


The Neurodevelopmental Gene *MSANTD2* Belongs to a Gene Family Formed by Recurrent Molecular Domestication of *Harbinger* Transposons at the Base of Vertebrates

Éma Etchegaray ¹, Dominique Baas,² Magali Naville,¹ Zofia Haftek-Terreau,¹ and Jean-Nicolas Volff^{*,1}

¹Institut de Génomique Fonctionnelle de Lyon, Ecole Normale Supérieure de Lyon, UCBL1, CNRS UMR 5242, Lyon, France

²Unité MeLiS, UCBL-CNRS UMR 5284, INSERM U1314, Lyon, France

*Corresponding author: E-mail: jean-nicolas.volff@ens-lyon.fr.

Associate editor: Patricia Wittkopp

Abstract

The formation of new genes is a major source of organism evolutionary innovation. Beyond their mutational effects, transposable elements can be co-opted by host genomes to form different types of sequences including novel genes, through a mechanism named molecular domestication. We report the formation of four genes through molecular domestication of *Harbinger* transposons, three in a common ancestor of jawed vertebrates about 500 million years ago and one in sarcopterygians approx. 430 million years ago. Additionally, one processed pseudogene arose approx. 60 million years ago in simians. In zebrafish, *Harbinger*-derived genes are expressed during early development but also in adult tissues, and predominantly co-expressed in male brain. In human, expression was detected in multiple organs, with major expression in the brain particularly during fetal development. We used CRISPR/Cas9 with direct gene knock-out in the F0 generation and the morpholino antisense oligonucleotide knock-down technique to study in zebrafish the function of one of these genes called *MSANTD2*, which has been suggested to be associated to neurodevelopmental diseases such as autism spectrum disorders and schizophrenia in human. *MSANTD2* inactivation led to developmental delays including tail and nervous system malformation at one day post fertilization. Affected embryos showed dead cell accumulation, major anatomical defects characterized by impaired brain ventricle formation and alterations in expression of some characteristic genes involved in vertebrate nervous system development. Hence, the characterization of *MSANTD2* and other *Harbinger*-derived genes might contribute to a better understanding of the genetic innovations having driven the early evolution of the vertebrate nervous system.

Key words: transposable elements, novel genes, vertebrates, nervous system.

Introduction

The formation of new genes is an important source of evolutionary innovation and adaptation for species (Kaessmann 2010). Indeed, they represent a major substrate for the emergence of new functions and contribute to the birth of novel phenotypic traits that are source of adaptation and speciation. For example, new genes can be generated de novo from scratch from initially nonfunctional sequences, a rare phenomenon, or through the duplication of preexisting genes, which can lead to new functions linked to mutations and relaxed selective constraints (Ohno 1972; Lynch and Conery 2000; Knowles and McLysaght 2009; Toll-Riera et al. 2009). Another source of new genes is the recruitment, also called molecular domestication, of transposable element (TE)-coding sequences (Volff 2006; Moran and Malik 2009; Alzohairy et al. 2013).

TEs are repeated DNA sequences that can insert into novel genomic locations and thus can cause genomic instability through insertion and recombination (Kato et al. 2002). TEs have been found in every species that have been investigated. However, the quantitative and

qualitative composition of TEs in genomes is variable depending on the species (Huelsenbeck and Ronquist 2001). While TEs are mutagenic agents that can have neutral or deleterious effects on genomes (Ohno 1972; Doolittle and Sapienza 1980; Orgel and Crick 1980), they can also serve as material for the formation of new regulatory sequences, new exons or even new genes (Kidwell and Lisch 2000; Warren et al. 2015; Chuong et al. 2017). TEs have been source of major innovations during evolution, as exemplified by vertebrate development (Etchegaray et al. 2021). By the process of molecular domestication, TEs can give rise to new functional genes positively selected in host genomes. Major examples of TE domestication have been documented in vertebrates, such as the RAG genes involved in the adaptive immune system and the SYNCYTIN genes necessary for placenta development in mammals (Mallet et al. 2004; Dupressoir et al. 2011; Kapitonov and Koonin 2015; Etchegaray et al. 2021). Thus, TE molecular domestication can lead to important adaptive innovations. In the human genome, which is composed at least of 45% of TEs (Lander et al. 2001), a

© The Author(s) 2022. Published by Oxford University Press on behalf of Society for Molecular Biology and Evolution.

This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial License (<https://creativecommons.org/licenses/by-nc/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

Open Access

hundred cases of protein-coding genes derived from TEs have been identified so far (Volf 2006). However, most of these genes have been poorly characterized, particularly at the functional level. Considering the quantity and diversity of TEs in genomes, their role in the diversification and adaptation of organisms is probably still underestimated (Brandt et al. 2005; Britten 2006; Volf 2006; Alzohairy et al. 2013). Therefore, the identification and functional characterization of new cases of TE-derived genes is important to better understand the formation of novel genes and the factors driving genetic innovation.

In the course of a study aiming to assess the impact of TE molecular domestication on the early evolution of the vertebrate lineage, we have identified several genes domesticated from *Harbinger* TEs through the comparison of human protein sequences to a vertebrate-wide TE sequence database. *Harbinger* transposons are DNA transposons present in the genome of protists, plants, insects, worms and vertebrates but absent from mammals (Kapitonov and Jurka 2004). They are generally flanked by terminal inverted repeat sequences (TIRs) and encode two proteins, a transposase with a aspartate/aspartate/glutamate (DDE) endonuclease motif and a SANT-Myb-trihelix motif-containing protein, which we will now refer to as the Myb-like protein. Both genes have been shown to be necessary for *Harbinger* transposition (Sinzelle et al. 2008; Hancock et al. 2010). The Myb-like protein contains a trihelix motif with conserved bulky aromatic residues that allows DNA and protein binding. Myb-like proteins are responsible for the nuclear import of the transposase through interaction with its N-terminal end. Thanks to the tri-helix motif, they also bind the TIRs of the transposon, allowing the recruitment of the transposase and thus the excision/insertion of the sequence (Sinzelle et al. 2008).

Two cases of *Harbinger*-derived genes have been previously identified in vertebrates: *HARBI1* and *NAIF1* (Kapitonov and Jurka 2004; Sinzelle et al. 2008). *HARBI1* is derived from the transposase gene, while *NAIF1* has been formed from the second gene encoding the Myb-like protein. The *HARBI1* and *NAIF1* proteins can directly interact and form a protein complex, *NAIF1* allowing the nuclear import of *HARBI1*. *NAIF1* can also bind DNA, but not at the *HARBI1* sequence (Sinzelle et al. 2008). *NAIF1* has been linked to apoptosis in the context of several cancers and proposed to have antitumoural effects (Lv et al. 2006; Luo et al. 2011; Fu and Cao 2015; Zhao et al. 2015; Kong and Zhang 2018). However, the biological roles of both genes remain largely unknown.

This study describes a family of genes derived from Myb-like genes of *Harbinger* DNA transposons in jawed vertebrates. We have identified four new genes that have been formed through three to four independent molecular domestication events during vertebrate evolution, three at the base of jawed vertebrates about 500 million years ago and a fourth one possibly in a common ancestor of sarcopterygians ca. 430 million years ago. The *Harbinger*-derived genes are expressed during zebrafish embryonic development and in zebrafish adult tissues, predominantly in

male brain, as well as in human brain during fetal development. Inactivation of one of these genes, *MSANTD2*, by CRISPR/Cas9 direct knock-out in F0 and morpholino antisense oligonucleotide knock-down techniques in zebrafish led to embryos with severe brain developmental defects and modification of the expression of characteristic genes involved in vertebrate nervous system development. Interestingly, *MSANTD2* has been suggested to be associated with neuro-developmental diseases such as autism spectrum disorders and schizophrenia in human (Schizophrenia Working Group of the Psychiatric Genomics Consortium 2014; Lim et al. 2017; O'Brien et al. 2018; Zhang et al. 2020).

Results

Multiple Molecular Domestication Events of *Harbinger* Transposons Have Formed a New Gene Family in Vertebrates

Identification of New *Harbinger* Myb-Like-Derived Genes in Jawed Vertebrates

To the best of our knowledge, we performed the first systematic comparison of human protein sequences to an extensive TE sequence database generated during a broad survey in vertebrates, which particularly included many TE families from fish that are absent from mammals and birds (Chalopin et al. 2015). We have identified, in addition to genes already described as TE-derived genes (Volf 2006; Alzohairy et al. 2013), a gene called *MSANTD2* (Myb/SANT DNA Binding Domain Containing 2) as a new potential case of molecular domestication from a *Harbinger* DNA transposon. Indeed, the *MSANTD2* predicted protein sequence presented homologies with the Myb-like protein of a *Harbinger* transposon from the genome of the medaka fish *Oryzias latipes*, with a conservation score (considering both residue identities and conservation of physico-chemical classes) of 54% in the Myb-like domain region (150 aa). The *MSANTD2* gene is located on chromosome 11 in the human genome and is 2,384 base pairs (bp) in length, with four exons encoding a protein of 559 amino-acids (aa). Prediction of conserved domains on the whole sequence of the *MSANTD2* protein revealed a single domain, a Myb-like DNA binding domain containing a trihelix motif. *MSANTD2* was different from the two other genes derived from *Harbinger* transposons previously described in human, *HARBI1* that has been formed from a transposase gene and *NAIF1* from a Myb-like gene (Kapitonov and Jurka 2004; Sinzelle et al. 2008).

We further identified three additional cases of *Harbinger*-derived genes in the human genome, called *MSANTD1*, *MSANTD3* and *MSANTD4*. These three genes encode predicted proteins with similarities to Myb-like proteins of *Harbinger* transposons. In human, *MSANTD1* is on chromosome 4 and 3,164 bp in length, with three exons coding for a 278 aa protein. *MSANTD3*, on chromosome 9, is a 1,880 bp gene with three exons encoding a 275 aa protein. Finally, *MSANTD4*, on chromosome 11 and 4103 bp in length, contains three exons coding for a 345

aa protein. Prediction of conserved domains on each MSANTD1, MSANTD3 and MSANTD4 protein also revealed a Myb-like domain containing a tri-helix motif (fig. 1). This suggested functional homology of the MSANTD proteins with the Myb-like proteins of *Harbinger* transposons, with possible DNA- and/or protein-binding properties.

The phylogenetic distribution of *Harbinger*-derived genes including *NAIF1* and *HARBI1* was determined using the Ensembl and NCBI databases and verified by blast analysis on metazoan genomes (fig. 2A) (Altschul et al. 1990). All genes were detected only in jawed vertebrates from cartilaginous fishes to mammals, except for *MSANTD3*, which was absent from both cartilaginous and ray-finned fish genomes but present in sarcopterygians. This suggested the formation of *MSANTD1*, *MSANTD2*, *MSANTD4*, *NAIF1* and *HARBI1* genes early during vertebrate evolution at the base of jawed vertebrates around 500 Mya, and the occurrence of *MSANTD3* in a common ancestor of sarcopterygians around 430 Mya (Lu et al. 2016). Synteny analysis showed that each *MSANTD* gene was present at the same position in the genome of divergent vertebrate species, indicating that they corresponded to *bona fide* genes and not to mobile transposon sequences anymore (fig. 2B).

Moreover, *MSANTD2P1*, an intronless processed pseudogene (according to the loss of its protein-coding capacity), probably originating from the retrotransposition of *MSANTD2* mRNA, was detected in simians, i.e., from human (on chromosome 21) to marmoset but neither in macaques nor in baboons (supplementary fig. 1, Supplementary

Material online). This suggested that *MSANTD2P1* appeared at the base of simians about 36–50 Mya (Perelman et al. 2011).

Vertebrate Myb-Like-Derived Genes Originated From Three to Five Independent Molecular Domestication Events of Harbinger Transposons

Predicted Myb-like-derived proteins were compared to *Harbinger* TE sequences collected from the Repbase database or annotated from sequenced genomes. Multiple sequence alignments were built, comparing the most similar (i.e., with the lowest *E*-value, all of them $<10^{-5}$) *Harbinger* transposon Myb-like proteins from different species with each Myb-like-derived protein (fig. 1). This revealed conservation between the Myb-like domain region of the MSANTD proteins and the Myb-like proteins of *Harbinger* transposons (covering 115 to 167 aa). The conservation scores (considering both residue identities and conservation of physico-chemical classes) between each MSANTD protein and its closest *Harbinger* Myb-like proteins were calculated all along the Myb-like domains. For the different MSANTD proteins these scores were estimated between 45 and 57%. Putative alpha helices and aromatic residues, which are essential for Myb-like domain function, were also conserved (fig. 1).

