched expression of PGBD1 in the cerebellum and cerebellar hemisphere.

(C) Tissue specific gene expression profile of the human PGBD2 proteins based on www.gtexportal.org. Note that PGBD2 is mostly enriched in spleen and thyroid tissues.

(D) Genome activation status of PGBD1 promoter (1kb upstream to TSS) in human embryonic stem cells (ESC) neural progenitor cells (NPC) and neurons, characterized by selected histone marks for active and poised states. Yellow lines represent the count per million (CPM) values at log2FC.

(E) PGBD1 expression is the highest in ESCs, followed by NPC and differentiated neuron. Transcript levels of PGBD1 in human embryonic stem cells (ESC), neural progenitor cells (NPC) and neurons (GSE118106).

(F) Expression level of the PGBD1 protein in ESC, NPC and differentiated neuron (Western blot). PGBD1:actin ratio is shown at the bottom. ESC (Embryonic stem cell H1_ESC), H1_mDAN1; human H1_ECS-derived neuron, N1H1; human H1_ECS derived neural progenitor cell (NPC). The numbers indicate the ratio of the endogenous PGBD1 and actin proteins (determined by densitometry analysis).

Fig. S6. MS-SILAC analysis of the PGBD1 proteome

(A) The highly confident PGBD1 interactome. Scatter plot of the H/L ratios of the SILAC based AP-MS experiments (label-switch) in HEK293 cells. The normalized H/L ratios are log and z transformed. Significance A was calculated according to (Cox and Mann 2008) with modifications (see methods). The red dots represent the highly significant interacting partners (FDR < 0.05), whereas the black dots are considered as background. Note that the approach does not distinguish between HSAPA6/HSPA7 or DTNA/DTNB, as the detected peptide is shared between both proteins.

(B) Relative transcriptional changes of interactors upon knockdown and overexpression of PGBD1. Heatmap of the TPM normalized expression from RNA-seq in wild type, PGBD1 knockdown and PGBD1 overexpression NPCs. Only interactors with > 10 TPM in wild type NPCs are shown in the heatmap.

(C) Enriched gene sets in the PGBD1 interactome (Molecular Signature Database v7.2). The colour intensities of the dots indicate the P-value, the size of the dot represents the effect size.

Fig. S7. Genome-wide analyses of PGBD1 binding

(A) Genome-wide distribution of the PGBD1 ChIP-exo peaks in the protein coding and non-protein coding regions in human NPCs and neurons.

(B) Heatmap of PGBD1 binding sites merged with H3K4me1 histone mark signal (2kb upstream from RefSeq annotated TSS).

(C) Comparative plot of the significant ($p < 0,05$) Gene Ontology terms (GO- Biological Function) of the protein coding genes, targeted by PGBD1 binding (ChIP-exo).

(D) Heatmap of PGBD1 binding sites merged with H3K4me3 and H3K27me3 histone mark signals (1kb upstream from RefSeq annotated TSS).

(E) Heatmap of PGBD1 binding sites merged with H3K36me3 and H3K27ac histone mark signals (1kb upstream from RefSeq annotated TSS).

(F) Correlation plot of PGBD1 ChIP-exo peaks and ATAC-seq signals from NPCs and differentiated neurons (GSE95023).

Fig. S8. Depletion strategies of PGBD1 expression

(A) Knockout (KO) strategy interfered with cell renewal preventing stable maintenance of a colony. Representative images of knock-out (KO) CRISPR cells. Note that none of the used gRNA constructs was suitable to establish a stable KO-PGBD1 cell line. Long-term (two months) culturing resulted in a differentiated morphology.

(B) Determination of the endogenous protein levels upon knocking down (KD) PGBD1 in NPCs (miRNA/SB100X (RNAi strategy)). (Left panel) Total protein visualized by the BioRad-ChemiDoc™ MP Imaging System. (Right panel) Western blotting. The quantification was performed by normalizing the PGBD1 level with the total protein amount. KD-efficiencies of the unique samples are indicated in %.

(C) Stable knock-down (KD) and overexpressing (OE) of PGBD1 in neuroblastoma SHEP cells. (Left panels) Representative confocal microscopic images of stable miRNA based KD-PGBD1 cell

lines (KD1-PGBD1, KD2-PGBD1) used for functional studies (miRNA/SB100X (RNAi strategy). Scr, scrambled control. mCherry is a marker of the integrated KD miRNA constructs in the stable SHEP cells. (Right panel) Representative confocal microscopy image shows the overexpressed HA-tagged PGBD1 protein in a stable human SHEP cell line that was used in functional experiments.

(D) (Left panels) Western blot analysis of the PGBD1 depletion in stable KD-PGBD1_SHEP cell lines (miRNA/SB100X (RNAi strategy). Red boxes mark the selected lines used for further studies. Numbers indicate the efficacy of the KD determined by the ratio of endogenous PGBD1 and actin. (Right panel) Western blot analysis of stable TET-inducible (doxycyclin) SHEP cell lines overexpressing HA-tagged PGBD1 protein in the presence and absence of doxycycline. IB. immunoblot; aPGBD1: anti-PGBD1 antibody; aActin: anti- α -actinin antibody.

(E) Quantification of transcription levels of the NEAT1_1, NEAT1_2 isoforms by qPCR in stable PGBD1 (KD) SHEP cells (miRNA/SB100X (RNAi strategy). Graph shows the results of three independent experiments in triplicates. P-values: * $p < 0.05$, ** $p < 0.005$.

Fig. S9. The CRISPR-KRAB-MeCP2 repressor approach for depleting PGBD1 expression

(A) The position of the gRNA sequences (Oli-1 and Oli-3) used to target PGBD1 expression (-50 and +200 bp window from transcription start site (TSS)). Double stranded oligonucleotides encoding the gRNA (Oli-1 and Oli-3) were cloned into pSB700 plasmid (Addgene #64046) vector. To deplete the PGBD1 expression level using the dCas9-KRAB-MeCP2 method (Yeo, et al. 2018), the pCAG dCas9-KRAB-2A-EGFP (Addgene #92396) and pSB700_Oli-1/3 constructs were co-transfected into NPCs (ratio 1:1. Scale 400 μ M.

(B) Representative fluorescence microscopy images the NPCs transfected with control Oli-1 and Oli-3 constructs (3 days post-transfection).

(C) Quantification of PGBD1 protein levels and the knock-down efficiency determination by the BioRad ChemiDoc™ MP Imaging System. (Left panel) total protein; (Middle panel Western blot; (Right panel) KD efficiency of the independent oligos, Oli-1 and Oli-3. The KD-efficiency was calculated as the ratio of PGBD1 and actin. Specifically, in control the actin level was 10127441 as compared to 398152 for PGBD1. In Oli-1 and Oli-3 the same figures are 10490207/230295 and 13221556/265594 respectively. In control the PGBD1 to control ratio was thus 0.0393, while it was

0.021 and 0.02 in Oli1 and Oli3 respectively. With PGBD1 levels at 56% and 51% of control level, respectively, the KD efficiency is 44% and 49% for Oli-1 and Oli-3 respectively. For qPCR verification of reduced mRNA levels of PGBD1 see S Table 4.

