



HAL
open science

Domestication moléculaire des éléments transposables chez les Vertébrés : étude évolutive et fonctionnelle de gènes dérivés de transposons Harbinger

Ema Etchegaray

► **To cite this version:**

Ema Etchegaray. Domestication moléculaire des éléments transposables chez les Vertébrés : étude évolutive et fonctionnelle de gènes dérivés de transposons Harbinger. Evolution [q-bio.PE]. Université de Lyon, 2022. Français. NNT : 2022LYSEN018 . tel-03783570

HAL Id: tel-03783570

<https://theses.hal.science/tel-03783570v1>

Submitted on 22 Sep 2022

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Numéro National de Thèse : 2022LYSEN018

THÈSE DE DOCTORAT DE L'UNIVERSITÉ DE LYON

opérée par

l'École Normale Supérieure de Lyon

**École doctorale N° 340 :
Biologie Moléculaire, Intégrative et Cellulaire (BMIC)**

Discipline : Sciences de la vie et de la santé

Soutenue publiquement le 9 Juin 2022, par :

Ema ETCHEGARAY

Domestication moléculaire des éléments transposables chez les vertébrés : étude évolutive et fonctionnelle de gènes dérivés de transposons *Harbinger*

Devant le jury composé de :

BÉTERMIER,	Mireille	Directrice de recherche	I2BC, Univ. Paris Saclay	Rapporteuse
GILBERT,	Clément	Chargé de recherche	EGCE, Univ. Paris Saclay	Rapporteur
RUGGIERO,	Florence	Directrice de recherche	IGFL, ENS de Lyon	Examinatrice
CORDAUX,	Richard	Directeur de recherche	EBI, Univ. Poitiers	Examinateur
SÉMON,	Marie	Professeure des Universités	LBMC, ENS de Lyon	Examinatrice
VOLFF,	Jean-Nicolas	Professeur des Universités	IGFL, ENS de Lyon	Directeur de thèse
NAVILLE,	Magali	Docteur en biologie	Lycée Lalande	Co-encadrante

Résumé de la thèse

La formation de nouveaux gènes est une source majeure d'innovation pour les organismes. Au-delà de leurs effets mutagènes, les éléments transposables peuvent être source de nouveaux gènes. J'ai étudié les transposons à ADN *Harbinger*, que j'ai pu caractériser chez les poissons téléostéens. J'ai de plus identifié chez les vertébrés par criblage bioinformatique quatre nouveaux gènes dérivés de transposons *Harbinger*, trois formés à la base des vertébrés à mâchoires il y a 500 millions d'années et un chez les sarcoptérygiens il y a 430 millions d'années. Chez le poisson-zèbre, ces gènes sont exprimés pendant le développement précoce et dans les tissus adultes, avec une co-expression dans le cerveau mâle. Ils sont également activés dans le cerveau humain, en particulier pendant le développement fœtal. Afin d'étudier la fonction des gènes dérivés de transposons *Harbinger*, leur inactivation a été réalisée par CRISPR/Cas9 et oligonucléotides antisens de type morpholino. L'inactivation du gène *MSANTD2*, qui a été associé à des maladies neurodéveloppementales comme l'autisme et la schizophrénie, produit des embryons présentant des retards de développement et des malformations de la queue et du système nerveux, en particulier des défauts de formation des ventricules du cerveau et de patterns neuronaux. Ainsi, cette thèse a permis de mettre en évidence des domestications moléculaires récurrentes de transposons *Harbinger* qui ont abouti à la formation d'une nouvelle famille de gènes chez les vertébrés. L'étude d'un des membres, *MSANTD2*, contribue à une meilleure compréhension des innovations génétiques qui ont déterminé l'évolution précoce du système nerveux des vertébrés.

Just keep swimming...

Dory, Finding nemo

REMERCIEMENTS

Pour commencer, je souhaite remercier Jean-Nicolas. Merci d'avoir cru en moi dès ma L3 et de m'avoir permis de réaliser cette thèse. Merci de m'avoir guidée, en me laissant une grande liberté scientifique et pour ta confiance en mon travail.

Je poursuivrai avec Magali. Merci de m'avoir encadrée et guidée dès mon stage de L3 et jusqu'à la fin de ma thèse. Merci Magali pour ton soutien quotidien dans notre ancien bureau, tes précieux conseils et nos discussions scientifiques ou non-scientifiques.

Je souhaite remercier chaleureusement Dominique, qui m'a permis de bien terminer ma thèse. Merci pour ton aide sans condition, tes précieux conseils, et ton implication. Je remercie également Abdou, Fred et François pour leur aide, leurs conseils et leur soutien.

Merci aux membres de l'équipe Volf qui m'ont partagé leurs connaissances : Delphine, Fred, Thibault. Et merci Fabien et bien sûr Corentin pour vos discussions et votre soutien. Je remercie particulièrement Laure, merci pour ton aide immense avec les poissons, et merci pour ta gentillesse, nos discussions, ta présence et ton soutien. Merci aux jeunes stagiaires qui m'ont fait confiance pour les encadrer : Théo, Candice et Antonin. Merci également aux membres de passage : Sara et Sho.

Merci Martine, Fabienne et toute l'équipe administrative, pour votre gentillesse, votre patience et vos conseils pour trouver le chemin à travers les méandres de l'administration.

Plus généralement, je remercie l'ensemble de l'IGFL pour le cadre, l'ambiance de travail, les retours constructifs et le soutien dont j'ai pu bénéficier. Merci à l'équipe Ruggiero pour vos conseils et votre aide. Merci à Benjamin et Sandrine pour votre aide pour mon projet de séquençage. Merci également à Robert, pour ta gentillesse et ton aide précieuse au PRECI.

Merci aussi aux doctorants (ou non) pour leur soutien et les happy hours : Camille, Cindy, Nawal, Augustin, Théodore, Amélie, Jonathan, Houssam, Lies, Juliette, Jessika, Yanis... Je remercie également Jean-Nicolas, Nicolas Goudemand, Cyril Charles, Stéphane Vincent et Pradeep Das pour m'avoir fait confiance pour participer à des activités d'enseignements.

Je souhaite également remercier Richard et Abdou pour leurs encouragements, leur écoute et les précieux conseils dispensés durant mes comités de suivi de thèse. Merci à tous les membres du jury d'avoir accepté d'évaluer ce travail et particulièrement à Mireille Bétermier et Clément Gilbert, rapporteurs de cette thèse.