In order to investigate the evolutionary origin(s) of MSANTD genes, phylogenetic trees were built based on protein alignments using the Bayesian method (Huelsenbeck and Ronquist 2001) (fig. 3). While the sequence of the transposase of *Harbinger* transposons is highly conserved

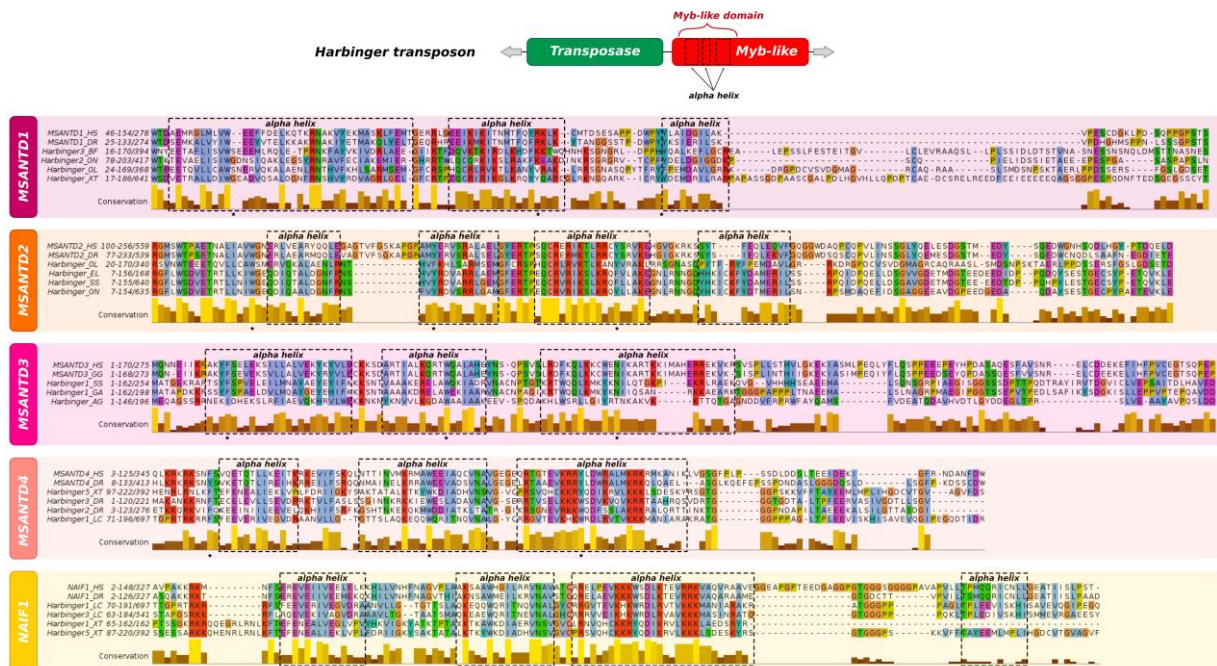


FIG. 1. Multiple alignments of the Myb-like domain of MSANTD proteins and their closest Myb-like proteins from *Harbinger* transposons. Predicted alpha-helix motifs are represented with dashed squares; bulky aromatic residues, which are essential for alpha helix structure stabilization, are indicated by black stars. The conservation score represented for each residue is measured considering both residue identities and conservation of physico-chemical classes. AG, *Anopheles gambiae*; DR, *Danio rerio*; EL, *Esox lucius*; GA, *Gasterosteus aculeatus*; GG, *Gallus gallus*; HS, *Homo sapiens*; LC, *Latimeria chalumnae*; OL, *Oryzias latipes*; ON, *Oreochromis niloticus*; SS, *Salmon salar*; TF, *Takifugu flavidus*; XT, *Xenopus tropicalis*.

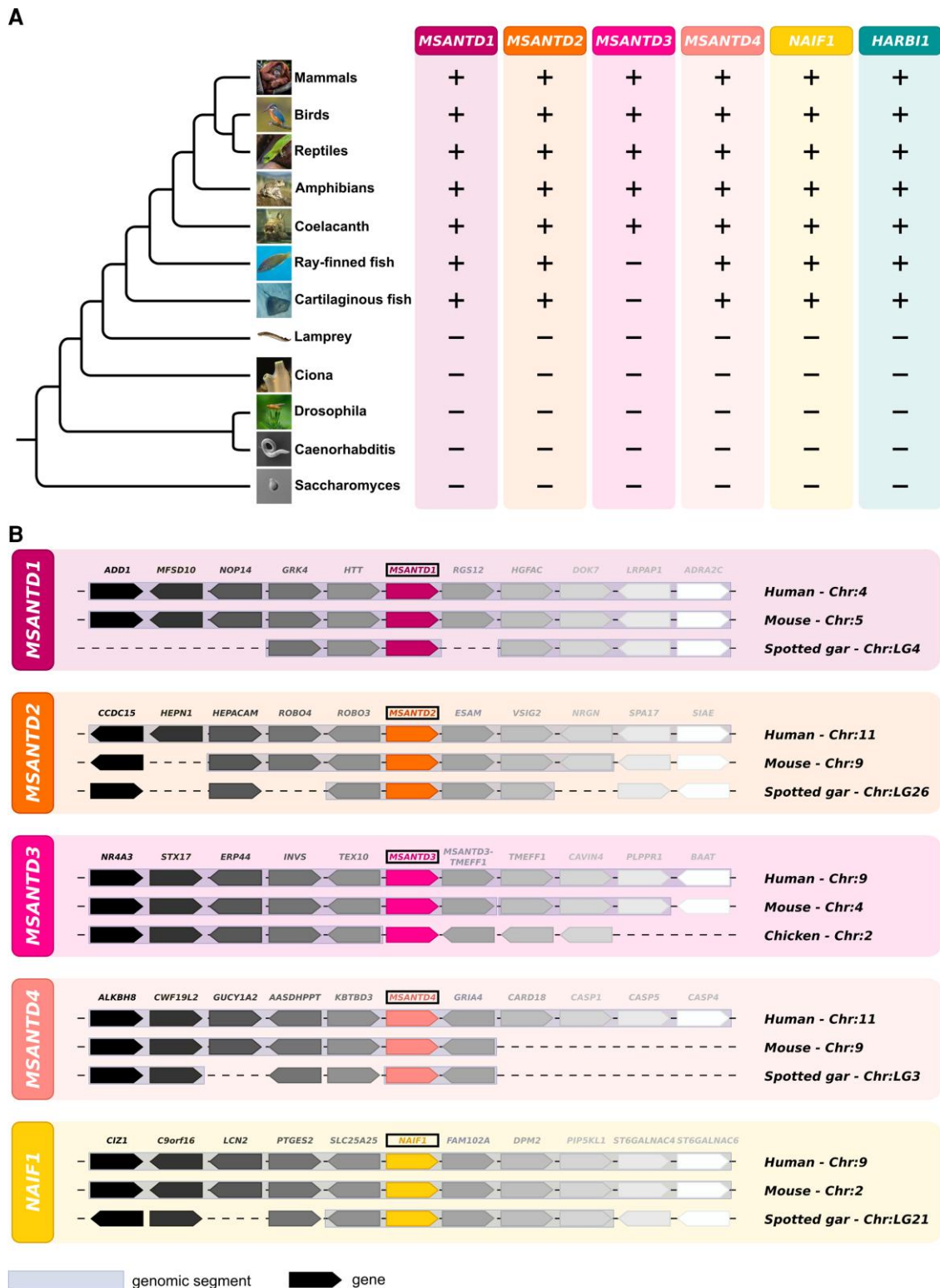


FIG. 2. (A) Phylogenetic distribution of *Harbinger*-derived genes and *Harbinger* transposons. Presence (+) or absence (-) of these genes in the different lineages is indicated. (B) Synteny analysis of *Myb*-like-derived genes between human, mouse and spotted gar (nonteleost ray-finned fish) or chicken (for *MSANTD3*, which is absent from both cartilaginous and ray-finned fish). Species names and genomic locations are shown on the right. For each gene the same color stands for orthologous genes.

between different families, this is not the case for the *Myb*-like transposon proteins, which are much more divergent (Kapitonov and Jurka 2004). Such an important sequence divergence was also observed between most

MSANTD proteins. Therefore, we were not able to reconstruct reliable general sequence alignment and phylogeny for all *MSANTD* and transposon *Myb*-like proteins together. However, *MSANTD1* and *MSANTD2* were most similar to

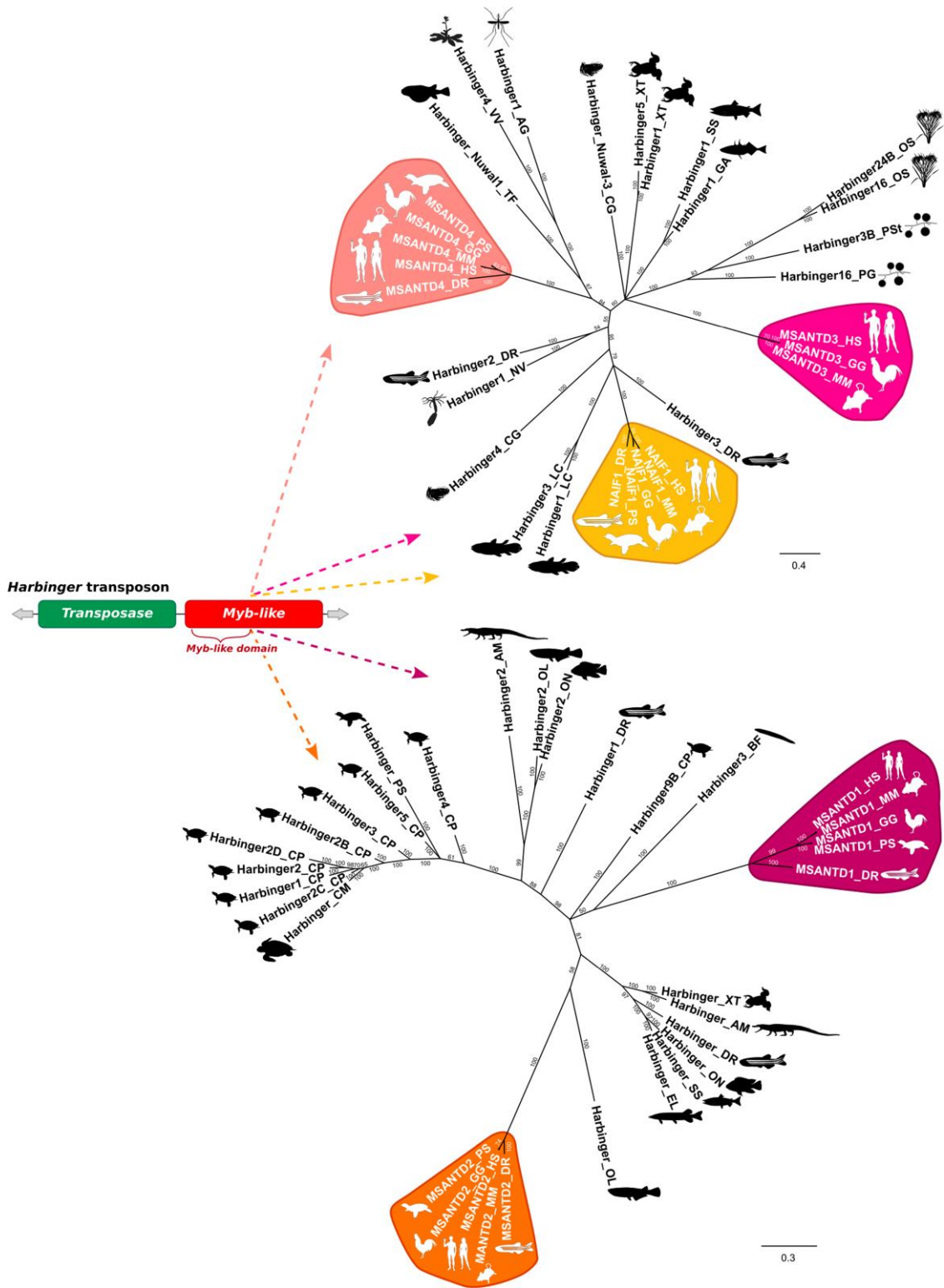


Fig. 3. Phylogenetic relationships between MSANTD proteins and their closest Myb-like proteins from Harbinger transposons. Trees were constructed using the Bayesian method (Huelsenbeck and Ronquist 2001). Only branch support values higher than 50% are shown. AG, *Anopheles gambiae*; AM, *Alligator mississippiensis*; BF, *Branchiostoma floridae*; CG, *Crossostrea gigas*; CM, *Chelonia mydas*; CP, *Chrysemys picta*; DR, *Danio rerio*; EL, *Esox lucius*; GA, *Gasterosteus aculeatus*; GG, *Gallus gallus*; HS, *Homo sapiens*; LC, *Latimeria chalumnae*; MM, *Mus musculus*; NV, *Nematostella vectensis*; OL, *Oryzias latipes*; ON, *Oreochromis niloticus*; OS, *Oryza sativa*; PG, *Puccinia graminis*; PS, *Puccinia striiformis*; SS, *Salmon salar*; TF, *Takifugu flavidus*; VV, *Vitis vinifera*; XT, *Xenopus tropicalis*). Silhouette images from phylopic.org.

the same group of *Harbinger* sequences, and MSANTD3, MSANTD4 and NAIF1 to the same other group of transposon sequences. This allowed generating two different

sets of multiple sequence alignments and phylogenies: one for MSANTD1 and MSANTD2 with related *Harbinger* transposon proteins, and another one for

MSANTD3, MSANTD4, NAIF1 and related *Harbinger* sequences (fig. 3). Phylogenies indicated that each MSANTD sequence from different species formed an independent monophyletic group, and that the closest related *Harbinger* transposons were different for each MSANTD sequence. Hence, this supported five independent events of molecular domestication, four at the base of jawed vertebrates and a fifth one later in a common ancestor of sarcopterygians for MSANTD3. Phylogenies were also constructed with the Maximum Likelihood method and showed similar results (supplementary fig. 2, Supplementary Material online). However, in this analysis, MSANTD3 and MSANTD4 did not clearly group with a specific *Harbinger* transposon, and the clustering with transposon was not highly statistically supported for NAIF1. Hence, Maximum Likelihood analysis suggested at least three events of molecular domestication, two for MSANTD1 and MSANTD2 and at least a third one for MSANTD3, MSANTD4 and NAIF1.

