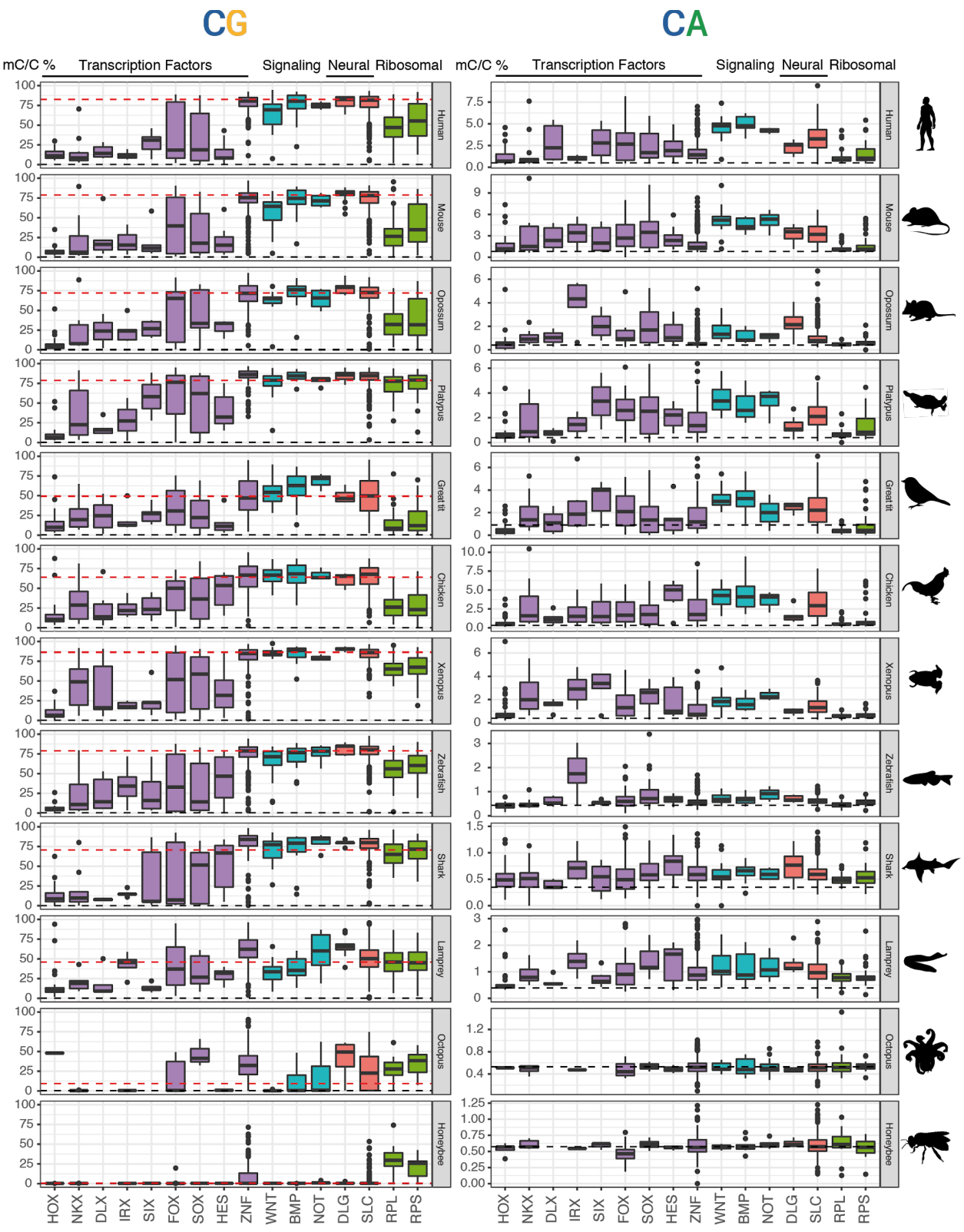
Supplementary material

The emergence of the brain non-CpG methylation system in vertebrates

Alex de Mendoza, Daniel Poppe, Sam Buckberry, Jahnvi Pflueger, Caroline B. Albertin, Tasman Daish, Stephanie Bertrand, Elisa de la Calle Mustienes, Jose Luis Gomez-Skarmeta, Joseph R. Nery, Joseph R. Ecker, Boris Baer, Clifton W. Ragsdale, Frank Grützner, Hector Escriva, Byrappa Venkatesh, Ozren Bogdanovic, Ryan Lister