Merci aussi à Sandrine, Ludivine, Denise, Natacha et Cathy pour leur précieux soutien ces derniers mois.

En s'éloignant du labo, je tiens à remercier ceux qui me sont chers et sans qui je ne serais jamais arrivée jusqu'ici. Quelques mots ne sauraient résumer à quel point vous comptez pour moi. Tout d'abord merci à Florence et Marie-Charlotte! Et bien sûr à Corentin, Pauline, Axel et Max, grâce à vous je sais maintenant reconnaître un rat un singe dans un arbre. Merci à Baptiste, Audrey, Sarah et Kévin et aux plus anciennes Joana, Tamara et Axelle. Je remercie toute ma (grande) famille qui m'est si chère, et plus particulièrement Xabi et mes parents, sans qui je ne serais pas arrivée jusqu'ici. Merci tout spécialement à toi Maman, pour ta présence malgré la distance.

Pour finir, merci Jérémy, sans toi je n'écrirais pas ces lignes aujourd'hui.

Tu sais bien que la rédaction ce n'est pas mon fort, alors j'ai plutôt choisi quelques notes pour toi :

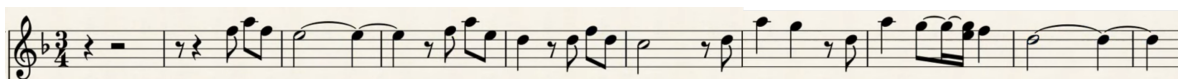


Table des matières

Resumé / Summary	i
Remerciements	v
Table des matières	vii
Liste des figures	ix
Liste des tableaux	xi
1 Introduction	1
1.1 Les vertébrés...	3
1.2 Les éléments transposables : de l'ADN parasite vers des éléments qui boostent l'évolution des génomes	12
1.3 Les vertébrés et les éléments transposables	18
1.4 Le poisson-zèbre comme modèle d'étude fonctionnelle des gènes de vertébrés	47
1.5 Objectifs de la thèse	54
2 Les transposons <i>Harbinger</i> chez les poissons téléostéens	55
2.1 Avant-propos	56
2.2 Article : Diversity of <i>Harbinger</i> -like transposons in teleost fish genomes	57
3 Étude d'une famille de gènes dérivés de transposons <i>Harbinger</i> chez les vertébrés	77
3.1 Introduction	78
3.2 Domestication moléculaire récurrente des transposons <i>Harbinger</i> chez les vertébrés	78
3.3 Mise au point d'un protocole expérimental de KO-direct par CRISPR/Cas9 pour le criblage de phénotype à la première génération	79
3.4 Criblage des fonctions des gènes dérivés de transposons <i>Harbinger</i> chez le poisson-zèbre	86
3.5 Article : The neurodevelopmental gene <i>MSANTD2</i> belongs to a gene family formed by recurrent molecular domestication of <i>Harbinger</i> transposons at the base of vertebrates	89
3.6 Conclusion	139
3.7 Annexe	140
4 Conclusion et perspectives générales	141
4.1 Caractérisation des transposons <i>Harbinger</i> chez les poissons téléostéens	142
4.2 Chez les vertébrés les transposons <i>Harbinger</i> sont à l'origine d'une nouvelle famille de gènes	144

4.3	Les gènes dérivés de transposons <i>Harbinger</i> chez les vertébrés pourraient être impliqués dans le développement du système nerveux	145
4.4	L'utilisation d'une nouvelle technique d'inactivation de gène pour la caractérisation fonctionnelle des gènes	147
4.5	Encore d'autres gènes dérivés de transposons <i>Harbinger</i> ?	149
4.6	Conclusion	157
A	Annexes	I
	Liste complète des références	V

Liste des figures

Figure 1.1 :	Phylogénies simplifiées des bilatériens et des vertébrés	4
Figure 1.2 :	Comparaison du développement embryonnaire des vertébrés	5
Figure 1.3 :	Exemple du xénope pour illustrer l'embryogenèse des vertébrés.	6
Figure 1.4 :	La neurulation	8
Figure 1.5 :	La crête neurale	9
Figure 1.6 :	Les placodes et leurs structures dérivées	10
Figure 1.7 :	Composition en éléments transposables des génomes	13
Figure 1.8 :	Classification et structure des éléments transposables	15
Figure 1.9 :	Les éléments transposables dans les génomes de vertébrés	19
Figure 1.10 :	Diversité des superfamilles d'éléments transposables chez les vertébrés .	20
Figure 1.11 :	Poissons-zèbres adultes femelle et mâle	47
Figure 1.12 :	Nombre de publications des différents organismes modèles	50
Figure 1.13 :	Développement embryonnaire du poisson-zèbre	51
Figure 1.14 :	Développement du poisson-zèbre des stades larvaires précoce et moyen au stade adulte	52
Figure 1.15 :	Cycle de développement du poisson-zèbre	53
Figure 2.1 :	Genome coverage and copy number of <i>Harbinger</i> and <i>ISL2EU</i> transposons in fishes	63
Figure 2.3 :	Multiple alignments of the <i>Harbinger</i> -like transposase proteins on the DDE domain	65
Figure 2.4 :	Multiple alignments of the <i>Harbinger</i> -like Myb-like proteins on the Myb- like domain	66
Figure 2.5 :	Phylogenetic relationships between <i>Harbinger</i> and <i>ISL2EU</i> transposase and Myb-like proteins from different fish species	67
Figure 2.6 :	Phylogenetic congruence between transposase and Myb-like proteins of <i>Harbinger</i> transposons from fish genomes	68
Figure 2.7 :	Expression analysis of <i>Harbinger</i> and <i>ISL2EU</i> transposons in spotted gar, zebrafish, cod and medaka using the PhyloFish database	69
Figure 2.8 :	MAplot representing the relative expression in male and female gonads of all TE families of the medaka genome	70
Figure 2.9 :	Distribution of <i>Harbinger</i> and <i>ISL2EU</i> transposons on medaka chromosomes	74
Figure 2.10 :	Phylogenetic relationships between <i>Harbinger</i> transposase and Myb-like proteins of different fish species by Maximum Likelihood	75
Figure 3.1 :	Comparaison des méthodes de génération de mutants par CRISPR/Cas9	81