(D) Representative confocal microscopic images of stable PGBD1-depleted SHEP cells (KD1-PGBD1, KD2-PGBD1) (miRNA/SB100X (RNAi strategy). Non-transfected (Ctrl) SHEP cells and cells transfected with the control KD-Scr miRNA scrambled construct are served as controls. The NEAT1_2 transcripts (gold) are visualized by fluorescent *in situ* hybridization (FISH), whereas PGBD1 (green) with immunostaining (blue, DAPI). mCherry is a marker of the integrated KD miRNA constructs in the stable SHEP cells. Note the elevated NEAT1_2 transcription in stable PGBD1-depleted SHEP cells. Scale 10 μ M (Ctrl SHEP), 5 μ M (rest of the images).

MATERIALS and METHODS

Cloning of plasmid DNA constructs

MER75 and piggyBac excision constructs: pCAGGS-Venus was cleaved with PstI, treated with T4 DNA polymerase and Antarctic phosphatase and finally gel-isolated. This vector was ligated to blunt, 5'-phosphorylated PCR fragment of the moth piggyBac transposon from pUC19-XL-Neo (with 'TTAA on both ends). The fragment was amplified with the high fidelity Pfu Ultra II Fusion HS DNA polymerase (Agilent) according to the manufacturer's recommendations using the primers: Piggy-forw (5'-P-TTAACCCTAGAAAGATAATCATATTGT-3') and Piggy-rev (5'-P-TTAACCCTAGAAAGATAGTCTGCG-3').

The pTR-HA-PGBD1 expression plasmid overproducing the PGBD1 protein utilized for MS-SILAC experiments was generated as follows: The PGBD1 ORF was amplified from a cDNA library prepared from HeLa cells using DNA polymerase Pfu Ultra II Fusion HS (Agilent Technologies) and PGBD1 Forw (5'-ATGTATGAAGCTTTGCCAGGC-3') and PGBD1 Rev (5'-TTGCGGCCGCCTAATCTGACAG-ATGAGCATTTGT-3') primer pairs. A NotI restriction site was added to the 3' end of the fragment for further cloning processes. PCR program was as follows: 95°C 150s pre-denaturation, then 35 cycles of 95°C for 30s, 63°C for 30s, 72°C for 60s, and finally 60s final extension at 72°C. A PCR product ~ 2500 bp was digested with NotI, gel isolated (Qiaprep Gel Isolation Kit) and cloned into pHA5 expression vector (EcoRV and NotI), resulting in pTR-HA-PGBD1. In pTR-HA-PGBD1, the expression of PGBD1 ORF is driven by CAG, and the

expressed protein is tagged with an N-terminal hemagglutinin (HA). To generate a non-tagged version (as a control construct for MS-SILAC), the DNA sequence encoding the HA-tag was eliminated by NotI/EcoRI digestion from pHA5. Before NotI digestion the EcoRI site was filled up using Klenow DNA polymerase. The PCR product encoding the PGBD1 ORF was inserted.

The pTOV-T11-TR-HA-PGBD1, inducible expression construct was generated by NcoI and NotI fragment replacement from pTR-HA-PGBD1 plasmid into previously cleaved pTOV-T11 plasmid (Heinz, et al. 2011) by SalI and filled up using Klenow DNA polymerase.

For knocking down PGBD1, miR-expressing SB transposon vectors were generated by inserting the following elements into the pT2/HB transposon plasmid (Cui, et al. 2002): MPSV promoter and 5' intron of the retroviral vector MP71 (Engels, et al. 2003), mCherry as a marker gene followed by the posttranscriptional regulatory element (PRE) of woodchuck hepatitis virus, and the poly(A) signal (PAS) of psiCHECK2 (Promega, Mannheim, Germany). Redirected miRs targeting human PGBD1 were generated as described previously (Bunse, et al. 2014) and introduced between PRE and PAS. In short, RNAi target sites were identified using BLOCK-iT RNAi Designer (Thermo Fisher Scientific) and redirected miRs were generated by overlap PCR using synthesized DNA oligos encoding the 21-nt antisense sequences and a plasmid encoding mouse miR-155 (Chung, et al. 2006) as template.

Construct	Name	Antisense sequence (5'->3')
620	Control	TAG GTG CTC TTC ATC TTG TTG
621	PGBD1-638	AGC CAT TGA TGA CAA AGT TCT
622	PGBD1-749	TCC AGA GGA TAT GTC TTC ACA
623	PGBD1-801	TCC ACT TCC TGT CTC TAG ATT
624	PGBD1-867	ATA TTC TGC CAC CAT TGG GTG
625	PGBD1-1009	TTT CCT GGA GAT GAG CTG ATC

Depletion of PGBD1

First, we used a CRISPR/Cas9-mediated knock-out (KO) approach. However, no stable, proliferative KO line could be generated in either hESCs or neuroprogenitor cells (NPC). As an alternative, we applied knock-down (KD) RNA depletion strategies using (i) dCas9-CRISPR-KRAB-MeCP2 (Yeo, et al. 2018) and (ii) miRNA/SB100X (RNAi) methods (Bunse, et al. 2014). While (i) was used as a proof of concept, (ii) 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 overexpression (OE) cell lines were also generated in human SHEP neuroblastoma cells.

For **knocking out PGBD1**, a CRISPR/Cas9 strategy was used. The pU6-sgRNA-CAGGS-Cas9-Venus-bPA (Addgene #86986) vector was cleaved using BbsI and ligated to the ds oligonucleotides encoding gRNA sequence. The following gRNA sequences were used: Guide-79 sequence in the pU6-sgRNA79 plasmid: 5'-GGC GGC AAA TCT CCT GAG TG-3'; Guide-87 sequence in the pU6-sgRNA87 sequence plasmid: 5'-ATG ACA AAG TTC TCG GAG TT-3'; Guide-91 sequence in the pU6-sgRNA91 sequence plasmid: 5'-AGA TTT GCC GCC TGC GCT TT-3'. The vectors were transfected into hESC_H9 and neural progenitor cells, respectively. 5 days post-transfection, the cells were FACS-sorted for the Venus fluorescence marker gene intensity. The cells were then cultured additional 2 months.