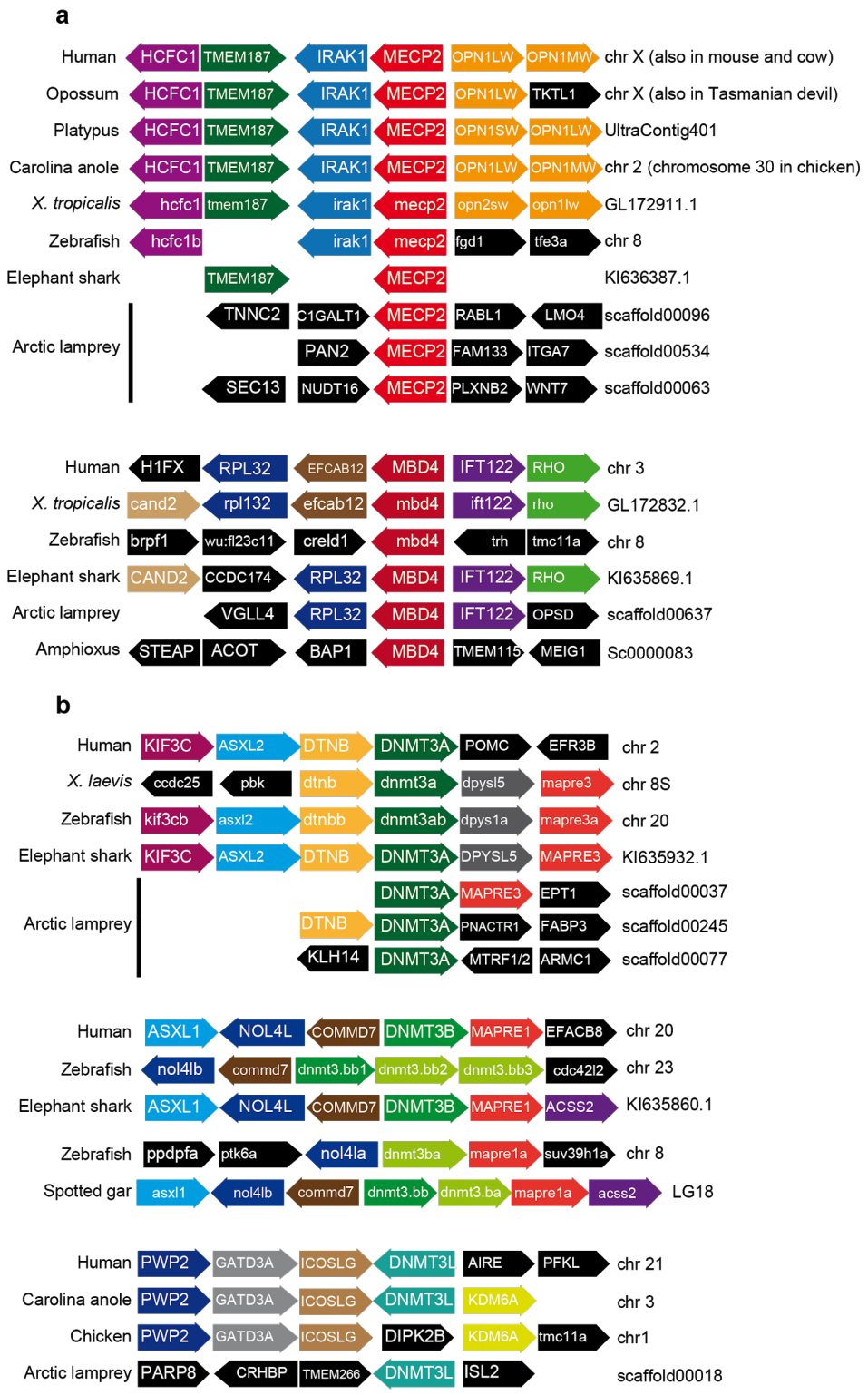
Correspondence to: ryan.lister@uwa.edu.au

Supplementary Figures 1-3.