In order to test if some MSANTD genes might have been formed through larger segmental genomic duplications, their flanking genomic regions were compared by synteny analyses. No evidence for paralogous sequences that might have been co-duplicated with MSANTD genes was found, consistent with more local events (fig. 2B).

Taken together, the results indicated that MSANTD1, MSANTD2, MSANTD3 and MSANTD4 are four new cases of vertebrate genes derived from *Harbinger Myb-like* transposon sequences, in addition to NAIF1. These genes arose from three to four independent molecular domestication events at the base of jawed vertebrates around 500 Mya, with another potential one at the base of sarcopterygians around 430 Mya that generated MSANTD3.

Vertebrate Myb-Like-Derived Genes Evolved under Negative Selection

To further investigate the evolutionary constraints having acted on vertebrate MSANTD genes, we performed a positive/negative selection test using CODEML (Yang et al. 2007). We calculated the dN/dS ratio (ratio between nonsynonymous vs. synonymous substitution rates) as a proxy for selection pressure (supplementary Table 1, Supplementary Material online). All ratios were smaller than 1, reflecting a higher rate of synonymous than nonsynonymous substitutions. These ratios were comparable to those of other genes in genomes (0.066 on average in human-zebrafish comparisons) (Wolf et al. 2009). Hence, MSANTD sequences evolved under negative/purifying selection in vertebrates, i.e., these genes were functionally constrained. The ratio for the MSANTD2P1 pseudogene was closer to 1 compared to other MSANTD genes, in accordance with relaxed constraints and loss of protein-coding capacity.

Harbinger-Derived Genes Are Expressed in Zebrafish during Embryonic Development and Predominantly in Adult Male Brain

The expression of the MSANTD1, MSANTD2, MSANTD4, NAIF1 and HARBI1 *Harbinger*-derived genes was studied by quantitative PCR (qPCR) in zebrafish embryos (fig. 4A).

All these genes were expressed during zebrafish embryonic development, with HARBI1 being the most expressed gene. Except for MSANTD1, which is more expressed at later stages, most genes were more strongly expressed at the first stages of development before the midblastula transition (MBT), suggesting maternal effect. Using *in situ* RNA hybridization, MSANTD2, which was chosen for further functional analyses (see below), was found to be expressed during zebrafish embryonic development in the whole embryo from 1.25 h post fertilization (hpf) to 17hpf (fig. 4C). From 6hpf, MSANTD2 was more strongly expressed in the anterior side of the embryo, the region leading to the head and the central nervous system. At 24hpf, the expression of MSANTD2 was specifically restricted to the head region, and more particularly to the forebrain, midbrain and hind-brain regions of the brain.

Expression of *Harbinger*-derived genes was also studied in zebrafish adult tissues by qPCR (fig. 4B). We observed both a sex- and tissue-biased expression of these genes. Particularly, *Harbinger*-derived genes were predominantly co-expressed in male but not female brain. As observed in embryos, HARBI1 was also the most expressed gene in adult tissues, with stronger expression in liver and muscle of both males and females.

Harbinger-Derived Genes Are Expressed in Human, Particularly during Brain Development

According to the National Institutes of Health (NIH) genotype-tissue expression (GTEx) project (dbGaP Accession phs000424.v8.p2; GTEx Consortium 2013), all *Harbinger*-derived genes appeared to be expressed in human brain as well as in some other tissues depending on the gene. Using the BrainSpan Atlas of the Developing Human Brain (www.brainspan.org) (Miller et al. 2014), expression was detected in human brain before and after birth, except for the MSANTD2P1 pseudogene (fig. 5). MSANTD genes were expressed in the whole brain particularly during early fetal development at the first/second trimesters of pregnancy, with decreasing expression in the third trimester (around 10 weeks before birth) (fig. 5A). MSANTD3 and MSANTD4 were the most expressed genes and MSANTD2 and NAIF1 presented the same expression pattern but with lower expression. HARBI1 had a more ubiquitous expression, with higher expression in early fetal development as observed for the MSANTD genes, but also later after 13 years. MSANTD1 presented a more localized expression in a specific brain structure, the striatum, during the second trimester of pregnancy (from 13 to 24 postconceptional weeks [pcw]) (fig. 5B). These results showed that *Harbinger*-derived genes are particularly expressed during fetal brain development, in whole human brain for most genes or in a more specific brain region (striatum) for MSANTD1.

MSANTD2 Inactivation Leads to Zebrafish Embryos with Severe Neuro-Developmental Defects

The biological function of vertebrate *Harbinger*-derived genes was further investigated by gene inactivation.

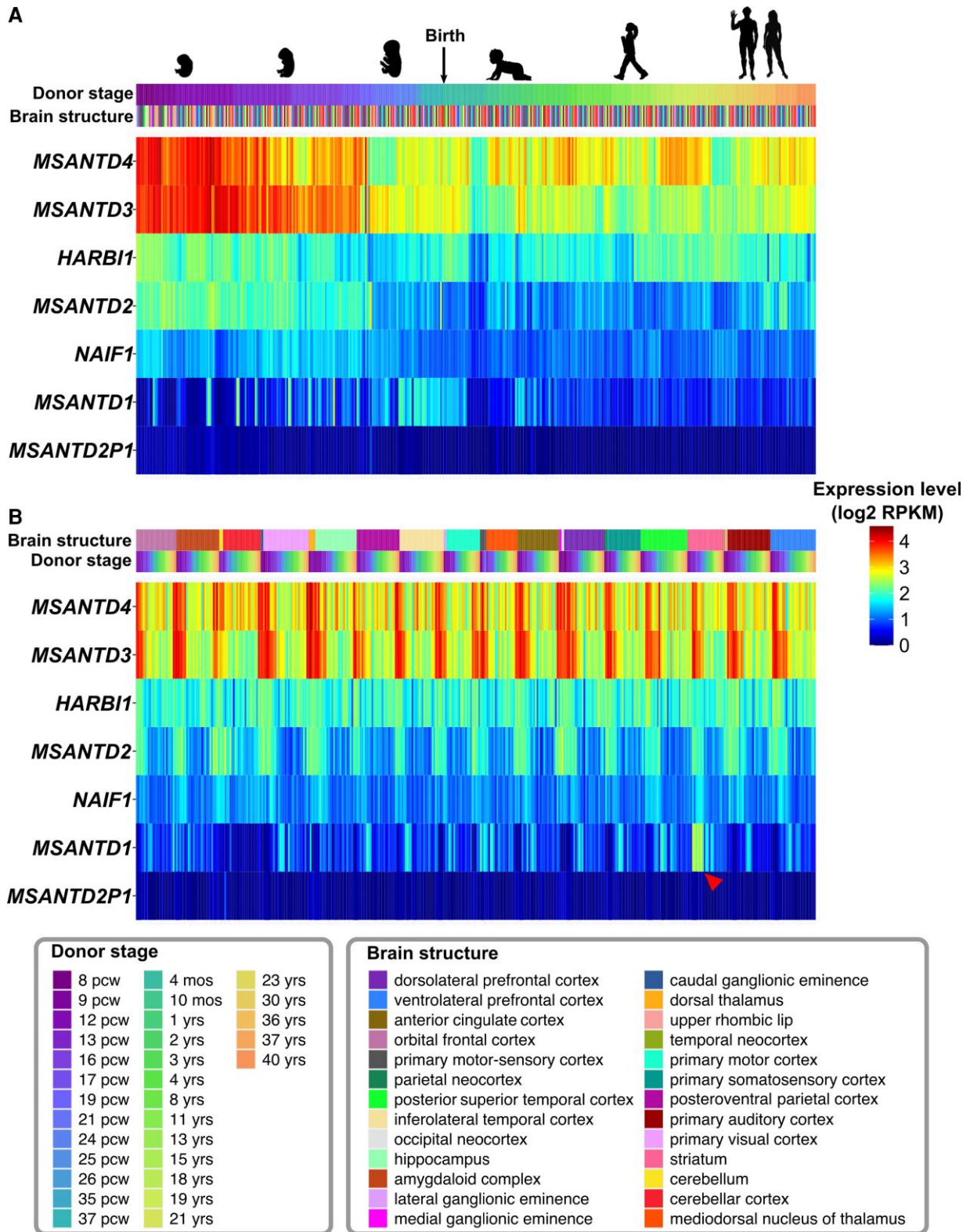


FIG. 5. Expression of *Harbinger*-derived genes in human brain before and after birth according to donor stages (A) or brain structures (B). For each gene, the expression is shown in log₂ reads per kilobase per million (RPKM) for different donor stages (pcw, postconceptional weeks; mos, months; yrs, years) and in different brain structures, represented with multiple colors. Data were obtained from the BrainSpan Atlas (www.brainspan.org) (Miller et al. 2014). The striatum-specific expression of *MSANTD1* is indicated with a red arrowhead. Silhouette images are from lifesizesilhouette.com and (Haniffa et al. 2021).

such as autism spectrum disorders and schizophrenia (Schizophrenia Working Group of the Psychiatric Genomics Consortium 2014; Lim et al. 2017; O'Brien et al. 2018; Zhang et al. 2020).

In order to inactivate *MSANTD2* by CRISPR/Cas9, zebrafish embryos were injected with one sgRNA or different combinations of two or four sgRNAs. The observed phenotypes were similar with almost all the different combinations of sgRNAs, although the penetrance was variable (fig. 6G; supplementary fig. 3, Supplementary Material online). Therefore, only three treatments (combination of four sgRNAs, combination of sgRNAs 1 and 4, sgRNA 4 only) will be further detailed (fig. 6G). Sequencing of injected embryos revealed mutations at all sgRNA loci, with multiple frameshift nucleotide deletions leading to premature stop codons (supplementary figs. 4–5, Supplementary Material online). When looking at embryos injected with four sgRNAs, read coverage analysis showed that almost all reads (ca. 90%) showed mutations in the first exon at the sgRNA 1 locus, in addition to mutations at the three other sgRNA loci.

Embryos injected with four sgRNAs showed developmental delays as well as tail and nervous system malformations compared to control embryos at 24hpf (fig. 6A–F). Heads were smaller and tails curved with not well-defined somites (fig. 6D, black arrowheads). Moreover, *MSANTD2* CRISPR/Cas9 embryos presented defects in neural tube folding (fig. 6F, white arrowheads), with cell aggregates (i.e., cells that are loosely grouped together) visible around the nervous system (fig. 6E–F, black arrows). Similar phenotypes were observed when embryos were injected with sgRNAs 1 and 4 together, and intermediate phenotypes with sgRNA 4 alone (supplementary fig. 6, Supplementary Material online). In a typical experiment (fig. 6), about 25% of embryos injected with all four sgRNAs presented strong phenotypes (developmental delays, tail malformations, nervous system malformations, cell aggregates) and around 50% intermediate phenotypes (developmental delays, nervous system malformations, no or few cell aggregates, no tail malformation) (fig. 6G). Moreover, *MSANTD2* inactivation appeared to severely compromise development, since 40–85% of injected embryos with phenotypes died few days or weeks post injection (compared to 6–16% for control embryos). Similar phenotypes were also observed after injection of morpholino antisense oligonucleotide directed against *MSANTD2* (supplementary fig. 7, Supplementary Material online). Developmental delays as well as tail and nervous system malformations were observed in *MSANTD2* CRISPR-Cas9 zebrafish embryos at 15hpf and 19hpf too (supplementary fig. 8, Supplementary Material online).