(i) Depletion with dCas9-CRISPR-KRAB-MeCP2

For depletion of PGBD1 level by dCAS9- KRAB-MeCP2 method (Yeo, et al. 2018) we utilized the pCAG dCas9-KRAB-2A-EGFP (Addgene #92396) and pBS700 (Addgene #64046) plasmid vectors. The pBS700 vector was cleaved with the BsmBI endonuclease and ligated to the ds oligonucleotides encoding the gRNA sequence. For constructing the pBS700-dCas-PGBD1-Oli1 (Oli-1) and pBS700-dCas-PGBD1-Oli3 (Oli-3), 5'-AGCTAAGTGAAGCTTTAGCC-3' and 5'-CACTCAGGAGATTTGCCGCC-3' gRNA sequences were used, respectively. The generated constructs were verified by sequencing. To deplete PGBD1, 6×10^5 smNPC (p27) cells were transfected by pCAGdCas9-KRAB-2A-EGFP and Oli-1 / Oli-3 plasmids (1-1 μ g), using Lipofectamine-3000, following the recommended protocol. For the control, the 6×10^5 smNPC (p27) cells were transfected by mixture of pCAGdCas9-KRAB-2A-EGFP and pBS700 (with a 21 bp scrambled sequence) plasmid vectors (1-1 μ g). After 3 days post-transfection, the cells were washed twice with DPBS and collected by centrifugation. **The efficacy of the depletion was determined both on RNA (qPCR) and protein levels (fig. 4B, supplementary fig. 9C and supplementary table S4).** To

determine the protein levels (total, PGBD1 and actin), we utilized the densitometry analysis software of BioRad-ChemiDoc™ MP Imaging System (<https://www.bio-rad.com/de-de/product/chemidoc-mp-imaging-system?ID=NINJ8ZE8Z>). This approach was used to determine the total protein amounts and the PGBD1 and actin levels following Western blotting. The relative KD efficiency was calculated as a ratio of PGBD1 and actin protein levels. *Note that the dCas9-CRISPR-KRAB-McCP2 method was used as a proof of concept to deplete PGBD1. Following transfection, the cells were not further enriched or cultured. We used the miRNA/SB100X (RNAi) strategy to generate stable KD clones.*

(ii) Generating stable KD PGBD1 clones using the miRNA/SB100X strategy

For the KD, 300,000 human NPCs were transfected by different pmiRNA-mCherry constructs together with pSB100X plasmid, encoding SB100X transposase (Mates, et al. 2009) in ratio 10:1. To enrich for the presence of the pmiRNA-mCherry, 3 days post-transfection the cells were FACS sorted for the mCherry marker. The sorted single cells were cultured for an additional three weeks, and subjected to another round of FACS sorting. Using the strategy, five different stable KD PGBD1 clones were established (smNPC-miRNA621/622/623/624/625). The control KD clone (smNPC-miRNA620/scrambled sequence) was generated by a similar strategy. PGBD1 protein levels were determined by Western blotting from the cell lysates of the stable individual clones.

For the **transcriptome analysis**, the KD-PGBD1 NPC sample was generated as follows. To generate replicates, stable smNPC-miRNA621 cells (6×10^5) were transfected with a combination of miRNA constructs of 623/624/625 (1-1-1 μ g) (in 3 independent replicates, KD1/2/3-PGBD1). These cells were estimated to have around $\sim 65\%$ depletion of PGBD1 on protein level (supplementary fig. S8B). For the control (KD-Scr), the stable scrambled smNPC-miRNA620 cells (6×10^5) were transfected with miRNA620 scrambled plasmid in 3 replicates (KD1/2/3-Scr).

To **overexpress (OE) PGBD1**, NPC cells were transfected by 3 μ g pTR HA-PGBD1 plasmid in 2 replicates. 3 days post-transfection, the cells were washed by DPBS. For one sample, cells from 2 wells of the plate were collected by centrifugation at 10,000 rpm for 2 min generating (3 replicates). To determine the PGBD1 level, 1/4 part of the collected cells (per samples) was tested by Western blotting. To generate over-expression (OE) HA-PGBD1 samples, NPCs were transfected by using pTR-HA-PGBD1 plasmid.

For RNA-sequencing, total RNA was extracted from the (KD/OE) PGBD1 cells by using the Trizol kit (Invitrogen) following the manufacturer's instructions. The sequencing libraries were prepared by using TruSeq RNA Indexes Set A, (Illumina 20020492) and TruSeq®Stranded mRNA LT kit 48 Samples (Illumina, 20020594), according to manufacturer's instructions. High throughput single-indexed, paired-end 150 bp sequencing was performed in a HiSeq4000 instrument (Illumina, USA).

For the **transposon excision assay**, mPB is a plasmid vector expressing the mammalian codon optimized version of the insect piggyBac transposase, kindly provided by Allan Bradley (Wellcome Trust Sanger Institute, UK) (*repository number: pCyL43*; (Wang, et al. 2008) *PB-PGK-Puro vector*: The pCyL50 PB cloning vector (*from Allan Bradley, Wellcome Trust Sanger Institute, UK*; (Wang, et al. 2008) was used to insert the human phosphoglycerate-kinase (PGK) promoter driven puromycin resistance gene unit between the PB inverted repeat sequences. The transcription unit was PCR-amplified and ligated into the PB transposon cloning vector. *PGBD1 expression vector (pSB-cmv-GFP-cag-HUCEP4)*: The coding sequence of the human PGBD1 fusion gene was PCR-amplified from cDNA prepared from HEK-293 total RNA. The primer sequences: PGBD1-forward: 5'-ggctagcgcgcacatgtatgaagctttgccaggccctg-3' (**NheI**-Kozak sequence-PGBD1 coding); PGBD1-reverse: 5'-tgcagatctctaactctgacagatgagcattgtg -3' (**Bgl II**-PGBD1 coding). After **NheI** /**Bgl II** digestion, the PCR product was ligated into an expression vector after a CAG promoter. The vector contains a separate CMV-GFP expression cassette for monitoring transfection efficiency. In addition, it also has the *Sleeping Beauty* (SB) inverted terminal repeat sequences in appropriate positions, in order to allow for establishing stable cell lines using the SB100X transposon system.

MER75B-puro / MER85-Puro vectors

To select potentially the best MER sequences, the most important criteria were the high similarity to the consensus sequences and the presence of intact inverted terminal repeats (ITRs) at both 5' and 3' ends (Sarkar, et al. 2003). The selected human MERs with their short flanking genomic sequences were PCR-amplified from HEK293 gDNA preparations, using the following primers: MER75B-cloning-For: 5'- GGATCCTTTCCTTCACCCTCCCTGC -3'

MER75B-cloning-Rev: 5'- GGATCCAGTCACCCCAAGGAGAAAAGG -3' (*from human chromosome 4, Sequence ID: NC_000004.12, nt: 143088382 to 143088750*);

MER85-cloning-For: 5'- GGATCCTTAACACATCAGACATGGAGGG -3';

MER85-cloning-Rev: 5'- GGATCCGGCCAAAGCATATGTTCTTAATC -3' (*from human chromosome 21, Sequence ID: NC_000021.9, nt: 34685694 to 34686125*)

The MER75 and MER85 sequences (verified by Sanger sequencing) were cloned into the pGEM®-TVectorSystems, then BsmAI and HindIII restriction sites were used to insert a PGK-Puro-polyA expression cassette into the MER75B and MER85 sequences, respectively. In the formed expression constructs, the ITRs at both ends of the MER sequences remained intact, allowing the potentially active transposase excising the expression cassette as a transposon. Note that MER75 is a predicted substrate of PGBD4, whereas MER85 is a predicted substrate of PGBD3 (Pace and Feschotte 2007; Pace, et al. 2008).

Excision repair reporter assay

Human HeLa cells were seeded in 12-well cell culture plates (100,000 cells/well). Transfections were performed using plasmids encoding HA-tagged PGBD expression construct (500 ng each) together with 500ng pEXC-mPB plasmid (in 1:1 ratio). 48 hours post-transfection cells were washed three times by DPBS and treated by trypsin for 2 min. After treatment 1 ml DPBS supplemented by 10% fetal serum was added to the cells and filtered through FALCON 5 ml Polystyrene round-bottom tube with cell-strainer cap. 60,000 cells GFP intensity was then measure by FACS Calibur™ System.