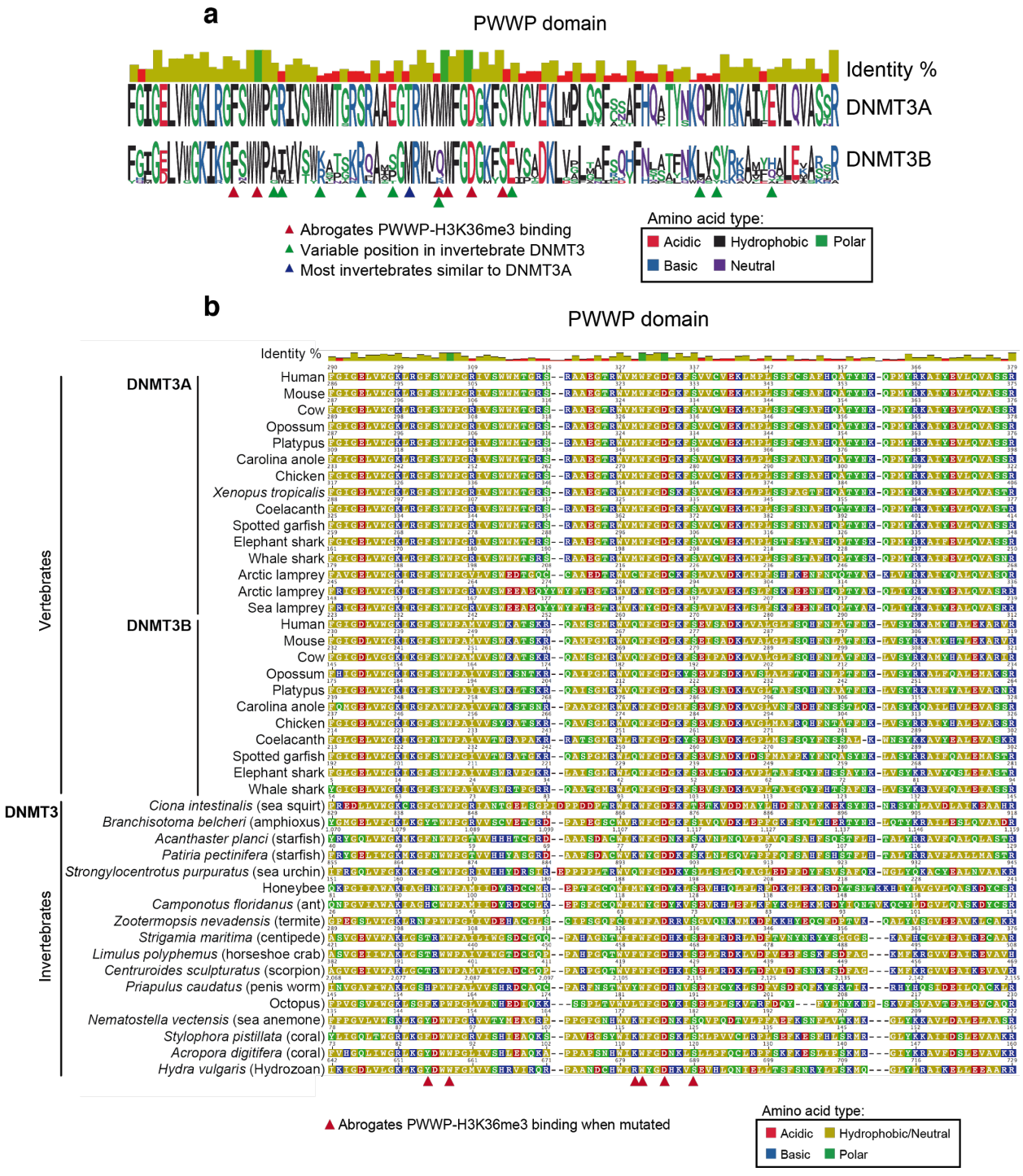


Supplementary Figure 1 | Gene body methylation on gene families/classes. Distribution of gene body methylation on genes belonging to specific transcription factor classes, signaling

molecules, neural related and ribosomal functions. Black dashed lines indicate the bisulfite non-conversion rate for each WGBS library, red dashed lines indicate the global CpG methylation levels in each species. HOX, NKX, DLX, IRX and SIX are Homeobox transcription factor (TF) families. FOX are forkhead domain TFs. SOX are HMG-box TFs belonging to the Sox family. HES are bHLH TFs belonging to the Hes gene family. ZNF are zinc finger C2H2 TFs. WNT, BMP and NOT (Notch) are signaling peptides and receptors. DLG (Discs Large Homologue) and SLC (Solute Carriers) are genes enriched in neural functions. RPL (L Ribosomal Proteins) and RPS (S Ribosomal Proteins) are genes with ribosomal functions. The median CpA methylation levels of invertebrate genes overlaps the non-conversion rate. Boxplot centre lines are medians, box limits are quartiles 1 (Q1) and 3 (Q3), whiskers are 1.5 × interquartile range (IQR) and points are outliers.