In order to assess the specificity of the phenotypes observed, we performed a rescue experiment by co-injection of *MSANTD2* mRNA with the four *MSANTD2*-directed sgRNAs in zebrafish embryos (supplementary fig. 9, Supplementary Material online). We observed that the mutant phenotypes were strongly rescued, since most embryos presented wild-type (WT) phenotypes or only slight developmental delays. Particularly, we observed a well-developed nervous system in the rescued embryos.

Concerning the nervous system, the midbrain-hindbrain boundary (MHB), which is a well-defined structure of the 24hpf stage of zebrafish development, was not well formed (fig. 6A–F, white stars). In order to characterize nervous system malformations, dextran Texas Red was injected into zebrafish brain ventricles at 24hpf and 30hpf (fig. 6H–O) (Gutzman and Sive 2009). At each stage, three pictures of *MSANTD2* CRISPR/Cas9 embryos were compared to a control. All *MSANTD2* CRISPR/Cas9 embryos presented brain abnormalities with neural tubes misfolding particularly in the MHB region (fig. 6H–O, white arrowheads), as well as defects in forebrain, midbrain and hindbrain inflation. These phenotypes were observed from 24hpf and were still present at 30hpf. Dead cells were marked in 24hpf embryos with acridine orange staining (fig. 7A–D). We observed numerous dead cells in *MSANTD2* CRISPR/Cas9 embryos compared to control embryos. Hence, the cell aggregates we observed might correspond to dead cell areas (fig. 7C–D, regions of cell aggregates are indicated with white stars).

To further study the role of *MSANTD2* in nervous system development, this gene was inactivated using the same protocol of CRISPR/Cas9 with four sgRNAs in the Tg(elavl3:GCaMP6s) zebrafish line (fig. 7E–I). This is a transgenic line containing a green fluorescent protein (GFP)-based calcium sensor, with the *elav3* promoter fused to the *GCaMP6s* genetically encoded calcium indicator, marking fluorescently all differentiated neurons (Park et al. 2000; Panier et al. 2013). At each stage, two pictures of *MSANTD2* CRISPR/Cas9 embryos were compared to a control to assess the variability of neuronal patterns. From 24hpf to 72hpf, general anatomical defects characterized by aberrant patterning of early neurons were visible in *MSANTD2* CRISPR/Cas9 embryos (fig. 7E–I, white arrowheads). Moreover, the abnormal pattern of neuronal marking was not only explained by developmental delay, since *MSANTD2* mutated embryos were still different from control embryos at later stages of development.

We studied in 24hpf *MSANTD2* CRISPR/Cas9 embryos the expression of characteristic genes involved in vertebrate nervous system development (fig. 7J–N). Because the MHB is a well-defined structure at the 24hpf stage of zebrafish development, we characterized it with the expression of the *FGF8*, *PAX2A* and *HER5* genes. *FGF8* encodes a fibroblast growth factor involved in several processes, including nervous system development particularly for two brain structures, the tectum and cerebellum. *FGF8* is also responsible for the maintenance of the MHB together with *PAX2A* (Nakamura 2001; Chi et al. 2003). In 24hpf control embryos, *FGF8* was expressed in telencephalon, dorsal diencephalon, optic stalks, otic vesicle and MHB regions. In addition to MHB, *PAX2A* was expressed in optic stalks, otic vesicle and hindbrain neurons at 24hpf (fig. 7J, K, M). The MHB is also characterized by the expression of *HER5*, which is involved in multiple developmental processes and particularly brain development (fig. 7L). Finally, as we observed dead cell aggregates around the nervous system in the *MSANTD2* CRISPR/Cas9 embryos (figs. 6 and 7), we questioned whether they might correspond to dead neural crest cells

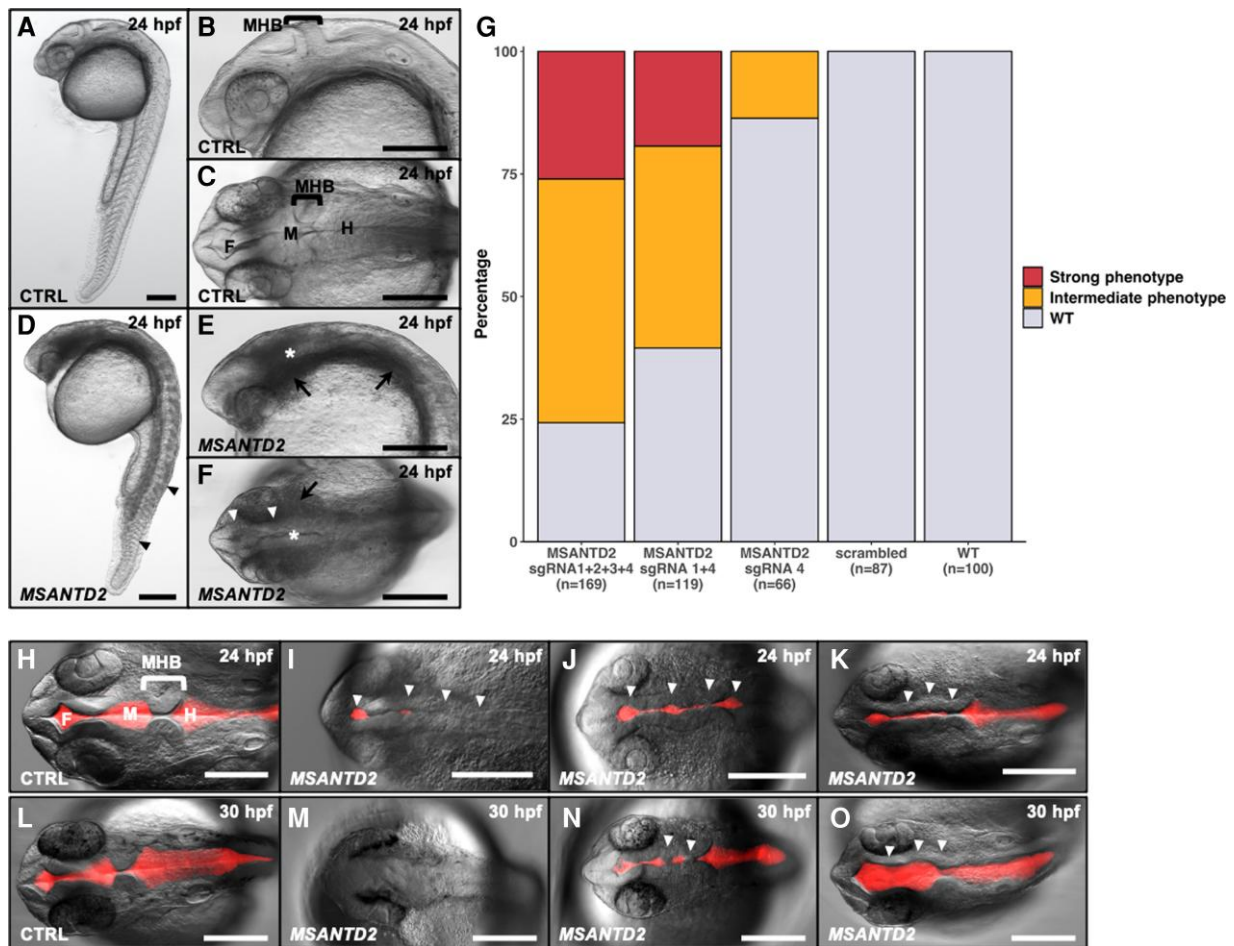


FIG. 6. *MSANTD2* CRISPR/Cas9 embryo phenotypes. Embryos injected with control (A–C, CTRL) or four *MSANTD2*-directed sgRNAs (D–F, *MSANTD2*). A/D and B/E present lateral views of whole embryos and of the head region, respectively. C/F show dorsal views of the embryo head region. Embryos injected with four sgRNAs showed developmental delays as well as tail and nervous system malformations compared to control embryos at 24hpf (A–F). Tail curvature and not well-defined somites are shown with black arrowheads (D). Default in neural tube folding and cell aggregates around the nervous system are indicated with white arrowheads and black arrows, respectively (E–F). Not well-formed MHB are shown with white stars (*) (E–F). (G) Proportions of F0 zebrafish phenotypes. Embryos were injected with sgRNA 1 + 2 + 3 + 4, sgRNA 1 + 4 and sgRNA 4 targeting *MSANTD2*, four scrambled sgRNAs or noninjected (WT), and were scored for phenotypes at 24hpf. Percentages with strong (red), intermediate (orange) or no phenotypes (gray–WT) are shown. Strong phenotypes: developmental delays, tail malformations, nervous system malformations, aggregates; intermediate phenotypes: developmental delays, nervous system malformations, no or few aggregates, no tail malformation. (H–O): Dextran Texas Red injection in brain ventricles of 24hpf (H–K) or 30hpf (L–O) embryos injected with control (H, L, CTRL) or four *MSANTD2*-directed sgRNAs (I–K, M–O, *MSANTD2*) (dorsal view). At each stage, three pictures of *MSANTD2* CRISPR/Cas9 embryos were compared to a control picture in order to represent phenotype variability. The proportion of embryos presenting the same phenotypes are the following: I—31%, J—47%, K—22% at 24hpf and M—14%, N—43%, O—43% at 30hpf. All cases illustrate neural tubes misfolding (shown with white arrowheads) particularly in the MHB region. We also observed smaller red fluorescent areas in the forebrain and midbrain regions, indicating reduction of these ventricles. F, forebrain, M, midbrain; H, hindbrain. Scale bar 200 μ m.

by studying the expression of *DLX2*, which marks the cranial migratory neural crest cells that form the pharyngeal arches and migrate into the forebrain (Akimenko et al. 1994; Yan et al. 2005; Sperber et al. 2008; Dai et al. 2013).

In *MSANTD2* CRISPR/Cas9 embryos, we observed alterations of *FGF8*, *PAX2A* and *HER5* expression particularly in the MHB region (fig. 7J–M). These genes were still expressed, but the marked areas were different between control and *MSANTD2* CRISPR/Cas9 embryos. The expression bands were narrower for *FGF8* and *PAX2A* (fig. 7K, M) and also less deep for *FGF8* (fig. 7J). For *HER5*, the staining into two distinct areas was lost in *MSANTD2* CRISPR/Cas9

embryos (fig. 7L). Moreover, the expression of *FGF8* in telencephalon and optic stalks was not separated into two different zones but formed a unique and larger area, suggesting defects in the definition and individualization of these structures (fig. 7J, K). *MSANTD2* CRISPR/Cas9 embryos lacked the expression of *PAX2A* in hindbrain neurons (fig. 7M). Finally, the expression of *DLX2* was markedly reduced in the telencephalon and pharyngeal arch regions (fig. 7N). In conclusion, these results indicated anatomical defects of the *MSANTD2* CRISPR/Cas9 embryos in the MHB and telencephalon regions. Finally, the accumulation of dead cells and the expression patterns of *PAX2A* and