Cell lines and transfection conditions for transposon excision assay

Human embryonic kidney cells (HEK293) cells were cultured in Dulbecco's modified Eagle's medium supplemented with 10% of fetal calf serum, 1% of L-glutamine, and 1% of penicillin–streptomycin (Thermo Fisher Scientific). Cell were transfected using the FuGENE 6 reagent (BioScience Ltd., Hungary), according to the manufacturer's instruction. Briefly, 4×10^5 cells were seeded onto 6-well plates and 1 day later, cell were transfected with the lipid-DNA mixture as specified by the protocol, using 500 ng of total DNA, composed of 250 ng transposon containing plasmid mixed with 250 ng of a transposase/putative transposase/control expression vector. Transfection efficiency was checked by fluorescence microscopy detecting GFP signal where appropriate, and quantified by FACS measurements using as FACSCanto instrument (BD Biosciences).

Transposon excision assay

To detect transposon excision events from the donor plasmids, plasmid DNA was isolated from the transfected cells 48 hours post-transfection, using standard phenol/chloroform extraction method, followed by ethanol precipitation. To detect those plasmids that were re-circularized following transposon excision, followed by cellular DNA repair, the extracted plasmids were subjected to a

nested PCR in 2 rounds. Primers used for the assay are as follows: Excision, first round, forward: 5'-GCGAAAGGGGGATGTGCTGCAAGG -3'; Excision, first round, reverse: 5'-TCTTTTCCTGCGTTATCCCCCTGATTC -3'; Excision, second round, forward: 5'-CGATTAAAGTTGGGTAACGCCAGGG -3'; Excision, second round, reverse: 5'-CAGCTGGCACGACAGGTTTCCCG -3'. For normalization, PCR on the plasmid backbone (on the ampicillin resistance gene sequence) was carried out, resulting in a product, regardless of transposon excision. Excision, plasmid control, Amp64-Forward: 5'-TTTGCTCACCCAGAAACGC -3'; Excision, plasmid control, Amp403-Reverse: 5'-AGTTGGCCGCAGTGTATCAC -3'

Colony forming assay to detect stable integration

To detect stable integration after transposition, puromycin resistance gene expressing transposons were used for antibiotic selection experiments (Ivics and Izsvak 1997). After 48 hours post-transfection, 1% of transfected cells were seeded onto cell culture Petri dishes, selected for 2-3 weeks with 1 µg/ml of puromycin (Sigma-Aldrich), and surviving cells were fixed with methanol and stained with Giemsa (Sigma-Aldrich). Colonies were quantified in a 75S model gel imager, using the Quantity One 4.4.0 software (Bio-Rad).

Generation of stable PGBD1 knockdown and overexpression SHEP neuroblastoma cell lines

Note that to deplete PGBD1 in hESCs, first, we used a CRISPR/Cas9-mediated knocking out (KO) approach. However, using the KO strategy, no stable, proliferative KO line could be generated. This KO approach has also failed in neural progenitor cells (NPC), suggesting an essential function of PGBD1 in cell survival. As an alternative, we used a knock down (KD) RNA depletion strategy. To KD PGBD1, 300,000 human SHEP neuroblastoma cells were co-transfected using different miRNA constructs (marked by mCherry), pTOV-T11-TR-HA-PGBD1 together with pSB100X (Mates, et al. 2009) plasmid encoding the SB100X transposase in ratio 10:1. 3 days post-transfection, the cells were FACS-sorted for mCherry marker gene intensity. The sorted single cells were cultured for three weeks then subjected to an additional FACS sorting. To generate overproducing HA-PGBD1 stable cell lines, SHEP cells were transfected using pTOV-T11-TR-HA-PGBD1 inducible expression construct together with pSB100X plasmid encoding the SB100X transposase in ratio 10:1. The transfected cells were cultured for four weeks in Dubelcco's modified Eagle's medium (DMEM) with 4.5 g/l D-glucose containing 10% fetal bovine serum, supplemented with G418 (gentamicin) in final concentration 500 µg/ml. PGBD1 protein level was determined using Western-blotting. To induce HA-PGBD1 protein expression, the media was supplemented by 1µg/ml doxycycline in final concentration.

Cell culture and cell transfection for additional cell lines

Human cervical carcinoma cells (HeLa) and human neuroblastoma (SHEP) cells were cultured in Dubelcco's modified Eagle's medium (DMEM) with 4.5 g/l D-glucose containing 10% fetal bovine serum, supplemented with penicillin (100 µg/ml) and streptomycin (100 µg/ml) at 37°C and 5% CO₂. In general, the cells were seeded in 10 cm or 6-well plates and transfection was performed using JetPrime™ or *Lipofectamine®-3000* (Thermo Fischer) reagents, and following the recommended manufacturer's protocol.

Stable isotope labelling by amino acids in cell culture (SILAC)

Two populations of HEK293 cells were cultivated in cell culture for three 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 labelled with stable heavy isotopes (13C6-15N4 L-arginine; 13C6-15N2 L-lysine) ("heavy cell population"). In our experimental approach, untagged PGBD1 and HA-tagged PGBD1 were overexpressed in both conditions. The overexpressing plasmids encoding the HA-tagged or untagged PGBD1 were transfected into the two cell populations of HEK293 cells, respectively. 3 days post-transfection protein purification was performed using EZview™ red coloured Anti-HA agarose affinity gel, following the recommendations of the manufacturer. Purified protein mixtures were prepared as follows: (a) as forward experiment: protein mixture of Heavy HA-PGBD1 and protein mixture of Light PGBD1; (b) as reverse experiment: protein mixture of Light HA-PGBD1 and protein mixture of Heavy PGBD1. The purified protein mixtures including HA-tagged target proteins and their interacting partners were subjected to LS-MS/MS (mass spectrometry): Samples were processed by methanol-chloroform extraction, reduced, alkylated and digested with LysC and trypsin using standard protocols. After offline desalting, peptides were analysed by LC-MS/MS on a Proxeon EASY-nLC II system, connected to a Q Exactive mass spectrometer (Thermo Scientific). Chromatography was performed using a 120 min acetonitrile gradient on a 25cm long inhouse prepared column (ReproSil-Pur 120 C18-AQ, 3 µm (Dr. Maisch GmbH HPLC)). The instrument was operated in the data dependent mode with the following settings for the full scans: resolution 70,000, AGC target value 3E6, maximum injection time 20 ms. The following settings were chosen for the MS2 scans: resolution 17,500, AGC target value 1E6, maximum injection time 60 ms. Raw files were processed with MaxQuant (version 1.4.1.2).