Figure 3.2 :	Inactivation du gène <i>SAIYAN</i> par Crispr/Cas9 chez le poisson-zèbre . . .	82
Figure 3.3 :	Reproduction du phénotype de l'inactivation de <i>SAIYAN</i>	83
Figure 3.4 :	Comparaison de l'efficacité des différents sgRNA ciblant le gène <i>SAIYAN</i> par CRISPR/Cas9 chez le poisson-zèbre	84
Figure 3.5 :	Comparaison de la méthode d'injection des complexes sgRNA/Cas9 lors de l'inactivation du gène <i>SAIYAN</i> par CRISPR/Cas9 chez le poisson-zèbre	85
Figure 3.6 :	Comparaison de l'efficacité de différentes concentrations d'injection de sgRNA/Cas9 lors de l'inactivation du gène <i>SAIYAN</i> par CRISPR/Cas9 chez le poisson-zèbre	86
Figure 3.7 :	Phénotypes observés lors de l'inactivation des gènes dérivés de <i>Harbinger</i>	88
Figure 4.1 :	Relations phylogénétiques entre les protéines Myb-like de transposons <i>Harbinger</i> et les protéines MYPOP provenant de plusieurs espèces de vertébrés	150
Figure 4.2 :	Relations phylogénétiques entre les protéines Myb-like de transposons <i>Harbinger</i> et les protéines ZSCAN20 provenant de plusieurs espèces de vertébrés	151
Figure 4.3 :	Relations phylogénétiques entre les protéines Myb-like de transposons <i>Harbinger</i> et les protéines TSNARE1 provenant de plusieurs espèces de vertébrés	152
Figure 4.4 :	Relations phylogénétiques entre les protéines Myb-like de transposons <i>Harbinger</i> et les protéines PRDM11 provenant de plusieurs espèces de vertébrés	153
Figure 4.5 :	Distribution phylogénétique des gènes dérivés de transposons <i>Harbinger</i> chez les drosophiles et d'autres mouches	154
Figure 4.6 :	Relations phylogénétiques entre les protéines Myb-like de transposons <i>Harbinger</i> et les protéines MSANTD1-like de diptères	155
Figure 4.7 :	Relations phylogénétiques entre les protéines Myb-like de transposons <i>Harbinger</i> et les protéines NAIF2 de diptères	156

Liste des tableaux

TABLEAU 3.1 : Séquence des sgRNAs utilisés 140

1

Introduction

Sommaire

1.1 Les vertébrés...	3
1.1.1 ... forment un clade diversifié au considérable succès évolutif...	3
1.1.2 ... aux étapes de développement embryonnaire communes...	5
1.1.3 ... et ayant acquis de nombreuses innovations.	7
1.1.3.1 La crête neurale	7
1.1.3.2 Les placodes	10
1.1.3.3 Le système nerveux	10
1.1.3.4 Le système endocrinien	11
1.1.3.5 Les os et cartilages	11
1.1.3.6 Les membres	11
1.1.3.7 Conclusion	12
1.2 Les éléments transposables : de l'ADN parasite vers des éléments qui boostent l'évolution des génomes	12
1.2.1 Découverte des éléments transposables	12
1.2.2 Définition des éléments transposables	14
1.2.2.1 Classe I : Les rétrotransposons	14
1.2.2.2 Classe II : Les transposons à ADN	16
1.2.2.3 Les transposons <i>Harbinger</i>	16
1.2.3 Influence des éléments transposables sur les génomes	17
1.2.3.1 Les éléments transposables : des éléments aux effets délétères	17
1.2.3.2 Les éléments transposables : des éléments qui boostent l'évolution	17
1.3 Les vertébrés et les éléments transposables	18
1.3.1 Les éléments transposables dans les génomes de vertébrés	18
1.3.2 Les éléments transposables à l'origine d'innovations développementales chez les vertébrés	21
1.4 Le poisson-zèbre comme modèle d'étude fonctionnelle des gènes de vertébrés	47
1.4.1 Le poisson-zèbre...	47
1.4.2 ... est un organisme modèle de choix.	48
1.4.3 Développement du poisson-zèbre	51
1.4.3.1 Développement embryonnaire du poisson-zèbre	51
1.4.3.2 Développement larvaire du poisson-zèbre	52
1.4.3.3 Le poisson-zèbre à l'état adulte	52

1.5 Objectifs de la thèse 54

1.1 Les vertébrés...

1.1.1 ... forment un clade diversifié au considérable succès évolutif..

Des plus petits poissons comme *Paedocypris progenetica*, d'une taille de 10mm et vivant dans la forêt de Sumatra, et des reptiles miniatures comme *Brookesia nana*, de moins de 30mm et trouvés dans la forêt Malgache, jusqu'aux gigantesques dinosaures *Australotitan cooperensis*, de plus de 6m de haut et 30m de long et aux immenses baleines bleues *Balaenoptera musculus*, d'une longueur de 30m, les vertébrés présentent des variabilités morphologiques extrêmes (Glaw et al., 2021; Hocknull et al., 2021; Kottelat et al., 2006; Sears & Perrin, 2009). Ils ont également colonisé la plupart des milieux de vie sur Terre, parfois extrêmes. Le fennec *Vulpes zerda* vit dans le désert du Sahara auquel il est particulièrement adapté grâce à ses oreilles caractéristiques fonctionnant comme un système de régulation thermique (Williams et al., 2004). Les manchots empereurs *Aptenodytes patagonicus* ont, eux, un plumage extrêmement dense leur permettant de vivre dans le froid polaire de l'Antarctique (Duchamp et al., 2002). Même dans les profondeurs abyssales, le poisson Black Dragonfish (*Idiacanthus atlanticus*) a pu s'y développer et est capable de produire sa propre lumière grâce à des photophores lui permettant de chasser activement ses proies. Ainsi, les vertébrés présentent de multiples traits génétiques et phénotypiques adaptés à leur environnement, ayant fait des vertébrés l'un des phyla les plus étendus sur la planète (Zimmer, 2000).