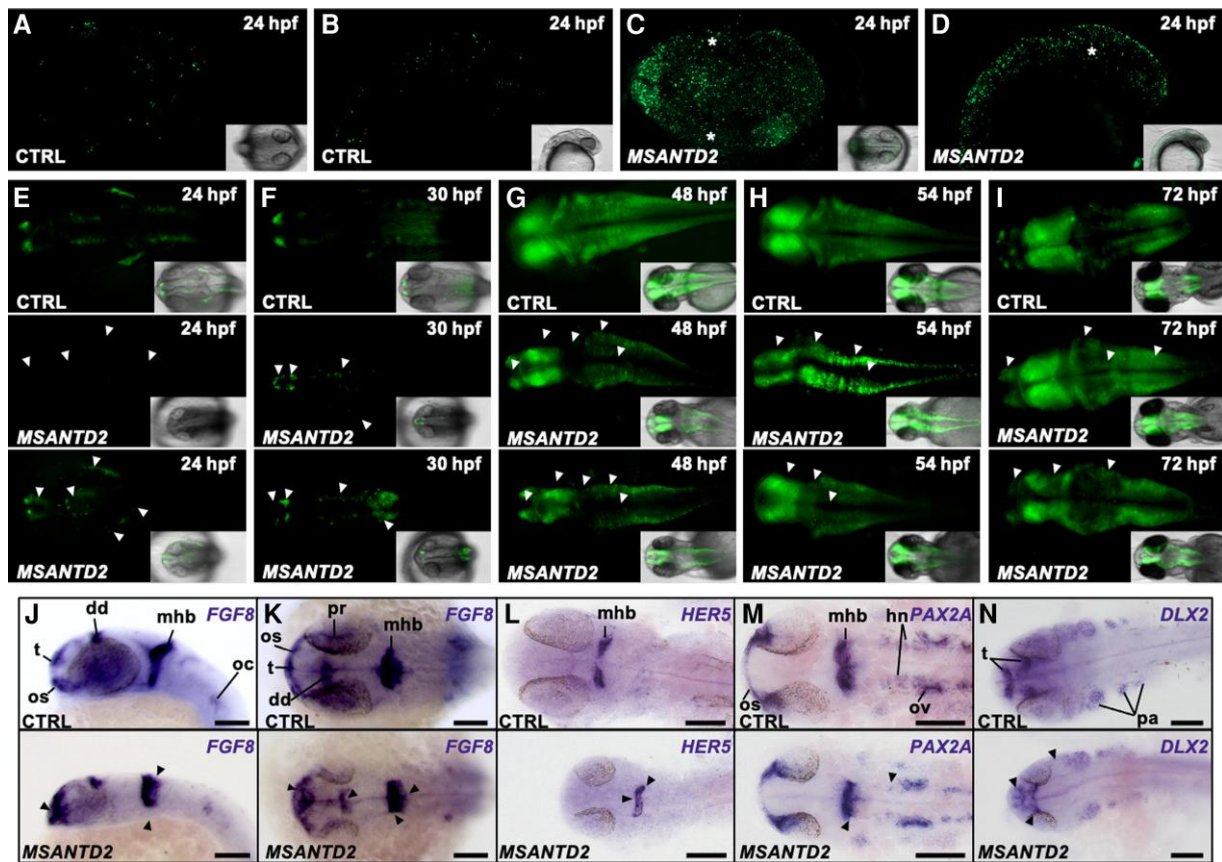


Fig. 7. Effects of *MSANTD2* inactivation by CRISPR/Cas9 in zebrafish embryos. (A–D) Acridine orange staining of embryos injected with control (CTRL) or four *MSANTD2*-directed sgRNAs (*MSANTD2*) at 24hpf. Numerous dead cells were visible in *MSANTD2* CRISPR/Cas9 embryos compared to control embryos. Regions of cell aggregates were indicated with white stars. (E–I) Embryos injected with control (CTRL) or four *MSANTD2*-directed sgRNAs (*MSANTD2*) in Tg(elavl3:GCaMP3) zebrafish embryos. At each stage, two pictures of *MSANTD2* CRISPR/Cas9 embryos were compared to a control in order to represent the variability of impaired neuronal pattern. Differentiated neurons were marked by green fluorescence. The results showed general anatomical defects characterized by different patterns of fluorescence between *MSANTD2*-mutated and control embryos (shown with white arrowheads). Abnormal pattern of neuronal marking was not only explained by developmental delay, since *MSANTD2* mutated embryos are still different from control embryos at later stages. (J–N): Expression of *FGF8* (J, K), *HER5* (L), *PAX2A* (M) and *DLX2* (N) at 24hpf in embryos injected with control (CTRL, top) or four *MSANTD2*-directed sgRNAs (*MSANTD2*, bottom). J present lateral and K/L/M/N dorsal views of the head region of the embryos, respectively. Differences in gene expression patterns between *MSANTD2*-mutated and control embryos are indicated with black arrowheads. Dd, dorsal diencephalon; hn, hindbrain neurons; oc, otic capsule; os, optic stalks; ov, otic vesicle; pa, pharyngeal arches; pr, proximal part of retina; t, telencephalon; tp, thyroid primordium. For each picture at least $n = 10$ embryos were observed.

DLX2 also suggested potential implication of *MSANTD2* in neural crest cell migration, homing or differentiation into neurons (visible for hindbrain neurons).

SOX10 (SRY-box transcription factor 10) is a marker of neural crest cells, particularly marking the neural crest cell streams at 15hpf and 19hpf (supplementary fig. 8, Supplementary Material online). The most anterior stream (S1), second stream (S2) and third stream (S3) correspond to the cranial neural crest cells that migrate into the mandibular arch, the hyoid arch and posterior pharyngeal arches, respectively (Piotrowski and Nüsslein-Volhard 2000). Cranial neural crest cells can then give rise to multiple derivatives: chondrocytes, osteocytes, cranial sensory ganglia, pigment cells, connective tissue, Schwann and satellite cells (Schilling and Kimmel 1994; Kague et al. 2012). Using a zebrafish Tg(sox10:mRFP) transgenic line, in which *SOX10* was fluorescently labeled, we observed defects in *SOX10* marking

in *MSANTD2* CRISPR/Cas9 embryos at both 15hpf and 19hpf, particularly in the S1 stream region. Moreover, we also detected impaired marking in the optic tectum region, a multi-tissue part of the midbrain involved in the visual system. These results suggest potential migratory delays or defects of the cranial neural crest cells in *MSANTD2* CRISPR/Cas9 mutant embryos, potentially leading to altered functions of their cellular derivatives.

Discussion

Harbinger Transposons Have Given Rise to a New Gene Family through Recurrent and Concomitant Molecular Domestication Events in Vertebrates

In this work, we report recurrent and concomitant molecular domestication of *Harbinger* transposons in early

vertebrate evolution. We have identified in jawed vertebrates a new family of genes derived from *Harbinger* elements, and more particularly from their *Myb-like* gene (figs. 1–3). Indeed, MSANTD1, MSANTD2, MSANTD3 and MSANTD4 presented sequence similarities with the *Myb-like* domain of proteins from *Harbinger* transposons (fig. 1). Each MSANTD gene is present as a single copy gene at a conserved position in vertebrate genomes (fig. 2). Hence, MSANTD genes are not transposons anymore but *bona fide* vertebrate genes. In order to investigate whether MSANTD genes arose from independent molecular domestication or/and sequential duplication events, sequence alignments and phylogenies were constructed (figs. 1 and 3). The sequences of all MSANTD genes and proteins could not be aligned unambiguously together due to high divergence. Indeed, the *Myb-like* proteins of different families of *Harbinger* transposons present significant similarities only restricted to a short part of their *Myb-like* domain. In contrast, *Harbinger* transposases are much more conserved. After constructing separate phylogenies through Bayesian analyses for MSANTD1/MSANTD2 and MSANTD3/MSANTD4/NAIF1, respectively, we observed that orthologous MSANTD sequences formed monophyletic groups and that for each of them the closest *Harbinger* transposon was different. However, Maximum Likelihood analysis failed to support some of these preferential phylogenetic relationships between MSANTD genes and *Harbinger* transposons. Taken together, we propose that the vertebrate family of *Myb-like* genes derived from *Harbinger* transposons originated from three to five independent molecular domestication events. Three to four domestications probably occurred at the base of jawed vertebrates about 500 Mya, and a more recent might have led to the formation of MSANTD3 at the base of sarcopterygians approx. 430 Mya.

Interestingly, after comparing 675 human proteins annotated as containing a *Myb*-related domain with *Harbinger* *Myb-like* proteins (data not shown), we identified four additional putative cases of *Harbinger*-derived genes in the human genome: *MYPOP*, *ZSCAN20*, *PRDM11* and *TSNARE1*, *TSNARE1* having already been suggested to originate from a *Harbinger* transposon (Smith et al. 2012). Even if these genes require further analysis, this suggests the presence of additional genes derived from the *Myb-like* genes of *Harbinger* transposons in vertebrate genomes.

A processed pseudogene resulted from a duplication by MSANTD2 mRNA retrotransposition (supplementary fig. 1, Supplementary Material online). Duplicated transcribed pseudogenes can directly regulate related functional genes by transcriptional interference through the production of small interfering RNAs, or by recruiting factors silencing the protein-coding gene transcript (Sen and Ghosh 2013). However, MSANTD2P1 is not expressed in human brain, and no more information is available on its expression. Its function, if any, remains to be further investigated.

Harbinger DNA transposons have given rise to multiple novel genes in divergent organisms: at least six in

vertebrates, nine in *Arabidopsis* and seven in *Drosophila* (Kapitonov and Jurka 2004; Casola et al. 2007; Sinzelle et al. 2008; Liang et al. 2015; Duan et al. 2017; Velanis et al. 2020; Zhou et al. 2021). Hence, it seems that these elements have high propensity to be recruited as new genes. The characteristics of *Harbinger* transposons, with their two protein-coding open reading frames (ORFs), may be an advantage. These two ORFs encode proteins with domains with widespread functions. Particularly, the *Myb-like* domain, a DNA- and protein-binding domain, could be repurposed in diverse ways for gene regulation as a transcription factor and/or as a member of a protein interactome (Sinzelle et al. 2008). In addition, separation of the different molecular activities, i.e., DNA breaking/recombination and DNA/protein binding, in two independent ORFs is uncommon for TEs. This might allow a more specific co-option by the host of a single molecular activity without interference of the other.

Harbinger-Derived MSANTD Genes Encode Potential DNA- and Protein-Binding Proteins

We observed that the secondary structure (tri-helix motif) of the *Harbinger* *Myb-like* protein has been conserved in the different MSANTD proteins, suggesting conservation of the original molecular properties (fig. 1). Generally, the *SANT/myb/trihelix* motifs have been shown to have DNA- and protein-binding capacities in multiple transcription factors (Boyer et al. 2004). The *Harbinger* *Myb-like* protein is able to bind both the transposon DNA and the transposase protein (Sinzelle et al. 2008). Thus, the MSANTD proteins could act as DNA- and protein-interactors. Accordingly, NAIF1, like the *Harbinger* *Myb-like* protein, is able to bind DNA and interact with the *Harbinger* transposase as well as with the transposase-derived HARB11 protein (Sinzelle et al. 2008). MSANTD3 has been suggested to work as a transcription factor that binds to DNA, where it can recruit the Polycomb Repressive Complex 2 to regulate neuronal differentiation in P19 mouse cells (Gou 2014). Outside of vertebrates, other genes derived from *Harbinger* transposons have been identified in *Arabidopsis* (Liang et al. 2015; Duan et al. 2017; Velanis et al. 2020; Zhou et al. 2021). ALP1 (Antagonist of Like Heterochromatin Protein 1), its paralog HHP1 (HDA6-associated *Harbinger* transposon-derived Protein 1) and HDP1 (*Harbinger* transposon-derived protein 1) are derived from transposases, while ALP2 (Antagonist of Like Heterochromatin Protein 2), HDP2 (*Harbinger* transposon-Derived Protein 2), SANT1, SANT2, SANT3, and SANT4 have been formed from *Harbinger* *Myb-like* genes. ALP2 and HDP2 interact with ALP1 and HDP1, respectively, and are involved in chromatin modifying complexes (Liang et al. 2015; Duan et al. 2017; Velanis et al. 2020; Zhou et al. 2021). ALP1 and ALP2 mediate Polycomb Repressive Complex 2 formation (Velanis et al. 2020). HDP1 and HDP2 are part of a histone acetyltransferase complex acting in DNA methylation through the DNA-binding capacity of HDP2 (Duan et al.

2017). Similarly, HHP1, SANT1, SANT2, SANT3, and SANT4 belong to a HDA6 histone deacetylase complex controlling flowering time (Zhou et al. 2021).

Overall, multiple genes derived from *Harbinger* transposons encode proteins that have kept the DNA- and protein-binding capacities ancestrally present in the transposon Myb-like proteins. Therefore, the *MSANTD* genes identified in this study may encode transcription factors or other proteins with DNA- and protein-binding activities.

Harbinger-Deriving Genes Are Expressed in Developing and Adult Vertebrate Brain

Expression results indicated that *Harbinger*-derived genes are expressed in zebrafish during embryonic development, particularly before the MBT for most of them, suggesting potential maternal effect. These genes are also transcribed in adult tissues. We observed that *HARBI1* is generally expressed at a higher level than the *MSANTD* genes. This could favor *HARBI1* interaction with multiple *MSANTD* proteins, as demonstrated with *NAIF1* (Sinzelle et al. 2008), particularly in the brain. *HARBI1* might also have *MSANTD*-independent functions, as suggested by the absence of co-expression in some other tissues.

Harbinger-derived genes are expressed in multiple tissues in zebrafish (fig. 4) and human (GTEx Consortium 2013). However, as observed in zebrafish adult male brain, we detected a common expression of *Harbinger*-derived genes in human brain particularly during early fetal development (figs. 4B and 5), which might favor functional interactions of their proteins in this organ in vertebrates. *Harbinger*-derived genes are predominantly expressed from 8–9 pcw through the two first trimesters of fetal development. Around 8–9 pcw, a process called neuronal migration starts in fetal brain (Métin et al. 2008; Rahimi-Balaei et al. 2018). Neurons are formed in the neuroepithelium, a neural tube layer, during embryonic development. Neuronal migration corresponds to the processes by which neurons will migrate from their germinal layer to all over the central nervous system, where they will establish connections with other cells. As more and more neurons migrate to their final localization, the different brain structures start to be formed throughout the first and second trimesters of fetal development. Disturbance of neuronal migration has been associated to neurological disorders such as schizophrenia, autism spectrum disorders and epilepsy (Fatemi 2005; Guerrini and Parrini 2010; Muraki and Tanigaki 2015; Pan et al. 2019).