SILAC-AP-MS Analysis

Calculation of significance

The significance was calculated as described previously (Cox and Mann 2008) with small modifications. First, we filtered out contaminants and proteins for which no unique peptide was detected. We performed logarithmic transformation (\log_2 with pseudo count 1) of the normalized H/L ratios and confirmed a Gaussian distribution of the transformed data (Shapiro-Wilk test). The 15.87th, 50th and 84.15th percentiles were calculated and called r_{-1} , r_0 and r_1 , respectively. The z -transformation is defined as:

$$z = \frac{r - r_0}{r_1 - r_0} \quad \text{for } r > r_0 \quad \text{and} \quad z = \frac{r - r_0}{r_0 - r_{-1}} \quad \text{for } r < r_0, \text{ where } r \text{ are the transformed H/L ratios.}$$

The p -values were calculated with significance A formula. In the original paper significance A gives two-sided p -values. Here we considered the sidedness of the test and adjusted one side of the distribution (one-sided):

$$\text{significance A} = \frac{1}{2} \operatorname{erfc}\left(\frac{z}{\sqrt{2}}\right) \quad \text{for } z \geq 0,$$

$$\text{and significance A} = 1 - \left(\frac{1}{2} \operatorname{erfc}\left(\frac{z}{\sqrt{2}}\right)\right) \quad \text{for } z < 0.$$

The sidedness of the label-switch experiment is reverse. Which means that $z \geq 0$ was adjusted for sidedness. The p -values were adjusted for multiple testing with the Benjamini-Hochberg method. We only considered proteins as significant if they were also significant outliers in the label switch experiment (intersection). P -values of both experiments were combined with the berger method of the *scraper* R package (Lun, et al. 2016). The script was written in Rstudio (R version 3.6.3), using *erfc* function from the *pracma* package.

Gene set enrichment analysis

All gene sets of the Molecular Signature Database v7.2 (Subramanian, et al. 2005; Liberzon, et al. 2011) were downloaded. Terms with at least 5 gene members and less than 500 were tested for enrichment. The enrichment was tested with a ranked gene list approach, the „CERNO“ algorithm from the *tmod* R package (Weiner 3rd and Domaszewska 2016). Some of the unique peptides map to multiple genes and the true origin remains unknown, for the analysis all genes have been considered.

Immunofluorescence microscopy

The cells were seeded on coverslips in 12-well cell culture plates (100,000 cells/well). 48 h 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.

RNA-FISH and immunofluorescence

Cells grown on coverslips were fixed using 4% paraformaldehyde and permeabilized with 70% ethanol overnight. For RNA-FISH, Stellaris RNA-FISH probes labelled with Quasar 570 Dye for NEAT1_2 (SMF-2037-1) (1:100, Biosearch Technologies) were used according to the instructions provided. For subsequent immunofluorescence staining, SFPQ antibody (1:60, WH0006421M2, Sigma) and PGBD1 antibody (1:200, Abcam, ab180598) were used. Finally, cells were counterstained with DAPI (4',6-diamidino-2-phenylindole) in water for 15 min at room temperature. After an additional washing steps, the samples were mounted using ProLong® Gold antifade reagent (Invitrogen). The images were taken using a Leica LSM710 point-scanning single photon confocal microscope. Paraspeckles were defined as NEAT1_2 RNA-FISH signals that are colocalizing with SFPQ.

Co-immunoprecipitation

In general, 500.000 non-transfected or transfected of human cells were lysed in 200 µl lysis buffer containing 50 mM TRIS HCl pH 8.0, 10 mM EDTA, 100 mM NaCl, 5% glycerol, 1% NP-40 supplemented with COMPLETE protease inhibitor cocktail (Roche) and Benzonase®Nuclease (Sigma) for 40 mins at 4°C. After incubation, the lysates were centrifuged for 10 min at 12,000 rpm at 4°C to remove unbroken cells and cell debris. Supernatant was collected and protein concentration was determined. In parallel, 60 µl Dynabeads™ ProteinG was washed as recommended by the manufacturer, and resuspended in 200 µl PBS+0,02 % Tween20. The mixture was incubated with 5 µg of primary antibodies for 1 h at room temperature on a turning wheel. After incubation, the magnebeads were collected and washed with 200 µl PBS+0,02 % Tween20, then 150-200 µg of cell lysates was added to the beads. Cell lysates together with the magnebeads were incubated for 3 hours at 4°C on a turning wheel. The beads were washed three times with 200 µl PBS+0,02 % Tween20. The proteins were eluted from the beads by adding a mixture of 20 µl 100 mM glycine pH 2,8 and 10

μl of 5 x SDS-loading dye (10% SDS, 10 mM DTT, 20 % glycerol, 0.2 M Tris-HCl pH 6.8, 0.05% Bromophenolblue) and boiling at 70°C for 5 min.

Western blotting

Cells were collected from 6-well plates for Western blot analysis. The cells were subsequently washed with phosphate-buffered saline (PBS), and lysed on ice for 40 min in lysis buffer containing 50 mM TRIS HCl pH 8.0, 10 mM EDTA, 100 mM NaCl, 5% glycerol, 1% NP-40 and Protease Inhibitor Mini Tablets, EDTA Free (Pierce™) and Benzonase Nuclease (NOVAGEN) as recommended by the manufacturer. Total lysates were resolved by 8-12 or 4-20 % Mini-Protean® TGX™ SDS-PAGE Gel, and transferred to PVDF membranes by Bio-Rad Trans-Blot®Turbo™ Transfer System. The membranes were blocked with TBS-T (TBS supplemented with 0.05% Tween-20) containing 5% non-fat dry milk and were incubated overnight at 4°C with primary antibodies in appropriate dilutions in TBS-T containing 5% non-fat dry milk. Membranes were washed with TBS-T buffer, incubated for 1 hour at room temperature with Alkaline Phosphatase-conjugated (Sigma-Aldrich) or Horseradish Peroxidase-Conjugated (Promega) secondary antibodies in TBS-T containing 5% non-fat dry milk. The blots were subsequently washed with TBS-T. Bands were detected by ECL™ Prime Western Blotting Detection Reagent (Amersham) and images were then analysed by Chemi Doc™MP Imaging System (Bio-Rad).

Antibodies and chemicals

Anti-PGBD1 antibody (1:1000, Abcam, ab180598), anti-HA monoclonal antibody (1:2000, Roche), monoclonal anti- α -actinin (1:1000, Sigma, A7811), aSFPQ (WH0006421M2, Sigma). Green fluorescence goat anti-rabbit antibody Alexa Flour® 488 (1:200, Invitrogen), green fluorescence donkey anti-rabbit antibody Alexa Flour® 488 (1:200, Invitrogen), red fluorescence goat anti-rat antibody Alexa Flour® 568 (1:200, Life Technology), red fluorescence donkey anti-mouse antibody Alexa Flour® 555 (1:200, Invitrogen), goat anti-rabbit IgG (1:5000, Thermo scientific, 31462), goat anti-mouse IgG (1:5000, Thermo scientific, 31432), goat anti-rat IgG (1:5000, Thermo scientific, 31470), red fluorescence donkey anti-mouse antibody Alexa Flour® 647 (1:200, Life Technology).

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 following the manufacturer's protocol.