Les premiers vertébrés sont apparus il y a plus de 500 millions d'années, et ont divergé de leurs groupes frères céphalochordés et urochordés (**Figure 1.1A**) (Janvier, 2011). Provenant initialement d'eau douce, les vertébrés ont pu se terrestrialiser il y a environ 400 millions d'années (Wang et al., 2021). Le groupe des vertébrés inclut les agnathes (vertébrés sans mâchoires), les chondrichthyens (poissons cartilagineux), les actinoptérygiens (poissons à nageoires rayonnées) et les sarcoptérygiens (vertébrés à membres charnus, composés des cœlacanthes, dipneustes et tétrapodes) (**Figure 1.1B**). Ainsi, selon les dernières estimations, ce groupe taxonomique se compose de plus de 73.000 espèces (www.iucnredlist.org, (Cazalis et al., 2022)). Il est important de souligner que parmi ces espèces de vertébrés, on estime que plus de 10.000 sont menacées par l'activité humaine, selon l'IUCN Red List (www.iucnredlist.org, Cazalis et al. (2022); Rodrigues et al. (2006)). Les amphibiens sont les plus impactés puisque 41 % de ces espèces sont menacées, ce taux s'élevant à 37 % pour les poissons cartilagineux, 26 % pour les mammifères, 21 % pour les reptiles et 13 % pour les oiseaux (ce taux est difficile à établir pour les poissons osseux par manque de données).

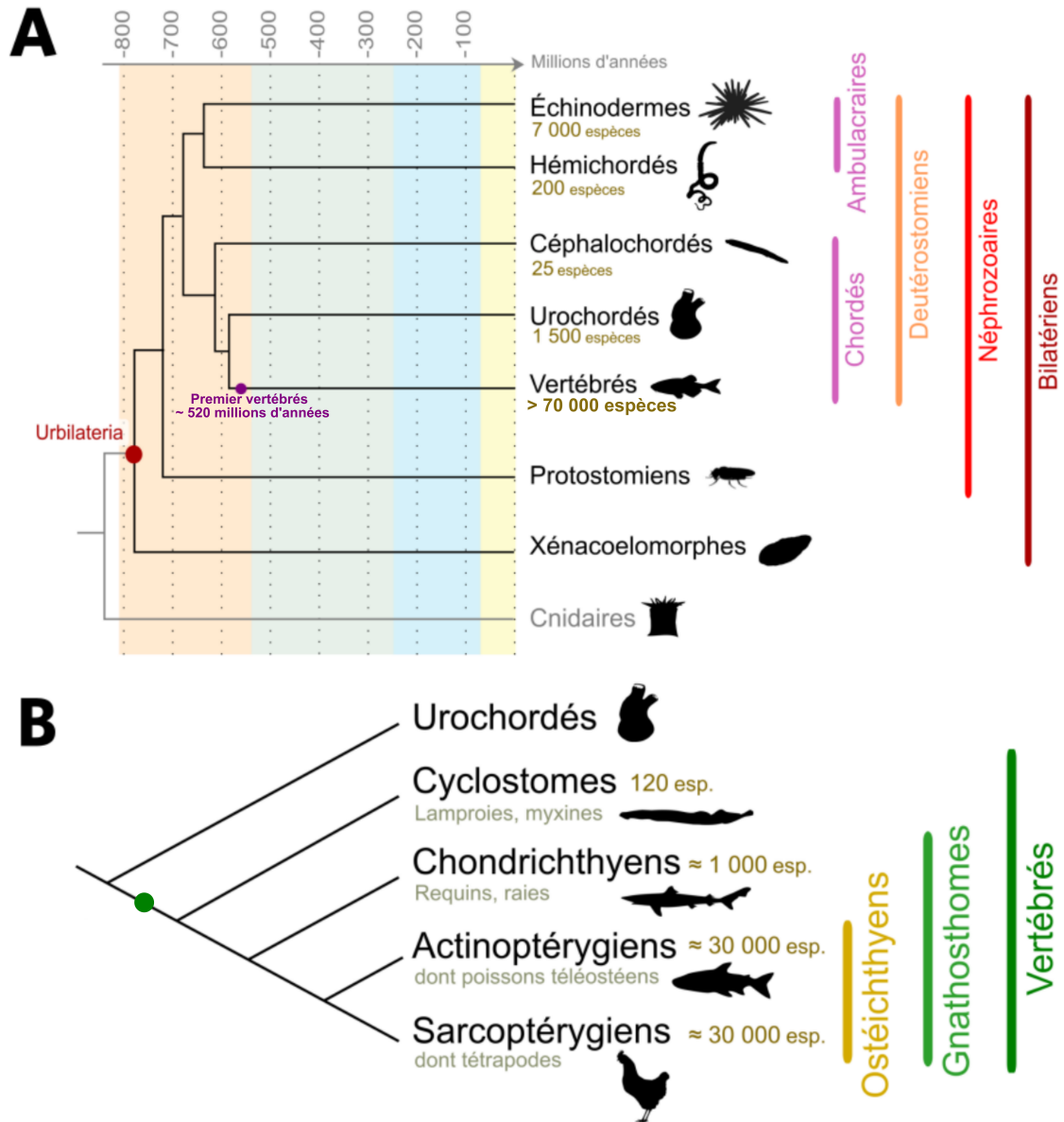


FIGURE 1.1 – **Phylogénies simplifiées des bilatériens (A) et des vertébrés (B).** (A) Phylogénie simplifiée des organismes bilatériens. L'ancêtre hypothétique des bilatériens est appelé Urbilateria. Celui-ci a donné naissance aux xénacoelomorphes, un groupe aujourd'hui réduit et dont la position systématique est discutée, et aux néphrozoaires, eux-mêmes divisés en protostomiens et en deutérostomiens. Les vertébrés appartiennent aux deutérostomiens chordés, et les urochordés sont, chez les chordés, le groupe frère le plus proche des vertébrés. Un des premiers vertébrés connu est *Mylokunmigia* (faune de Chengjiang, environ 520 millions d'années). Les nombres d'espèces de deutérostomiens sont indiqués. Les âges sont en millions d'années. D'après Cannon et al., 2016. (B) Phylogénie simplifiée des vertébrés. Les vertébrés sont divisés en deux grands taxons : les cyclostomes (aussi appelés agnathes, ou vertébrés sans mâchoire) et les gnathostomes (vertébrés avec mâchoire). Au sein des gnathostomes, les chondrichthyens (poissons cartilagineux) représentent le groupe frère des ostéichthyens (poissons osseux). Les tétrapodes constituent la majorité des sarcoptérygiens (vertébrés à membres charnus), le groupe frère des actinoptérygiens (poissons à nageoires rayonnées). Le nombre d'espèces sont indiqués. © Thibault Lorin (Lorin, 2018)

1.1.2 ... aux étapes de développement embryonnaire communes...