MSANTD1 presents a striatum-specific expression during the second trimester of fetal development in human. The striatum is part of the basal ganglia brain structure, mainly involved in voluntary motor control and related to rewards in social conditions (Báez-Mendoza and Schultz 2013). The general role of the basal ganglia on movement control is conserved in vertebrates (Grillner et al. 2013).

Together, the redundant expression in zebrafish and human brain suggests the potential implication of *Harbinger*-derived genes in vertebrate nervous system

development, potentially in neuronal migration (fig. 5). This is also compatible with works suggesting association of *MSANTD2* to schizophrenia and autism spectrum disorders in human (Schizophrenia Working Group of the Psychiatric Genomics Consortium 2014; Lim et al. 2017; O'Brien et al. 2018; Zhang et al. 2020).

MSANTD2, a Gene Involved in Vertebrate Nervous System Development

In order to better understand the biological roles of *Harbinger*-derived genes in vertebrates, we have further analyzed the effects of the inactivation of the *MSANTD2* gene in zebrafish. Expression analyses revealed *MSANTD2* expression in brain during development in human but also in zebrafish at 24hpf (figs. 4 and 5). This suggested a possible function of *MSANTD2* in vertebrate nervous system development.

Inactivation of *MSANTD2* by CRISPR/Cas9-direct gene knock-out in zebrafish produced embryos with severe developmental delays as well as tail and nervous system malformations (fig. 6A–F). We identified defects in neural tube folding, resulting in impaired ventricle formation in forebrain, midbrain and sometimes hindbrain regions (fig. 6H–O). These structural malformations were linked to cellular defects, as we observed accumulation of dead cells and multiple abnormalities in neuronal marking from 24hpf that lasted at least until 72hpf.

In *MSANTD2* CRISPR/Cas9 embryos we observed modified expression patterns for the *FGF8*, *DLX2*, *PAX2A* and *HER5* genes, which are involved in vertebrate nervous system development. These results revealed brain, and particularly MHB organization defects. Moreover, we found accumulation of dead cells in *MSANTD2* CRISPR/Cas9 embryos (fig. 7A–D). The altered expression of *DLX2*, a gene involved in cranial migratory neural crest cell development, suggested that dead cell accumulation could correspond to neural crest cells. Neural crest cells contribute to multiple cell lineages, including sensory and automatic neurons, glia cells, pigment cells and chondrocytes (Iulianella and Trainor 2003). In zebrafish, cranial neural crest cell migration starts around 13hpf (Rocha et al. 2020). In the Tg(*sox10*:mRFP) transgenic line, in which *SOX10* was fluorescently labeled, we observed defects in *SOX10* marking in *MSANTD2* CRISPR/Cas9 embryos at both 15hpf and 19hpf, suggesting migratory delays or defects of cranial neural crest cells (supplementary fig. 8, Supplementary Material online). Furthermore, the expression of *MSANTD2* in human brain during fetal development in a time lapse where neuronal migration arises, as well as the aberrant pattern of early neurons in *MSANTD2* CRISPR/Cas9 zebrafish embryos, might support a role of *MSANTD2* in neural crest cell or neuron migration.

The phenotypes observed in our analysis correspond to *MSANTD2* F0 generation mutants. Reproducible phenotypes were obtained with different combinations of sgRNAs as well as with morpholino oligonucleotides in a gene knock-down approach. CRISPR-Cas9 mutant phenotypes were rescued by *MSANTD2* mRNA. Finally,

inactivation of other *Harbinger*-derived genes in zebrafish did not produce similar phenotypes, indicating specificity of the phenotypes observed for *MSANTD2*. Hence, in addition to the strong mortality of *MSANTD2* mutated embryos, these results strongly support a role for this gene in the development of the vertebrate nervous system.

Conclusion

Vertebrate early evolution has been marked by the emergence of multiple major innovations, which have contributed to the evolutionary success of this lineage. Indeed, vertebrates present new and complex organs, which have allowed the improvement of their movement, sensing and adaptation to their environment. For example, vertebrates have a complex nervous system, which is composed of cranial nerves, spinal cord, ganglia and a brain organized in specialized regions. Bones, cartilages, paired appendages, a complex endocrine system, sensory placodes, the neural crest and an adaptive immune system are also major novelties acquired during early vertebrate evolution.

Ohno proposed that whole genome duplications, generating an extensive expansion of gene repertoires, are major events giving rise to massive innovations and important evolutionary transitions (Ohno 1999). Accordingly, two events of genome duplications have taken place at the base of vertebrates (Dehal and Boore 2005). However, new gene formation by duplication is not the unique mechanism allowing the apparition of major novelties. *SYNCYTIN* genes, involved in placenta formation in mammals, as well as *RAG* genes, implicated in the adaptive immune system in vertebrates, testify of the role of TE-derived novel genes in organismal innovation.

In this work, we propose that *Harbinger*-derived genes could have been contributors of early vertebrate evolution, notably through their role in the evolution of the nervous system development. Further analyses should look at the implication of other TE molecular domestication events in the emergence and evolution of other vertebrate innovations. Hence, the study of TE molecular domestication provides us with important clues on the functional and evolutionary characteristics of new genes, with a broader picture of the genetic basis and dynamics of the emergence and evolution of phenotypic traits.

Materials and Methods

Zebrafish Maintenance

Zebrafish of the strain AB/TU were raised according to standard procedures (PRECI, SFR Biosciences [UAR3444/CNRS, US8/INSERM, ENS de Lyon, UCBL]). Embryos were raised at 28°C. Developmental stages were expressed in hours post-fertilization (hpf) or days post-fertilization (dpf) based on morphological criteria (Kimmel et al. 1995). The Tg(elavl3:GCaMP6s) transgenic line, containing a modified GCaMP (GCaMP is a genetically-encoded calcium indicator) known as GCaMP3 under elavl3 regulatory region (ZFIN ID: ZDB-TGCONSTRUCT-180326-1), were also

raised according to the same procedures (Park et al. 2000; Panier et al. 2013). The zebrafish Tg(sox10:mRFP) transgenic line, obtained from Florence Ruggiero team (Institute of Functional Genomics, Lyon, France), was also raised according to the same standard procedures. The SOX10 protein, which is expressed in neural crest cells, is fluorescently marked in this transgenic line.

In Situ Hybridization

In situ hybridization (ISH) probes for *MSANTD2* were cloned from WT zebrafish cDNA by PCR using the GoTaq polymerase (Promega). *PAX2A* and *HER5* probes were given by Dr. Sebastian Dworkin lab, La Trobe University, Melbourne, Australia.

Zebrafish embryos were collected, removed from their chorion, sorted and fixed in paraformaldehyde (PFA) 4%, dehydrated in methanol and stored at -20°C. ISH was performed following the Thisse Lab protocol (Thisse and Thisse 2008; 2014). Embryos were rehydrated and washed in phosphate buffered saline (PBS)—Tween (PBT) solution. They were permeabilized with proteinase K and fixed in PFA. Each embryo was incubated with probes overnight at 65°C in hybridization mix supplemented with 5% Dextran Sulfate (Millipore). Nonhybridized probes were removed with several washes in formamide and saline-sodium-citrate solutions. Embryos were incubated overnight at 4°C with α -DIG (digoxigenin) antibodies (Roche). Nonfixed antibodies were removed with PBT washes. Probes were revealed with nitro blue tetrazolium chloride - 5-bromo-4-chloro-3-indolyl-phosphate (NBT-BCIP) (Roche). Embryos were fixed in PFA. After removing of the background with ethanol bath, embryos were stored in glycerol 80% at 4°C. Pictures were taken under Leica stereomicroscope and Keyence VHX-7000 microscope.

qPCR

Pools of 3–5 zebrafish adults and 15–20 embryos were used for RNA extraction. RNAs were extracted with Trizol according to the Bio-Rad company protocol and treated with DNaseI. Reverse transcription was performed using the RevertAid First Strand cDNA Synthesis Kit (Thermo Scientific). The following specific primers were designed: for *NAIF1* TGAAT CACTTTAACGCGGGC, CCGTCTTCAGATCCGACCAT; for *HARBI1* CGCTGCGTTTCTAACGTAC, AGAGTCATCCGCA TTGGGAG; for *MSANTD1* CAAACCTCTCATCGTCTGGC, AGGCCGTATCCTCATCATT; for *MSANTD2* AGACCCGAG TTCTTCAGATACGAC, GAGAGAAGTCCGTCCACGTTTG; for *MSANTD4* TCAAGATGGAGGACGACGAG, GGGAGGA TGGAGGGAAAACA. qPCR was performed using SYBR Green following the Bio-Rad protocol. 18S housekeeping ribosomal RNA gene (TCGCTAGTTGGCATCGTTTATG, CGGAGGTTCAAGACGATCA) was used to normalize gene expression. Results were analyzed with the Δ Ct method (Schmittgen and Livak 2008).

Morpholino Knockdown

Two nonoverlapping morpholino antisense oligonucleotides targeting the 5'-UTR of *MSANTD2* (GCCATCTTGC

TTCTGTTGCTAAGGG, CAGACACGACTGACGGCTTCT TATG) and a control mismatched morpholino (CACACA CCAGTGACGCCTTGTATG) were purchased from Gene Tools and injected from 0.2M to 3M into one-cell embryos (Nasevicius and Ekker 2000). Morphants and mismatched controls were matched per cross. Morphological and phenotypic observations were performed at 1dpf under Zeiss Axio Zoom microscope.

CRISPR/Cas9

For each gene four nonoverlapping single guide RNAs (sgRNAs) were purchased from Synthego (supplementary fig. 2, Supplementary Material online). The sequences of the sgRNAs were selected from Wu et al. 2018 (Wu et al. 2018). For all sgRNAs used, no putative off target site with four or less mismatches and next to a PAM could be detected in the zebrafish genome using the CRISPOR program (Concordet and Haeussler 2018). The Cas9-GFP protein was purchased from TacGene. A mix of four sgRNAs (20 to 30 μ M in total) and Cas9-GFP protein (5 to 15 μ M) was injected into WT embryos at the one-cell stage (individual sgRNAs as well as combinations of two sgRNAs were also injected). For *MSANTD2*, sgRNA 1 and sgRNA 2 were located in the first exon and sgRNA 3 and sgRNA 4 in the third and fourth exons, respectively. For each gene knock-out, the experiment was performed at least in duplicate (at least six times for the mix of 4 sgRNAs). Scrambled (random sequence) sgRNAs were used as a negative control and sgRNAs targeting the tyrosinase (*TYR*) gene as a positive control (its inactivation led to individuals without melanic pigmentation). Crispants and scrambled controls were matched per cross. Embryo survival and phenotypic observations were monitored from 6hpf under Leica stereomicroscope and Zeiss Axio Zoom microscope.

MSANTD2 Mutant Rescue Experiment

MSANTD2 mRNA was transcribed in vitro (mMessage Machine SP6 Transcription Kit, Ambion) from the corresponding cDNA sequences in the pcS2 + vector synthesized by GenScript. For microinjections, the Cas9-sgRNAs mix was combined in a 1:1 volume ratio with *MSANTD2* mRNA at a 100 ng/ μ l concentration. The presence of mutations at the *MSANTD2* sgRNA loci in the rescued embryos was verified by Sanger sequencing. For that, injected embryos were collected and DNA extractions were conducted for the three conditions (*MSANTD2* sgRNA 1 + 2 + 3 + 4, *MSANTD2* sgRNA 1 + 2 + 3 + 4 + *MSANTD2* mRNA, scrambled). Multiple pics (mutations) were observed after sequencing at sgRNA loci for the *MSANTD2* sgRNA 1 + 2 + 3 + 4 and *MSANTD2* sgRNA 1 + 2 + 3 + 4 + *MSANTD2* mRNA conditions compared to the scrambled condition, for which only unique pics were observed (no mutations).

Brain Ventricle Imaging

Zebrafish brain ventricle injection was performed according to the protocol developed by Gutzman and Sive (2009). Briefly, embryos were anesthetized with Tricaine

(Sigma). Micro-injection was performed in hindbrain ventricle with 1–10nl of dextran Texas Red (5% in 0.2mol/L KCl, Invitrogen). 15 to 30min after injection, images were taken with transmitted and fluorescent lights under a Zeiss AxioZoom Microscope.

Acridine Orange Staining

Embryos were dechorionated and stained with 10 μ g/mL acridine orange solution for 30 min. Then, embryos were washed three times in E3 medium. Images were taken with transmitted and fluorescent lights under a Zeiss AxioZoom Microscope.