a
Human
HCFC1
TMEM187
IRAK1
MECP2
OPN1LW
OPN1MW
chr X (also in mouse and cow)
Opossum
HCFC1
TMEM187
IRAK1
MECP2
OPN1LW
TKTL1
chr X (also in Tasmanian devil)
Platypus
HCFC1
TMEM187
IRAK1
MECP2
OPN1SW
OPN1LW
UltraContig401
Carolina anole
HCFC1
TMEM187
IRAK1
MECP2
OPN1LW
OPN1MW
chr 2 (chromosome 30 in chicken)
X. tropicalis
hcfc1
tmem187
irak1
mecp2
opn2sw
opn1lw
GL172911.1
Zebrafish
hcfc1b
irak1
mecp2
fgd1
tfe3a
chr 8
Elephant shark
TMEM187
MECP2
KI636387.1
Arctic lamprey
TNNC2
C1GALT1
MECP2
RABL1
LMO4
scaffold00096
PAN2
MECP2
FAM133
ITGA7
scaffold00534
SEC13
NUDT16
MECP2
PLXNB2
WNT7
scaffold00063
Human
H1FX
RPL32
EFCAB12
MBD4
IFT122
RHO
chr 3
X. tropicalis
cand2
rpl132
efcab12
mbd4
ift122
rho
GL172832.1
Zebrafish
brpf1
wu:fi23c11
creld1
mbd4
trh
tmc11a
chr 8
Elephant shark
CAND2
CCDC174
RPL32
MBD4
IFT122
RHO
KI635869.1
Arctic lamprey
VGLL4
RPL32
MBD4
IFT122
OPSD
scaffold00637
Amphioxus
STEAP
ACOT
BAP1
MBD4
TMEM115
MEIG1
Sc0000083
b
Human
KIF3C
ASXL2
DTNB
DNMT3A
POMC
EFR3B
chr 2
X. laevis
ccdc25
pbk
dtnb
dnmt3a
dpysl5
mapre3
chr 8S
Zebrafish
kif3cb
asxl2
dtnbb
dnmt3ab
dpys1a
mapre3a
chr 20
Elephant shark
KIF3C
ASXL2
DTNB
DNMT3A
DPYSL5
MAPRE3
KI635932.1
Arctic lamprey
DNMT3A
MAPRE3
EPT1
scaffold00037
DTNB
DNMT3A
PNACTR1
FABP3
scaffold00245
KLH14
DNMT3A
MTRF1/2
ARMC1
scaffold00077
Human
ASXL1
NOL4L
COMMD7
DNMT3B
MAPRE1
EFACB8
chr 20
Zebrafish
nol4lb
commd7
dnmt3.bb1
dnmt3.bb2
dnmt3.bb3
cdc42l2
chr 23
Elephant shark
ASXL1
NOL4L
COMMD7
DNMT3B
MAPRE1
ACSS2
KI635860.1
Zebrafish
ppdpfa
ptk6a
nol4la
dnmt3ba
mapre1a
suv39h1a
chr 8
Spotted gar
asxl1
nol4lb
commd7
dnmt3.bb
dnmt3.ba
mapre1a
acss2
LG18
Human
PWP2
GATD3A
ICOSLG
DNMT3L
AIRE
PFKL
chr 21
Carolina anole
PWP2
GATD3A
ICOSLG
DNMT3L
KDM6A
chr 3
Chicken
PWP2
GATD3A
ICOSLG
DIPK2B
KDM6A
tmc11a
chr1
Arctic lamprey
PARP8
CRHBP
TMEM266
DNMT3L
ISL2
scaffold00018