Dès le 19^{ème} siècle, les études embryonnaires ont révélé que malgré la variabilité morphologique adulte des différentes classes de vertébrés, de grandes ressemblances sont observées à l'état embryonnaire (**Figure 1.2**) (Hopwood, 2007). En effet, les étapes majeures d'embryogenèse sont communes aux vertébrés et aboutissent à la formation d'un plan d'organisation commun (**Figure 1.3**) (Hopwood, 2007, 2011). Il est à noter que les mécanismes moléculaires se déroulant lors des différentes phases du développement embryonnaire, détaillés ci-dessous, ne sont toutefois pas toujours les mêmes entre les espèces. Les caractères communs ou généraux des vertébrés sont établis en premier au cours du développement embryonnaire, puis les traits divergents entre les espèces sont mis en place à la fin de l'embryogenèse voire après la naissance.

Après la fécondation d'un oocyte par un spermatozoïde, l'embryogenèse commence par une phase de segmentation, lors de laquelle des étapes de divisions cellulaires synchrones vont se succéder pour passer d'une cellule œuf à un stade morula pluricellulaire. Les cellules ainsi formées se nomment blastomères. L'embryon se compose alors de blastomères qui se positionnent sur le vitellus (ou sac vitellin) représentant les réserves énergétiques de l'embryon. Ensuite, l'embryon entre dans un stade blastula, caractérisé par l'apparition d'une cavité (blastocœle) entre les blastomères et le vitellus. À ce stade, se produit la transition blastuléenne (mid-blastula transition) correspondant à l'activation de l'expression du génome de l'embryon, ce qui permettra de remplacer les ARN messagers (ARNm) maternels. Suite à cela, la gastrulation démarre, c'est-à-dire que des mouvements morphogéniques (mouvements de cellules) vont avoir lieu pour mettre en place les différents feuilletts embryonnaires : l'ectoderme, l'endoderme et le mésoderme. Chez

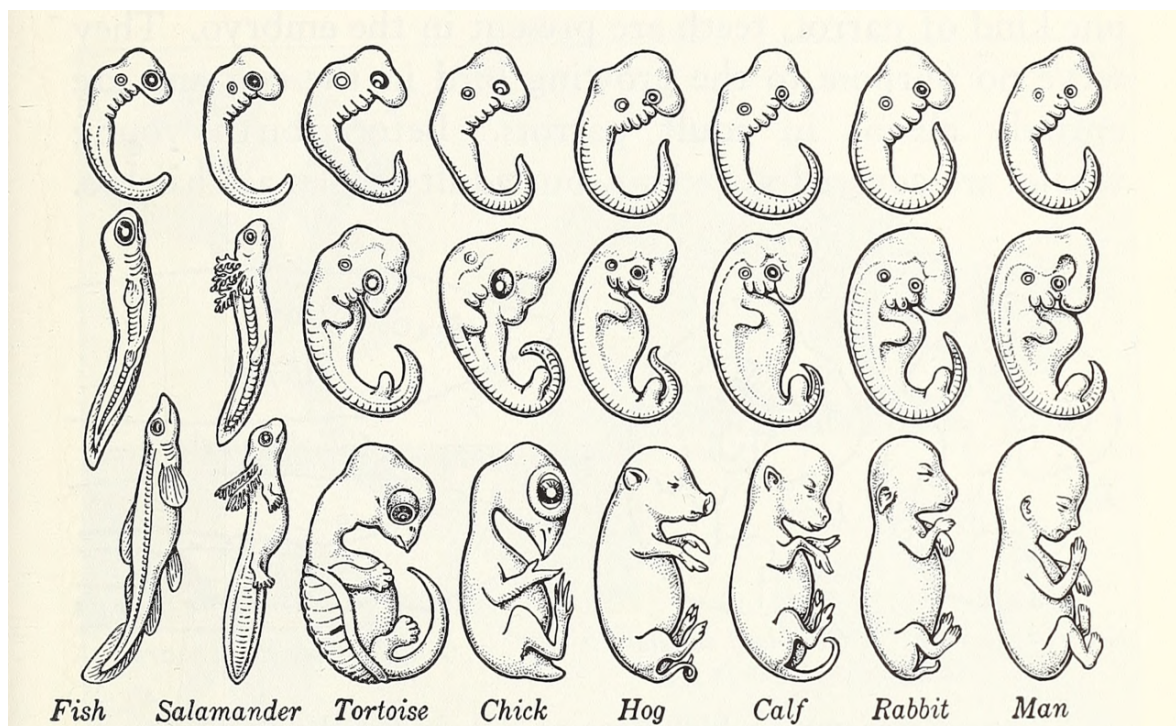


FIGURE 1.2 – **Comparaison du développement embryonnaire des vertébrés.** De gauche à droite des embryons de poisson, salamandre, tortue, poulet, porc, veau, lapin et humain sont représentés. © Life science 1941