DNA Extraction, PCR Amplification, NGS Sequencing and Sequencing Data Analyses

In order to search for mutations after application of the CRISPR/Cas9 protocol, injected embryos were collected at 24hpf and five DNA extraction replicates were conducted starting from a single embryo for the four conditions (4sgRNAs, sgRNA1–4, sgRNA4 and scrambled). Lysis of embryos was performed in lysis buffer (10mM Tris-HCl pH8–2 mM EDTA pH8–0.2% Triton x-100) with 250 μ g/ μ l proteinase K (Invitrogen) 12 h at 55°C, followed by proteinase K inactivation of 10 min at 95°C. Three fragments of the *MSANTD2* gene (exon 1, exon 3 and exon 4) were amplified by PCR. PCR reactions were performed in 25 μ l using the GoTaq G2 DNA polymerase kit (Promega), 2 μ l of DNA extract and 0.5 μ M of each primer set with the following PCR program: 2 min at 94°C, 35 cycles at 94°C 30 s, 60°C 30 s and 72°C 30 s, with a final extension step at 72°C for 5 min. For each condition and for each exon, five PCR tubes (each PCR corresponding to one embryo DNA amplification) were pooled. PCR product purification was carried out according to manufacturer's recommendations (Nucleospin Gel and PCR Clean-up, Macherey Nagel) and eluted in 30 μ l of elution buffer (NE buffer). For each condition, equimolar amounts of the three purified amplicons were used to create a bar-coded library with an input of 50ng using the NEBNext Ultra II DNA Library Prep Kit protocol for Illumina. Quantitation and quality assessment of each library was performed on a 4150 TapeStation analyzer using the High Sensitivity D5000 ScreenTape kit (Agilent Technologies). Libraries were mixed at the same equimolar proportions, spiked with approximately 5% PhiX control and sequenced with the Illumina MiSeq sequencer using the Nano Kit v2 reagent (pair-end reads, R1 and R2 read lengths, 260 bp and 259 bp respectively). More than 800K reads were obtained and analyzed using the Galaxy platform (Afgan et al. 2018) using the FastQC, Cutadapt, Bowtie2 and Sort tools to assess the quality of reads, remove adapter sequences, map reads against reference and store aligned sequences, respectively.

TE and Gene Sequence *in Silico* Analyses

TE-derived genes were identified through sequence similarity with TE sequences from the Repbase database

(www.girinst.org) and from annotation of various sequenced vertebrate genomes (Chalopin et al. 2015) using blastp, blastn and tblastn (Altschul et al. 1990). Additional *Harbinger* Myb-like protein sequences were recovered through blast analysis of the NCBI Genomes (RefSeq Genomes) database (www.ncbi.nlm.nih.gov) using MSANTD and *Harbinger* transposon sequences as queries. Blast was used with default parameters. A first threshold was set for protein-protein comparisons (E -value $<10e-5$), which was followed by manual inspection of sequence alignments on the whole predicted sequences, inspection of predicted conserved protein motifs and construction of molecular sequence phylogenies (see below). NCBI, Ensembl, Censor (www.girinst.org) and Genomicus (Muffato et al. 2010) were used to determine the copy number, sequence alignments, phylogeny and synteny of TE-derived genes. Conserved protein motifs were detected with the NCBI Conserved Domain Search (Marchler-Bauer et al. 2011), InterPro (Apweiler et al. 2000) and PROSITE (Sigrist et al. 2002). Genes and ORFs were predicted with Augustus (Stanke and Morgenstern 2005) and ORFfinder (www.ncbi.nlm.nih.gov/orffinder). NPS-PRABI (Combet et al. 2000) and Jpred4 (Drozdetskiy et al. 2015) were used to predict the secondary structure of proteins. For positive selection tests, the protein-coding sequence of genes from different species were collected from the Ensembl database, aligned as proteins using MUSCLE (Edgar 2004) and then converted back into a nucleic sequence alignment. A phylogenetic tree was then built with the PhyML package (see below). Positive/negative selection tests were performed using CODEML (Yang 2007). The tests were run based on an alignment of coding sequences from spotted gar, zebrafish, tetraodon, stickleback, platyfish, coelacanth, chinese soft-shell turtle, mouse, macaca, marmoset, human, chimpanzee and chicken.

For phylogenetic analysis, nucleotide and amino-acid sequences were aligned with MAFFT (Katoh et al. 2002). Phylogenetic trees were built using maximum likelihood with PhyML (Guindon and Gascuel 2003) and with MrBayes (Huelsenbeck and Ronquist 2001) using a mixed model (estimated by the Prottest-3 software (Darriba et al. 2011) and 500,000 generations of Bayesian inferences.

Gene accessions numbers are HGNC:33741, HGNC:26266, HGNC:23370, HGNC:29383, HGNC:25446 and HGNC:26522 for *MSANTD1*, *MSANTD2*, *MSANTD3*, *MSANTD4*, *NAIF1* and *HARBI1*, respectively. Transposon sequences and alignments are available upon request.

Supplementary Material

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

Acknowledgments

This work was supported by grants from the Agence Nationale de la Recherche (ANR) and from the Ecole

Normale Supérieure de Lyon. We acknowledge the contribution of the PRECI fish facility of the SFR Biosciences (UAR3444/CNRS, US8/INSERM, ENS de Lyon, UCBL) and of the IGFL's PSI Sequencing platform. We are grateful to Marilyne Malbouyres, Sandrine Bretaud and Florence Ruggiero (Matrix biology and pathology team, Institut de Génomique Fonctionnelle de Lyon) for their help in Crispr-Cas9 methodology, and to Marie Sémon (ENS Lyon) and Christoph Winkler (University of Singapore) for discussions. EE thanks the 'Fondation pour la Recherche Médicale' (FRM) for financial support through an end of PhD program fellowship.

Author Contributions

Experiments were designed by E.E. and J.N.V. and performed by E.E., manuscript was drafted by E.E. and amended by J.N.V., the project was supervised by J.N.V. and co-supervised by D.B., M.N. and Z.H.T.

Data Availability

Data are available on request.

References

- GTE Consortium. 2013. The genotype-tissue expression (GTEx) project. *Nat Genet.* **45**:580–585.
- Schizophrenia Working Group of the Psychiatric Genomics Consortium. 2014. Biological insights from 108 schizophrenia-associated genetic loci. *Nature* **511**:421–427.
- Afgan E, Baker D, Batut B, van den Beek M, Bouvier D, Čech M, Chilton J, Clements D, Coraor N, Grüning BA, et al. 2018. The galaxy platform for accessible, reproducible and collaborative bio-medical analyses: 2018 update. *Nucleic Acids Res.* **46**:W537–W544.
- Akimenko MA, Ekker M, Wegner J, Lin W, Westerfield M. 1994. Combinatorial expression of three zebrafish genes related to distal-less: part of a homeobox gene code for the head. *J Neurosci.* **14**:3475–3486.
- Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *J Mol Biol.* **215**:403–410.
- Alzohairy AM, Gyulai G, Jansen RK, Bahieldin A. 2013. Transposable elements domesticated and neofunctionalized by eukaryotic genomes. *Plasmid.* **69**:1–15.
- Apweiler R, Attwood TK, Bairoch A, Bateman A, Birney E, Biswas M, Bucher P, Cerutti L, Corpet F, Croning MDR, et al. 2000. Interpro—an integrated documentation resource for protein families, domains and functional sites. *Bioinformatics.* **16**:1145–1150.
- Báez-Mendoza R, Schultz W. 2013. The role of the striatum in social behavior. *Front Neurosci.* **7**:233.
- Boyer LA, Latek RR, Peterson CL. 2004. The SANT domain: a unique histone-tail-binding module? *Nat Rev Mol Cell Biol.* **5**:158–163.
- Brandt J, Schrauth S, Veith A-M, Froschauer A, Haneke T, Schultheis C, Gessler M, Leimeister C, Volff J-N. 2005. Transposable elements as a source of genetic innovation: expression and evolution of a family of retrotransposon-derived neogenes in mammals. *Gene.* **345**:101–111.
- Britten R. 2006. Transposable elements have contributed to thousands of human proteins. *Proc Natl Acad Sci U S A.* **103**:1798–1803.
- Casola C, Lawing AM, Betrán E, Feschotte C. 2007. PIF-like transposons are common in drosophila and have been repeatedly

- domesticated to generate new host genes. *Mol Biol Evol.* **24**: 1872–1888.
- Chalopin D, Naville M, Plard F, Galiana D, Volff J-N. 2015. Comparative analysis of transposable elements highlights mobile diversity and evolution in vertebrates. *Genome Biol Evol.* **7**: 567–580.
- Chi CL, Martinez S, Wurst W, Martin GR. 2003. The isthmic organizer signal FGF8 is required for cell survival in the prospective mid-brain and cerebellum. *Development* **130**:2633–2644.
- Chuong EB, Elde NC, Feschotte C. 2017. Regulatory activities of transposable elements: from conflicts to benefits. *Nat Rev Genet.* **18**: 71–86.
- Combet C, Blanchet C, Geourjon C, Deléage G. 2000. NPS@: network protein sequence analysis. *Trends Biochem Sci.* **25**:147–150.
- Concordet J-P, Haeussler M. 2018. CRISPOR: intuitive guide selection for CRISPR/Cas9 genome editing experiments and screens. *Nucleic Acids Res.* **46**:W242–W245.
- Dai J, Kuang Y, Fang B, Gong H, Lu S, Mou Z, Sun H, Dong Y, Lu J, Zhang W, et al. 2013. The effect of overexpression of Dlx2 on the migration, proliferation and osteogenic differentiation of cranial neural crest stem cells. *Biomaterials* **34**:1898–1910.
- Darriba D, Taboada GL, Doallo R, Posada D. 2011. Prottest 3: fast selection of best-fit models of protein evolution. *Bioinformatics* **27**: 1164–1165.
- Dehal P, Boore JL. 2005. Two rounds of whole genome duplication in the ancestral vertebrate. *PLoS Biol.* **3**:e314.
- Doolittle WF, Sapienza C. 1980. Selfish genes, the phenotype paradigm and genome evolution. *Nature* **284**:601–603.
- Drozdetskiy A, Cole C, Procter J, Barton GJ. 2015. JPred4: a protein secondary structure prediction server. *Nucleic Acids Res.* **43**: W389–W394.
- Duan C-G, Wang X, Xie S, Pan L, Miki D, Tang K, Hsu C-C, Lei M, Zhong Y, Hou Y-J, et al. 2017. A pair of transposon-derived proteins function in a histone acetyltransferase complex for active DNA demethylation. *Cell Res.* **27**:226–240.
- Dupressoir A, Vernochet C, Harper F, Guégan J, Dessen P, Pierron G, Heidmann T. 2011. A pair of co-opted retroviral envelope syncytin genes is required for formation of the two-layered murine placental syncytiotrophoblast. *Proc Natl Acad Sci U S A.* **108**: E1164–1173.
- Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32**:1792–1797.
- Etchegaray E, Naville M, Volff J-N, Haftek-Terreau Z. 2021. Transposable element-derived sequences in vertebrate development. *Mob DNA* **12**:1.
- Fatemi SH. 2005. Reelin glycoprotein: structure, biology and roles in health and disease. *Mol Psychiatry* **10**:251–257.
- Fu Y, Cao F. 2015. MicroRNA-125a-5p regulates cancer cell proliferation and migration through NAIF1 in prostate carcinoma. *Oncotargets Ther.* **8**:3827–3835.
- Gou Y. 2014. MSANTD3, a novel transcription factor, recruits PRC2 complex to regulate neuron differentiation in mouse P19 cells. [cited 2022 Jan 10]. Available from: <https://scholarship.rice.edu/handle/1911/87825>
- Grillner S, Robertson B, Stephenson-Jones M. 2013. The evolutionary origin of the vertebrate basal ganglia and its role in action selection. *J Physiol.* **591**:5425–5431.
- Guerrini R, Parrini E. 2010. Neuronal migration disorders. *Neurobiol Dis.* **38**:154–166.
- Guindon S, Gascuel O. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst Biol.* **52**:696–704.
- Gutzman JH, Sive H. 2009. Zebrafish brain ventricle injection. *J Vis Exp.* **26**:1218.
- Hancock CN, Zhang F, Wessler SR. 2010. Transposition of the tourist-MITE mPing in yeast: an assay that retains key features of catalysis by the class 2 PIF/harbinger superfamily. *Mob DNA* **1**:5.
- Haniffa M, Taylor D, Linnarsson S, Aronow BJ, Bader GD, Barker RA, Camara PG, Camp JG, Chédotal A, Copp A, et al. 2021. A roadmap for the human developmental cell atlas. *Nature* **597**: 196–205.
- Huelsenbeck JP, Ronquist F. 2001. MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics* **17**:754–755.
- Iulianella A, Trainor PA. 2003. Hox gene control of neural crest cell, pharyngeal arch and craniofacial patterning. In: *Advances in developmental biology and biochemistry*. Vol. 13. Murine Homeobox Gene Control of Embryonic Patterning and Organogenesis. New York (NY): Elsevier. p. 155–206. Available from: <https://www.sciencedirect.com/science/article/pii/S1569179903130067>
- Kaessmann H. 2010. Origins, evolution, and phenotypic impact of new genes. *Genome Res.* **20**:1313–1326.
- Kague E, Gallagher M, Burke S, Parsons M, Franz-Odenaal T, Fisher S. 2012. Skeletogenic fate of zebrafish cranial and trunk neural crest. *PLoS One* **7**:e47394.
- Kapitonov VV, Jurka J. 2004. Harbinger transposons and an ancient HARBI1 gene derived from a transposase. *DNA Cell Biol.* **23**: 311–324.
- Kapitonov VV, Koonin EV. 2015. Evolution of the RAG1-RAG2 locus: both proteins came from the same transposon. *Biol Direct.* **10**:20.
- 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.
- Kidwell MG, Lisch DR. 2000. Transposable elements and host genome evolution. *Trends Ecol Evol.* **15**:95–99.
- Kimmel CB, Ballard WW, Kimmel SR, Ullmann B, Schilling TF. 1995. Stages of embryonic development of the zebrafish. *Dev Dyn.* **203**: 253–310.
- Knowles DG, McLysaght A. 2009. Recent de novo origin of human protein-coding genes. *Genome Res.* **19**:1752–1759.
- Kong D, Zhang Z. 2018. NAIF1 suppresses osteosarcoma progression and is regulated by miR-128. *Cell Biochem Funct.* **36**:443–449.
- Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, et al. 2001. Initial sequencing and analysis of the human genome. *Nature* **409**:860–921.
- Liang SC, Hartwig B, Perera P, Mora-García S, de Leau E, Thornton H, de Lima Alves F, de Lima Alves F, Rappsilber J, Rapsilber J, et al. 2015. Kicking against the PRCs—a domesticated transposase antagonises silencing mediated by polycomb group proteins and is an accessory component of polycomb repressive Complex 2. *PLoS Genet.* **11**:e1005660.
- Lim ET, Uddin M, De Rubeis S, Chan Y, Kamumbu AS, Zhang X, D’Gama AM, Kim SN, Hill RS, Goldberg AP, et al. 2017. Rates, distribution and implications of postzygotic mosaic mutations in autism spectrum disorder. *Nat Neurosci.* **20**:1217–1224.
- Lu J, Zhu M, Ahlberg PE, Qiao T, Zhu Y, Zhao W, Jia L. 2016. A Devonian predatory fish provides insights into the early evolution of modern sarcopterygians. *Sci Adv.* **2**:e1600154.
- Luo Q, Zhao M, Zhong J, Ma Y, Deng G, Liu J, Wang J, Yuan X, Huang C. 2011. NAIF1 is down-regulated in gastric cancer and promotes apoptosis through the caspase-9 pathway in human MKN45 cells. *Oncol Rep.* **25**:1117–1123.
- Lv B, Shi T, Wang X, Song Q, Zhang Y, Shen Y, Ma D, Lou Y. 2006. Overexpression of the novel human gene, nuclear apoptosis-inducing factor 1, induces apoptosis. *Int J Biochem Cell Biol.* **38**:671–683.
- Lynch M, Conery JS. 2000. The evolutionary fate and consequences of duplicate genes. *Science* **290**:1151–1155.
- Mallet F, Bouton O, Prudhomme S, Cheynet V, Oriol G, Bonnaud B, Lucotte G, Duret L, Mandrand B. 2004. The endogenous retroviral locus ERVWE1 is a bona fide gene involved in hominoid placental physiology. *Proc Natl Acad Sci U S A.* **101**:1731–1736.
- Marchler-Bauer A, Lu S, Anderson JB, Chitsaz F, Derbyshire MK, DeWeese-Scott C, Fong JH, Geer LY, Geer RC, Gonzales NR, et al. 2011. CDD: a conserved domain database for the functional annotation of proteins. *Nucleic Acids Res.* **39**:D225–D229.
- Métin C, Vallee RB, Rakic P, Bhide PG. 2008. Modes and mishaps of neuronal migration in the mammalian brain. *J Neurosci.* **28**: 11746–11752.