ChIP-seq for histone tail modifications - peak calling

MACSv2 (Model-based Analysis for ChIP-seq) ([Zhang, et al. 2008](#)) was utilized for the detection/analysis of genome-wide broad peaks representing histone tail modifications. Replicates were pooled separately for each histone modification. Peaks were called with input/mock DNA samples for identification of unspecific signals. Candidate peaks were selected according to the threshold values: q-value ≤ 0.01 and mfold = 10,100 (default 5, 10). The mfold parameter selects only those regions that are mfold or higher enriched for ChIP-seq reads compared to a random genome-wide distribution (fold enrichment for the peak summit against random Poisson distribution computed with the local lambda). Consensus peaks between biological replicates were calculated with DiffBind (v 2.10.0). Histone modification data for NPC and neurons were obtained from GSE119006 and GSE62193. Active promoters were defined as (H3K4me3+H3K27ac+H3K36me3), repressed promoters as (H3K4me3+H3K27me3-H3K27ac-H3K36me3) and poised enhancers as (H3K4me1+H3K27me3). The overlaps were calculated with BedTools intersectBed command with `-f 0.5 -r`. Coverage of the BAM files representing histones was calculated for 1KB upstream (promoter regions) of genes using deepTools ([Ramirez, et al. 2014](#)).

RNA-sequencing data analysis

Raw reads filtering

Software tools such as the FASTX-Toolkit and Trimmomatic ([Bolger, et al. 2014](#)) were used to discard low-quality reads, trim adaptor sequences, and eliminate poor-quality bases. Outliers with over 30 % disagreement were discarded.

Read alignment

Salmon ([Patro, et al. 2017](#)) was used to build index and align the reads using the following commands: `salmon index -t transcripts.fa -i transcripts_index --decoys decoys.txt -k 31, ./bin/salmon quant -i transcripts_index/ -l IU -1 fastq -2 fastq --validateMappings -o output`. Quantified data was checked for GC content and gene length biases using R package NOISeq ([Tarazona, et al. 2015](#)) to provide useful plots for quality control of count data.

Reproducibility

Mean variance and PCA were computed between biological replicates using the tximport package (Soneson, et al. 2015) in R using lengthscaledTPM (CPM cutoff >2 and sample cutoff 2 between the replicates) for the analyzed groups (NPC-KD, KD-Neuron). Batch effects were removed using the RUV package of Bioconductor (Risso, et al. 2014). Subsequently, the samples were normalized using TMM (weighted trimmed mean of M-values) method for differential expression analysis. The differential expression analysis was conducted using the DESeq2 package (Love, et al. 2014). Dataset for the analysis of schizophrenia and control was obtained from GSE145656 and analyzed as above.

ChIP-exo data analysis for PGBD1 peak calling

To map the sequencing reads, the FASTQ files were aligned to human genome hg19 using Bowtie2. To filter out PCR duplicates, we used Samtools v 1.9 (Li, et al. 2009) q in the following steps: (i) filter out low-quality (<20) reads; (ii) sort reads by name (sort -n); (iii) fix read pairs (fixmate); (iv) sort reads by chromosomal position (sort); (v) mark duplicates (markdup); (vi) extract only read_1 from each non-duplicated pair (view -f 0x40). As the majority of TFs bind as dimers (either homo or hetero), it was necessary to determine the optimal trim length. Therefore, we took 3 times the radius of the PGBD1 and converted that size into number of bps. After rounding, this resulted in the optimal trim length in bp. BEDTools v 2.30.0 (Quinlan and Hall 2010) “genome coverage” function was employed to generate the read profiles for both strands separately using the determined optimal trim length. Subsequently, both strands were combined and only base positions where there are reads on both strands were reported as the PGBD1 binding profile. After this step, the replicates were normalized based on their average background read count and then combined using their average read count per base position. GEMv3.4 (Guo, et al. 2012) was used to detect narrow peaks using the following command:

```
--d Read_Distribution_ChIP-exo.txt --expt PGBD1.bam --ctrl PGBD1.bam --g hg19.chrom.sizes --f BAM --genome hg/ --k_min 6 --k_max 18 --outBED --outNP --smooth 3 --mrc 20 -q 3
```

Transposable element identification and enrichment in PGBD1 peaks

The overlap of PGBD1 peaks with Repetitive elements (RepBase annotation) (Jurka, et al. 2005) was performed with BedTools intersectBed command (-f 0.1 -r). The enrichment of a particular TE class was calculated as following- 200 bp DNA-wide window was extracted from the center of the PGBD1 peak using Bedtools slop -b -l 200 -r 200 command (Quinlan and Hall 2010). Dataset of 10000 randomly chosen peaks, containing the same number of regions with the same length (200 bps) and the same nucleotide distribution as true regions was generated using the BedTools random

command. To do this, each chromosome was divided in genomic windows of 1,000,000 bps and, for every real peak, the corresponding random peak was taken with flat distribution from the same window. All repeats falling in the peaks belonging to the real dataset and in the 10000 datasets of random peaks were annotated. This method delivered, for each transposable element, mean and variance, which was used to calculate a z-score $z_r = (x_r - \mu_r) / \sqrt{s_r}$, where x_r is the occurrence of a particular transposable element r in the original dataset, while μ_r and s_r are respectively its mean occurrence and its variance in the 10000 random sets. Similar analysis was performed with classes of transposable elements and the z-score $z_c = (x_c - \mu_c) / \sqrt{s_c}$, where x_c is the occurrence of a particular class of transposable elements in the original dataset, while μ_c and s_c are respectively its mean occurrence and its variance in the 10000 random sets.

ATAC-seq data analysis in PGBD1 peak regions genome-wide

Genrich tool (<https://github.com/jsh58/Genrich>) was utilized for detection of peaks in the PGBD1 binding sites. Genrich calls peaks for multiple replicates collectively. First, it analyzes the replicates separately, with p-values calculated for each. At each genomic position, the multiple replicates' p-values are then combined by Fisher's method. The combined p-values are converted to q-values, and peaks are called. FASTQ data were processed for alignment parsing, removal of multimapping reads, PCR duplicate removal, Genome length calculation, Control/background pileup calculation and p-value calculation using the following command:

```
./Genrich -t ATACseq.bam -o ATACseq.narrowPeak -f ATACseq.log -r -x -q 0.05 -a 20.0 -v -c chrM,chrY -E hg19_Ns.bed,wgEncodeDukeMapabilityRegionsExcludable.bed.gz  
./Genrich -P -f ATACseq.log -o peaks.narrowPeak -p 0.01 -a 200 -v Peak-calling from log file: ATACseq.log.
```

GRO-Seq data analysis

GRO-seq data (sra format, GSE140486) were processed into the FASTQ format with the 'fastqdump' command (SRA toolkit) (Leinonen, et al. 2011). The resulting cDNAs were trimmed with Homer v 4.10 to remove 3' terminal A-stretches, which had been attached during library construction (homerTools trim) (Heinz, et al. 2010). Only cDNAs ≥ 25 bp entered the analysis. Datasets were quality filtered with the FASTX (v 0.0.13) software tool (-q 10 -p 97) (http://hannonlab.cshl.edu/fastx_toolkit/), and resulting GRO-seq cDNAs were aligned to the human genome assembly (hg19) using Bowtie version 0.12.9 (-v 2 -k 3 -m 1 -best). BAM files were utilized to calculate GRO-seq peaks using the annotatePeaks function from HOMER (v 4.10).