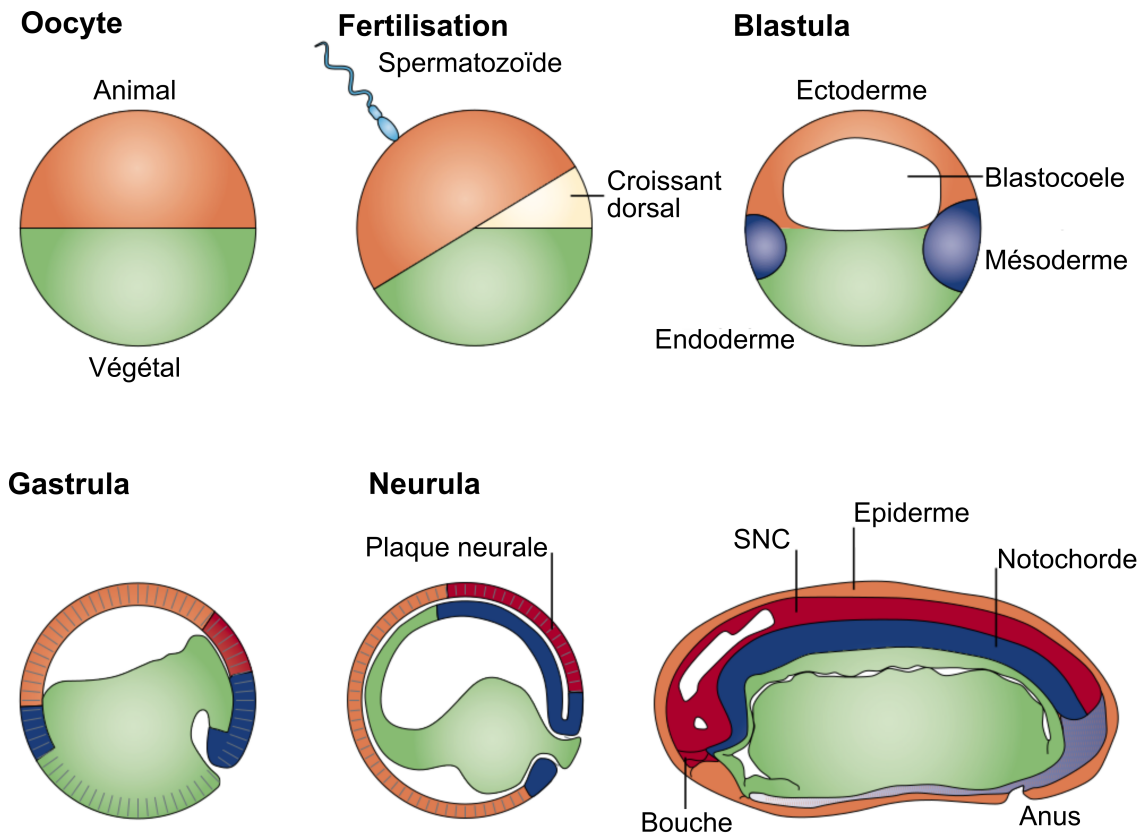


FIGURE 1.3 – Exemple du xénope pour illustrer l'embryogenèse des vertébrés.. L'oocyte est symétrique radialement et divisé en deux domaines appelés « animal » et « végétal ». Une heure après la fertilisation, le croissant dorsal (zone dépigmentée) se forme à l'opposé du point d'entrée du spermatozoïde. L'embryon se divise rapidement en cellules de plus en plus petites et une cavité appelée blastocoele se forme, définissant le stade blastula. Vers la fin du stade blastula (9 heures de développement) les trois feuilletts embryonnaires sont établis. Au stade gastrula (10 heures), l'invagination du mésoderme vers l'intérieur de l'embryon commence. Les mouvements morphogéniques de la gastrulation permettent l'établissement du plan corporel. Au stade neurula (14 heures), la plaque neurale, future système nerveux central (SNC), devient visible au niveau de l'ectoderme dorsal. Entre 24 et 42 heures, une larve avec un tube neural positionné entre l'épiderme et la notochorde se forme. Adapté de De Robertis et al. (2000)

les vertébrés, la gastrulation est caractérisée par quatre mouvements morphogéniques conservés évolutivement : l'invagination (mouvements de cellules vers l'intérieur de l'embryon), l'épibolie (mouvements cellulaires de recouvrement de cellules par d'autres), la convergence et l'extension. La convergence et l'extension vont simultanément affiner et étirer l'embryon. Lors de la gastrulation, les somites (segments de cellules dérivés du mésoderme) vont se former lors de la somitogenèse. Ces somites donneront par la suite les muscles squelettiques, les vertèbres et le derme. À la fin de la gastrulation, le système nerveux commence à se mettre en place lors d'une phase appelée neurulation (**Figure 1.4**). Cette étape initie la dernière phase du développement embryonnaire : l'organogenèse, correspondant à la différenciation des tissus en des structures fonctionnelles. Sous l'ectoderme, une structure du mésoderme appelée la notochorde va induire une partie de l'ectoderme, le neurectoderme, pour former le tube neural. La première étape de la neurulation consiste alors en la formation de la plaque neurale, qui est une simple couche de cellules du neurectoderme. La prolifération rapide de ces cellules va permettre la formation d'une gouttière neurale encadrée

par des pliures neurales. La poursuite des proliférations cellulaires va aboutir à l'agrandissement puis la fermeture de la gouttière neurale en un tube neural, qui formera par la suite le cerveau et la moelle épinière. Les cellules de la crête neurale dérivent des portions latérales de la plaque neurale et leur migration aboutira à la formation de nombreuses structures essentielles des vertébrés (voir **section 1.1.3.1**) (Gans & Northcutt, 1983).

1.1.3 ... et ayant acquis de nombreuses innovations.

Le succès des vertébrés se caractérise notamment par l'acquisition de nombreuses innovations (Zimmer, 2000). Durant la période du Cambrien, il y a environ 500 millions d'années, une lignée de crâniates a donné vie aux vertébrés (Janvier, 2011). Au cours de leur évolution, ces espèces ont acquis principalement un système nerveux plus complexe et un squelette plus développé (Khaner, 2007; Shimeld & Holland, 2000). Ceci leur a permis à la fois de perfectionner leurs capacités à obtenir de la nourriture et à éviter d'en devenir.

Tous les vertébrés ont en commun de multiples innovations : la crête neurale, les placodes, la complexification du système nerveux et du système endocrinien, les os et les cartilages. Les vertébrés à mâchoires (gnathostomes) qui composent la plupart des espèces actuels de vertébrés possèdent d'autres traits représentant des innovations majeures : une mâchoire, des dents, des paires de membres, un canal semi-circulaire horizontal (oreille), une gaine de myéline et un système immunitaire adaptatif (Brazeau et Friedman 2015). Les vertébrés à mâchoires ont également un cerveau plus complexe puisqu'ils possèdent une éminence médiane ganglionnaire permettant la formation du palladium servant à la régulation de mouvements volontaires ainsi qu'un cervelet stratifié (Brazeau & Friedman, 2015).

Les éléments qui seront évoqués ci-dessous sont détaillés dans (Khaner, 2007; Shimeld & Holland, 2000).

1.1.3.1 La crête neurale

L'origine de nombreuses structures considérées comme des innovations majeures des vertébrés est attribuée à un type de cellule embryonnaire en particulier : les cellules de la crête neurale (Gans & Northcutt, 1983; York & McCauley, 2020; Yu et al., 2008). Il s'agit de deux bandes bilatérales de cellules qui se trouvent proches de la gouttière neurale (**Figure 1.4**) (Hall, 2008). Au cours du développement embryonnaire, les cellules de la crête neurale vont migrer vers de multiples sites répartis dans tout l'embryon. Ces cellules migrantes vont donner lieu à de nombreuses structures spécifiques des vertébrés comme les os et cartilages du crâne, le squelette branchial, les ganglions sensoriels, le système nerveux périphérique etc. (**Figure 1.5**). Même si les céphalochordés et urochordés possèdent des tissus potentiellement homologues à la crête neurale, les structures dérivées de ces tissus restent très restreintes (York & McCauley, 2020).