- Miller JA, Ding S-L, Sunkin SM, Smith KA, Ng L, Szafer A, Ebbert A, Riley ZL, Royall JJ, Aiona K, et al. 2014. Transcriptional landscape of the prenatal human brain. *Nature* **508**:199–206.
- Moran JV, Malik HS. 2009. Diamonds and rust: how transposable elements influence mammalian genomes. Conference on mobile elements in mammalian genomes. *EMBO Rep.* **10**:1306–1310.
- Muffato M, Louis A, Poinsnel C-E, Roest Crolius H. 2010. Genomicus: a database and a browser to study gene synteny in modern and ancestral genomes. *Bioinformatics* **26**:1119–1121.
- Muraki K, Tanigaki K. 2015. Neuronal migration abnormalities and its possible implications for schizophrenia. *Front Neurosci.* **9**:74.
- Nakamura H. 2001. Regionalization of the optic tectum: combinations of gene expression that define the tectum. *Trends Neurosci.* **24**:32–39.
- Nasevicius A, Ekker SC. 2000. Effective targeted gene “knockdown” in zebrafish. *Nat Genet.* **26**:216–220.
- O’Brien HE, Hannon E, Hill MJ, Toste CC, Robertson MJ, Morgan JE, McLaughlin G, Lewis CM, Schalkwyk LC, Hall LS, et al. 2018. Expression quantitative trait loci in the developing human brain and their enrichment in neuropsychiatric disorders. *Genome Biol.* **19**:194.
- Ohno S. 1972. So much “junk” DNA in our genome. *Brookhaven Symp Biol.* **23**:366–370.
- Ohno S. 1999. Gene duplication and the uniqueness of vertebrate genomes circa 1970–1999. *Semin Cell Dev Biol.* **10**:517–522.
- Orgel LE, Crick FH. 1980. Selfish DNA: the ultimate parasite. *Nature* **284**:604–607.
- Pan Y-H, Wu N, Yuan X-B. 2019. Toward a better understanding of neuronal migration deficits in autism Spectrum disorders. *Front Cell Dev Biol.* **7**:205.
- Panier T, Romano SA, Olive R, Pietri T, Sumbre G, Candelier R, Debrégeas G. 2013. Fast functional imaging of multiple brain regions in intact zebrafish larvae using selective plane illumination microscopy. *Front Neural Circuits.* **7**:65.
- Park HC, Kim CH, Bae YK, Yeo SY, Kim SH, Hong SK, Shin J, Yoo KW, Hibi M, Hirano T, et al. 2000. Analysis of upstream elements in the HuC promoter leads to the establishment of transgenic zebrafish with fluorescent neurons. *Dev Biol.* **227**:279–293.
- Perelman P, Johnson WE, Roos C, Seuánez HN, Horvath JE, Moreira MAM, Kessing B, Pontius J, Roelke M, Rumpler Y, et al. 2011. A molecular phylogeny of living primates. *PLoS Genet.* **7**:e1001342.
- Piotrowski T, Nüsslein-Volhard C. 2000. The endoderm plays an important role in patterning the segmented pharyngeal region in zebrafish (*Danio rerio*). *Dev Biol.* **225**:339–356.
- Rahimi-Balaei M, Bergen H, Kong J, Marzban H. 2018. Neuronal migration during development of the cerebellum. *Front Cell Neurosci.* **12**:484.
- Rocha M, Singh N, Ahsan K, Beiriger A, Prince VE. 2020. Neural crest development: insights from the zebrafish. *Dev Dyn.* **249**:88–111.
- Schilling TF, Kimmel CB. 1994. Segment and cell type lineage restrictions during pharyngeal arch development in the zebrafish embryo. *Development* **120**:483–494.
- Schmittgen TD, Livak KJ. 2008. Analyzing real-time PCR data by the comparative C(T) method. *Nat Protoc.* **3**:1101–1108.
- Sen K, Ghosh TC. 2013. Pseudogenes and their composers: delving in the “debris” of human genome. *Brief Funct Genomics* **12**:536–547.
- Sigrist CJA, Cerutti L, Hulo N, Gattiker A, Falquet L, Pagni M, Bairoch A, Bucher P. 2002. PROSITE: a documented database using patterns and profiles as motif descriptors. *Brief Bioinform.* **3**:265–274.
- Sinzelle L, Kapitonov VV, Grzela DP, Jursch T, Jurka J, Izsvák Z, Ivics Z. 2008. Transposition of a reconstructed harbinger element in human cells and functional homology with two transposon-derived cellular genes. *Proc Natl Acad Sci U S A.* **105**:4715–4720.
- Smith JJ, Sumiyama K, Amemiya CT. 2012. A living fossil in the genome of a living fossil: harbinger transposons in the coelacanth genome. *Mol Biol Evol.* **29**:985–993.
- Sperber SM, Saxena V, Hatch G, Ekker M. 2008. Zebrafish *dlx2a* contributes to hindbrain neural crest survival, is necessary for differentiation of sensory ganglia and functions with *dlx1a* in maturation of the arch cartilage elements. *Dev Biol.* **314**:59–70.
- Stanke M, Morgenstern B. 2005. AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints. *Nucleic Acids Res.* **33**:W465–W467.
- Thisse C, Thisse B. 2008. High-resolution in situ hybridization to whole-mount zebrafish embryos. *Nat Protoc.* **3**:59–69.
- Thisse B, Thisse C. 2014. In situ hybridization on whole-mount zebrafish embryos and young larvae. *Methods Mol Biol.* **1211**:53–67.
- Toll-Riera M, Bosch N, Bellora N, Castelo R, Armengol L, Estivill X, Mar Albà M. 2009. Origin of primate orphan genes: a comparative genomics approach. *Mol Biol Evol.* **26**:603–612.
- Velanis CN, Perera P, Thomson B, de Leau E, Liang SC, Hartwig B, Förderer A, Thornton H, Arede P, Chen J, et al. 2020. The domesticated transposase ALP2 mediates formation of a novel polycomb protein complex by direct interaction with MS1, a core subunit of polycomb repressive complex 2 (PRC2). *PLoS Genet.* **16**:e1008681.
- Volf J-N. 2006. Turning junk into gold: domestication of transposable elements and the creation of new genes in eukaryotes. *Bioessays* **28**:913–922.
- Warren IA, Naville M, Chalopin D, Levin P, Berger CS, Galiana D, Volf J-N. 2015. Evolutionary impact of transposable elements on genomic diversity and lineage-specific innovation in vertebrates. *Chromosome Res.* **23**:505–531.
- Wolf JBW, Künstner A, Nam K, Jakobsson M, Ellegren H. 2009. Nonlinear dynamics of nonsynonymous (dN) and synonymous (dS) substitution rates affects inference of selection. *Genome Biol Evol.* **1**:308–319.
- Wu RS, Lam II, Clay H, Duong DN, Deo RC, Coughlin SR. 2018. A rapid method for directed gene knockout for screening in G0 zebrafish. *Dev Cell.* **46**:112–125.e4.
- Yan Y-L, Willoughby J, Liu D, Crump JG, Wilson C, Miller CT, Singer A, Kimmel C, Westerfield M, Postlethwait JH. 2005. A pair of *sox*: distinct and overlapping functions of zebrafish *sox9* co-orthologs in craniofacial and pectoral fin development. *Development* **132**:1069–1083.
- Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol.* **24**:1586–1591.
- Yang G, Zhang F, Hancock CN, Wessler SR. 2007. Transposition of the rice miniature inverted repeat transposable element mPing in *Arabidopsis thaliana*. *Proc Natl Acad Sci U S A.* **104**:10962–10967.
- Zhang Y, You X, Li S, Long Q, Zhu Y, Teng Z, Zeng Y. 2020. Peripheral blood leukocyte RNA-seq identifies a set of genes related to abnormal psychomotor behavior characteristics in patients with schizophrenia. *Med Sci Monit.* **26**:e922426.
- Zhao G, Liu L, Zhao T, Jin S, Jiang S, Cao S, Han J, Xin Y, Dong Q, Liu X, et al. 2015. Upregulation of miR-24 promotes cell proliferation by targeting NAIF1 in non-small cell lung cancer. *Tumour Biol.* **36**:3693–3701.
- Zhou X, He J, Velanis CN, Zhu Y, He Y, Tang K, Zhu M, Graser L, de Leau E, Wang X, et al. 2021. A domesticated harbinger transposase forms a complex with HDA6 and promotes histone H3 deacetylation at genes but not TEs in *Arabidopsis*. *J Integr Plant Biol.* **63**:1462–1474.