Protein coding genes were divided into 2 groups based on the number of PGBD1 peaks occurring in the entire gene body (1-2 vs. 3-7). Pausing using index for RNAPII in the dataset was calculated as follows: $S = \log_2(d(\text{RNAPII PGBD1 binding sites})) - \log_2(d(\text{RNAPII gene body}))$. Ratio of RNAPII read density at the PGBD1 binding sites within protein coding genes to that of the RNAPII read density in the gene body. d stands for the number of reads per nucleotide (nt) in the given region. The difference between the densities in log2 units equals the ratio of fold enrichment in these regions, meaning a value of 1 would represent a 2-fold greater enrichment of RNAPII signal at the promoter region rather than in the gene body (Muse, et al. 2007).

Gel-shift experiment with SCAN-12 motif

Approximately 1.1×10^6 HEK293 cells were transfected with 12 μg pTR -HA-PGBD1 plasmids encoding HA-PGBD1 fusion protein. Two days post-transfection, cells were collected and washed with DPBS. Cells were lysed in 500 μl lysis buffer (50mM Tris-HCl, pH8.0, 100mM NaCl, 10mM EDTA, 5% glycerine, 1% NP-40 and protease inhibitor cocktail (Roche)) for 40 min at 4°C. Following removal of the cell debris by centrifugation at 20,000 g, HA-PGBD1 protein was purified by using EZview™ Red Anti-HA Agarose beads. The HA-tagged PGBD1 was eluted by 500 μl 100 $\mu\text{g/ml}$ HA-peptide (Sigma) in 100 mM NaCl. The protein mixture was concentrated by Amicon Ultra-4, Ultracel -3K filter column. Centrifugation was performed at 7500 rpm for 40 min at 4°C. The protein mixture concentration was measured by standard BCA method and the level purity was analysed by loading 5 μl onto 10% SDS-PAGE. Binding reactions were performed in 25 μl volumes on ice for 20 min. DNA binding reactions contained FAM-labelled PGBD1-specific (SCAN-12-GS-PGBD1-UP: 5'-GCTTTCAATGGAATGGAATGCCITTCC-3'), complementary (SCAN-12-GS-PGBD1-LOW: 3'-CGAAAGTTACCTTACCTTAGGGAAGG-5') dsDNA oligonucleotides (PGBD1 oligo) 5 μM , purified HA-PGBD1 protein, 10 mM Tris-HCl pH 8.5, poly(dI-dC), 1 mM EDTA, 50 mM KCl, 10 mM 2-mercaptoethanol. As competitor dsDNA oligonucleotide we used the PGBD1-specific oligonucleotide without FAM labelling in concentration of 5 μM (+), 10 μM (++) (Fig2E middle left panel), 25 μM (+) and 50 μM (++) . The samples were loaded onto the gel after addition of 5 μl 86% glycerine. The gel buffer contained 25 mM Tris-borate pH8.3, 1 mM EDTA. Protein-DNA complexes were separated by electrophoresis in 6% non-denaturing polyacrylamide gels at 4°C. Electrophoresis was performed at constant voltage of 200V for 3 h. The fluorescent signal was detected by using a BioRad ChemiDoc™ MP Imaging System.

Gel-shift experiment with Motif1 and Motif2

Approximately 20×10^6 HEK293 cells were transfected with 20 μg pTR-HA-PGBD1 plasmids encoding HA-PGBD1 fusion protein. Two days post-transfection, cells were collected and washed with DPBS. The HA-PGBD1 protein was purified using the HA tagged Protein PURIFICATION KIT-BioZol (MBL-3320) following the manufacturer recommendation. The HA-tagged PGBD1 was eluted using 120 μl 2 $\mu\text{g}/\mu\text{l}$ HA-peptide. The protein concentration was measured by standard BCA method and the level of the purity was analysed by loading 6 μl onto 10% SDS-PAGE.

Binding reactions were performed in 25 μl volumes on ice for 20 min. DNA binding reactions contained FAM-labelled PGBD1-specific (motif-1-UP: 5'-ACCTCTCTTCACAGACTCAAATGACTCCAG-3'), complementary (motif-1-LOW: 3'-TGGAGAGAAGTGTCTGAGTTTACTGAGGTC-5'), PGBD1-specific (motif-2-UP: 5'-GCCGGTGGCCGTGGAGGAATCGTCCCGTTGAGCAAT-3'), complementary (motif-2-LOW: 3'-CGGCCACCGGCACCTCCTTAGCAGGGCAACTCGTTA-5') dsDNA oligonucleotides in 5 μM concentration, purified HA-PGBD1 protein and poly(dI-dC) as unspecific competitor. As a specific competitor we used the PGBD1-specific motif-1 dsDNA oligonucleotide without FAM labelling in equimolar concentration (5 μM). The samples were loaded onto the gel after addition of 5 μl 86% glycerine. The gel buffer contained 25 mM Tris-borate pH8.3, 1 mM EDTA. Protein–DNA complexes were separated by electrophoresis in 6% non-denaturing polyacrylamide gels at 4°C. Electrophoresis was performed at constant voltage of 200V for 3 h. The fluorescent signal was detected by using a BioRad ChemiDoc™ MP Imaging System.

Generation of NPCs and neurons from hESCs

We used the human embryonic stem cell (hESC) line H1 according to the German law under a license approved by the Robert Koch Institute (license to A. Prigione # AZ: 3.04.02/0077-E01). The derivation of NPCs was the same as reported ([Lorenz, et al. 2017](#)). We performed the generation of midbrain dopaminergic neurons (mDAN) following a previously published protocol ([Reinhardt, et al. 2013](#)). We first let the NPCs to differentiate onto Matrigel coated plates for 8 days using a medium containing: Neurobasal: DMEM/F12 (1:1), N2 (1x), B27 (1x), purmorphamine (1 μM), vitamin C (200 μM), and FGF8 (100 ng/ml). Afterwards, we cultured the cells for two additional days using a medium containing: Neurobasal: DMEM/F12 (1:1), N2 (1x), B27 (1x), purmorphamine (500 nM), and vitamin C (200 μM). We next split the cells using Accutase at 1:3 ratios and plated

them onto matrigel-coated dishes. Finally, we switched the culture conditions to the maturation medium containing: Neurobasal/DMEM-F12 (1:1), N2 (1x), B27 (1x), vitamin C (200 μ M), db-cAMP (500 μ M), BDNF (10 ng/ml), GDNF (10 ng/ml), and TGF β 3 (1 ng/ml). We kept the differentiating neurons in the maturation medium for eight weeks with medium changed every other day. We kept NPCs and neuronal cultures in a humidified atmosphere of 5% CO₂ at 37°C under atmospheric oxygen condition and regularly monitored them against mycoplasma contamination.

qRT-PCR validation of selected DEGs in PGBD1-depleted NPC

Total RNA was extracted from cells by using the Direct- zol™RNA MiniPrep kit following manufacturer's instructions. 1 μ g purified DNaseI-treated RNA was used for reverse transcription (RT) by High Capacity RNA-to-cDNA kit (Applied Biosystems). 1 μ g purified DNaseI-treated RNA was used for reverse transcription (RT) (High Capacity RNA-to-cDNA kit, Applied Biosystems). Quantitative RT-PCR (qRT-PCR) was performed using the Power SYBR Green PCR Master Mix (Applied Biosystems) on the ABI7900HT sequence detector (Applied Biosystems). Data were normalized to GAPDH expression using the $\Delta\Delta$ Ct method. Error bars represent the standard deviation (s.d.) of samples carried out in triplicates.