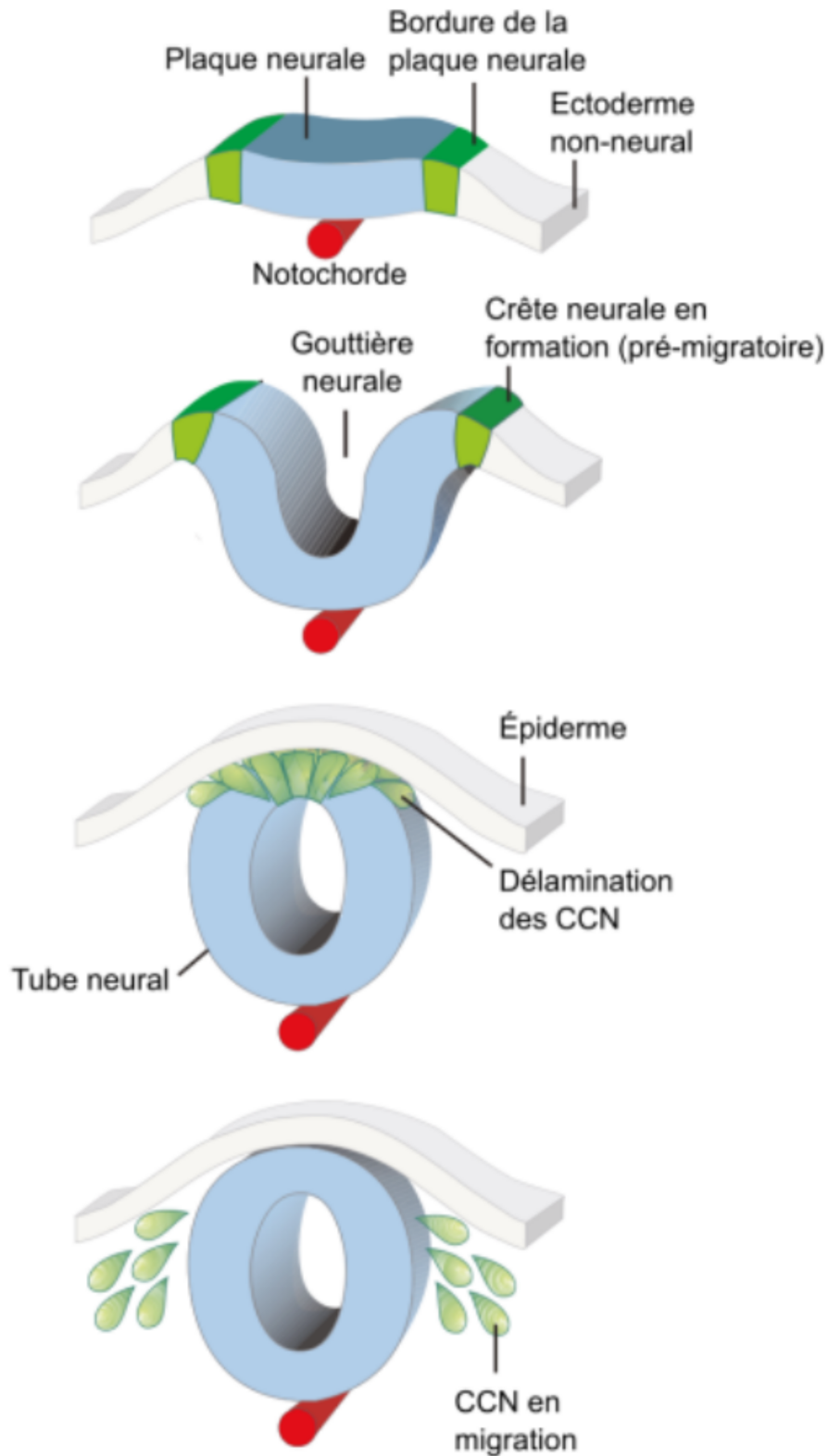


FIGURE 1.4 – **La neurulation.** La plaque neurale est un épaissement du neurectoderme (en bleu). Au cours de la neurulation, les bordures de la plaque neurale se replient et forment des « crêtes » entourant la gouttière neurale. Le tube neural se forme suite à l’invagination du neurectoderme. Les cellules de la crête neurale (CCN, en vert) sont originaires des bordures de la plaque neurale. Après l’invagination du neurectoderme, les cellules de la partie dorsale du tube neural se détachent (délamination) et commencent à migrer : ce sont les CCN. © Thibault Lorin (Lorin, 2018)