Supplementary Figure 2 | Syntenic regions surrounding MeCP2, MBD4, and DNMT3 genes. (**a**) Gene order and orientation around the MECP2 and MBD4 loci in the genomes of selected species. Lampreys encode three copies of MECP2. Amphioxus MBD4/MECP2 locus belongs to the *Branchiostoma lanceolatum* genome, and is consistent with previous synteny conservation observed previously in *B. floridae*[47]. (**b**) Gene order and orientation around the DNMT3 loci in selected species genomes. dnmt3ab and dnmt3bb are contiguous in the spotted gar genome assembly, whereas zebrafish has split paralogues in different scaffolds (dnmt3bb.1 is the conserved DNMT3B orthologue, whereas dnmt3ba, dnmt3bb.2 and dnmt3bb.3 are the teleost-specific duplication harbouring a Calponin homology domain). The DNMT3L syntenic region is conserved in bird genomes, however it does not contain the DNMT3L gene. Genes with the same colour are orthologous. Species with less than six genes are found in genomic scaffolds that do not contain more genes, either because it is too short or because it is truncated on one side. Gene names are based on UCSC genome browser annotation or best BLASTP hit against Uniprot in the case of the arctic lamprey.



Supplementary Figure 3 | PWWP domain conservation across DNMT3 genes. (**a**) Amino acid sequence motifs representing the PWWP domain of DNMT3A and DNMT3B orthologues across jawed vertebrates. Coloured triangles indicate specific positions that are known to abrogate PWWP binding to H3K36me3 (red), differ between DNMT3A and DNMT3B but are variable in

invertebrate DNMT3 PWWP domain (green), and positions in which invertebrate DNMT3 are similar to DNMT3A. Amino acid letters are colour coded per type as indicated in the legend. Identity shows the total identity in a multi-sequence alignment including DNMT3A, DNMT3B and invertebrate DNMT3 sequences. (**b**) Complete amino acid multi-sequence alignment of the PWWP (Pro-Trp-Trp-Pro) motif from DNMT3A, DNMT3B, and invertebrate DNMT3 genes. Alignment visualised using Geneious software.

Differences between mammalian CpH methylation patterns in neural and stem cells are attributed to the distinct binding affinities of DNMT3A and DNMT3B[46]. Stem cell CpH methylation and DNMT3B are preferentially localised on transcribed gene bodies and enriched in the CAG trinucleotide context. DNMT3B localisation to gene bodies is mediated by the PWWP domain binding of H3K36me3[88]. In contrast, DNMT3A does not show such a preference in human embryonic stem cells[88], despite encoding a conserved PWWP domain that is able to bind H3K36me2/3 *in vitro*[89]. Furthermore, neural CpH methylation is widespread across the genome, suggesting that DNMT3A does not exclusively bind to gene bodies[10]. Additionally, invertebrate CpG methylation is mostly restricted to gene bodies and has been shown to co-localise with H3K36me3[90], suggesting that DNMT3B might be more similar to the ancestral invertebrate DNMT3 enzymes. To test this idea, we constructed an alignment of DNMT3 enzymes (Supplementary Fig. 9). Counterintuitively, we found that the PWWP domains of invertebrate DNMT3 share higher identity with those of DNMT3A (44%) than DNMT3B (39.2%), highlighting that DNMT3A has accumulated less changes than DNMT3B. Five of the six mutations in the PWWP domain that have been shown to abrogate DNMT3A and DNMT3B preference for H3K36me3[17,44,88] are highly conserved between the ohnologues (Fig. 4d), and thus are not likely to mediate differences in histone tail modification affinity. However, glutamine in position 232 (Q232) of human DNMT3B is not shared with DNMT3A genes, which has a hydrophobic methionine in that position. Therefore, differences between PWWP domain key amino acid

positions are likely to explain the distinct binding affinities of DNMT3 family enzymes in vertebrates, likely fostering the specialisation of each ohnologue into distinct binding patterns and expression domains. Supporting the central role of PWWP in DNMT3 diversification, both DNMT3L and DNMT3C, a rodent-specific DNMT3B paralogue[91], have lost the PWWP domain, which likely allowed novel functions for these genes to emerge. A similar process of specialisation must have occurred at the methyltransferase domain, given the known preference for CAT trinucleotide of DNMT3B, instead of the CAC that characterises DNMT3A and neural CpH methylation. In summary, the redundancy of DNMT3 enzymes is a product of the vertebrate ancestral WGD that was fundamental to allow neofunctionalization of the ohnologues, likely conferring neural-specific functions to DNMT3A at the origin of vertebrates.