Sequence of the primers used in qPCR:

NEAT1-qFwd: 5'-TGTGTTCCAGAGCCCATGAT-3'

NEAT1-qRev: 5'-TGAAAACCTTTACCCCAGGA-3'

NEAT1_2-qFwd: 5'-GATCTTTTCCACCCCAAGAGTACATAA-3'

NEAT1_2-qRev: 5'-CTCACACAAACACAGATTCCACAAC-3'

MSI1-forward2: 5'-GTCACCTTCATGGACCAGGC-3'

MSI1-reverse2: 5'-CCGGTTGGTGGTTTGTCAA-3'

NEUROD1-forward1: 5'-GAGACGCATGAAGGCTAACG-3'

NEUROD1-reverse1: 5'-CTGAACGAAGGAGACCAGGT-3'

PAX3-forward2: 5'-GGCATGTTTCTAGCTGGGAAAT-3'

PAX3-reverse2: 5'-TGCTGTGTTTGGCCTTCTTC-3'

PAX6-forward1: 5'-ACCGGTTTCTCCTTCACAT-3'

PAX6-reverse1: 5'-GGAGTATGAGGAGGTCTGGC-3'

GAP43-forward2: 5'-CTCATAAGGCCGCAACCAAA-3'

GAP43-reverse2: 5'-GGTGCCTTCTCCCTTCTTCT-3'

ZNF24-forward1: 5'-ATGTCTGCACAGTCAGTGGGAAGAAGATTCA-3',
ZNF24-reverse1: 5'-CGGAAAATCTCTGGGTCTGGGAGATGGTTC-3'
GAPDH-forward1: 5'-CTTGTGGTATCGTGGGAAGGACTC-3'
GAPDH-reverse1: 5'-CTCTTCCTCTTGTGCTCTTGCT-3'

Statistics

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

REFERENCES

- Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 30:2114-2120.
- 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.
- Chung KH, Hart CC, Al-Bassam S, Avery A, Taylor J, Patel PD, Vojtek AB, Turner DL. 2006. Polycistronic RNA polymerase II expression vectors for RNA interference based on BIC/miR-155. *Nucleic Acids Res* 34:e53.
- Cox J, Mann M. 2008. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat Biotechnol* 26:1367-1372.
- Cui Z, Geurts AM, Liu G, Kaufman CD, Hackett PB. 2002. Structure-function analysis of the inverted terminal repeats of the sleeping beauty transposon. *J Mol Biol* 318:1221-1235.
- Engels B, Cam H, Schuler T, Indraccolo S, Gladow M, Baum C, Blankenstein T, Uckert W. 2003. Retroviral vectors for high-level transgene expression in T lymphocytes. *Hum Gene Ther* 14:1155-1168.
- Guo Y, Mahony S, Gifford DK. 2012. High resolution genome wide binding event finding and motif discovery reveals transcription factor spatial binding constraints. *PLoS Comput Biol* 8:e1002638.
- Heinz N, Schambach A, Galla M, Maetzig T, Baum C, Loew R, Schiedlmeier B. 2011. Retroviral and transposon-based tet-regulated all-in-one vectors with reduced background expression and improved dynamic range. *Hum Gene Ther* 22:166-176.
- Heinz S, Benner C, Spann N, Bertolino E, Lin YC, Laslo P, Cheng JX, Murre C, Singh H, Glass CK. 2010. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol Cell* 38:576-589.
- Ivics Z, Izsvak Z. 1997. Family of plasmid vectors for the expression of beta-galactosidase fusion proteins in eukaryotic cells. *Biotechniques* 22:254-256, 258.

- Jurka J, Kapitonov VV, Pavlicek A, Klonowski P, Kohany O, Walichiewicz J. 2005. Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet Genome Res* 110:462-467.
- Leinonen R, Sugawara H, Shumway M, International Nucleotide Sequence Database C. 2011. The sequence read archive. *Nucleic Acids Res* 39:D19-21.
- Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, Genome Project Data Processing S. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25:2078-2079.
- Liberzon A, Subramanian A, Pinchback R, Thorvaldsdottir H, Tamayo P, Mesirov JP. 2011. Molecular signatures database (MSigDB) 3.0. *Bioinformatics* 27:1739-1740.
- 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.
- Love MI, Huber W, Anders S. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* 15:550.
- Lun AT, McCarthy DJ, Marioni JC. 2016. A step-by-step workflow for low-level analysis of single-cell RNA-seq data with Bioconductor. *F1000Res* 5:2122.
- 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.
- Muse GW, Gilchrist DA, Nechaev S, Shah R, Parker JS, Grissom SF, Zeitlinger J, Adelman K. 2007. RNA polymerase is poised for activation across the genome. *Nat Genet* 39:1507-1511.
- Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C. 2017. Salmon provides fast and bias-aware quantification of transcript expression. *Nat Methods* 14:417-419.
- Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 26:841-842.
- Ramirez F, Dundar F, Diehl S, Gruning BA, Manke T. 2014. deepTools: a flexible platform for exploring deep-sequencing data. *Nucleic Acids Res* 42:W187-191.
- 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.
- Risso D, Ngai J, Speed TP, Dudoit S. 2014. Normalization of RNA-seq data using factor analysis of control genes or samples. *Nat Biotechnol* 32:896-902.
- 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. *Molecular genetics and genomics : MGG* 270:173-180.
- Serandour AA, Brown GD, Cohen JD, Carroll JS. 2013. Development of an Illumina-based ChIP-exonuclease method provides insight into FoxA1-DNA binding properties. *Genome Biol* 14:R147.
- Soneson C, Love MI, Robinson MD. 2015. Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences. *F1000Res* 4:1521.
- Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, et al. 2005. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A* 102:15545-15550.
- Tarazona S, Furio-Tari P, Turra D, Pietro AD, Nueda MJ, Ferrer A, Conesa A. 2015. Data quality aware analysis of differential expression in RNA-seq with NOISeq R/Bioc package. *Nucleic Acids Res* 43:e140.

Wang W, Lin C, Lu D, Ning Z, Cox T, Melvin D, Wang X, Bradley A, Liu P. 2008. Chromosomal transposition of PiggyBac in mouse embryonic stem cells. *Proc Natl Acad Sci U S A* 105:9290-9295.

Weiner 3rd J, Domaszewska T. 2016. T. tmod: an R package for general and multivariate enrichment analysis. *PeerJ Preprints*.

Williams RM, Senanayake U, Artibani M, Taylor G, Wells D, Ahmed AA, Sauka-Spengler T. 2018. Genome and epigenome engineering CRISPR toolkit for in vivo modulation of cis-regulatory interactions and gene expression in the chicken embryo. *Development* 145.

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.

Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, Nusbaum C, Myers RM, Brown M, Li W, et al. 2008. Model-based analysis of ChIP-Seq (MACS). *Genome Biol* 9:R137.