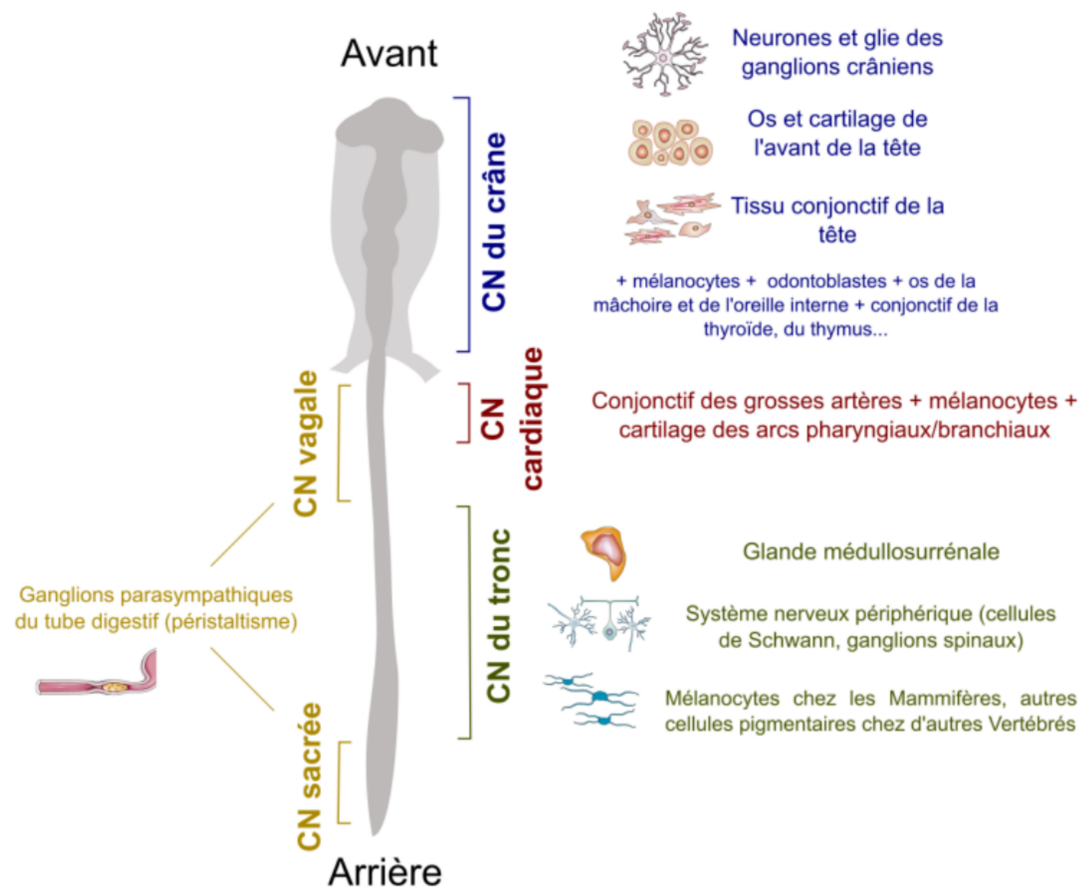


FIGURE 1.5 – **La crête neurale.** Les différents domaines de la crête neurale (CN) ainsi que leurs principaux dérivés. © Thibault Lorin (Lorin, 2018)

1.1.3.2 Les placodes

Les placodes sont des structures embryonnaires qui proviennent de l'épaississement de l'ectoderme. Elles sont à l'origine de la formation de nombreuses structures spécialisées des vertébrés (**Figure 1.6**) (Holland & Holland, 2001). Il en existe deux types : les placodes sensorielles (contribuant aux organes sensoriels) et les placodes neurogéniques (contribuant aux ganglions crâniens). Les céphalochordés et urochordés pourraient également posséder des structures sensorielles homologues, mais les organes sensoriels de vertébrés sont bien plus avancés (Shimeld & Holland, 2000). Ces organes permettent la détection et la transmission d'information concernant des milieux extérieurs et intérieurs complexes à un cerveau également complexifié. Par exemple, chez les poissons, les bulbes olfactifs, la ligne latérale, la placode otique et l'appareil vestibulaire sont tous des organes sensoriels innovants qui leur permettent de sentir chimiquement leur environnement, de mieux s'y déplacer et enfin d'entendre leur milieu environnant.

1.1.3.3 Le système nerveux

Les vertébrés présentent un système nerveux plus développé avec notamment un cerveau complexifié, organisé en régions spécialisées (Sugahara et al., 2017). Au cours du développement, le cerveau est composé de trois régions : prosencéphale (cerveau antérieur – forebrain en anglais), mésencéphale (cerveau moyen – midbrain en anglais) et rhombencéphale (cerveau postérieur – hindbrain en anglais). Ces régions donnent ensuite lieu au télencéphale et diencéphale (cerveau antérieur), tectum et tegmentum (cerveau moyen), et métencéphale et myélocéphale (cerveau postérieur). Chaque région se différencie ensuite en sous-domaines avec des types cellulaires et neuronaux particuliers, permettant la mise en place de réseaux de neurones complexes. Tout ceci est à l'origine de fonctions sophistiquées caractéristiques du cerveau des vertébrés (Holland, 2009; Northcutt, 1984).

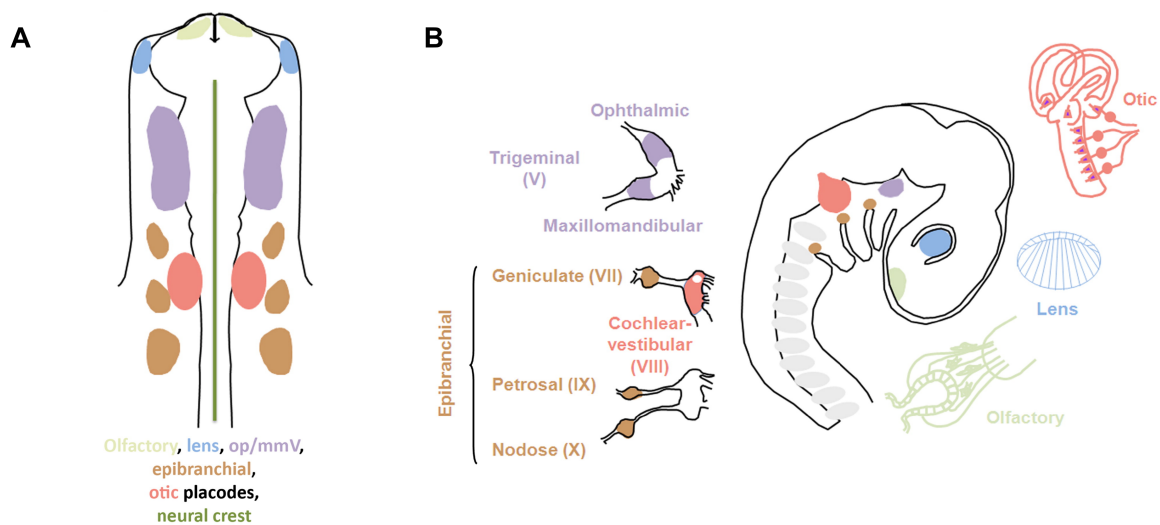


FIGURE 1.6 – **Les placodes et leurs structures dérivées.** (A) Représentation d'un embryon de poulet au stade 10-somites. A ce stade, les différentes placodes sont différenciées morphologiquement (représentées par différentes couleurs) et occupent des positions distinctes le long du tube neural. (B) Représentation d'un embryon au stade 3 jours. Les placodes et leurs structures dérivées sont représentées en différentes couleurs. D'après Grocott et al. (2012).

Il existe d'autres structures dérivées du système nerveux comprenant les nerfs crâniens, la moelle épinière, les ganglions et les systèmes nerveux viscéral et périphérique (Fritzscht & Northcutt, 1993; Northcutt, 1996).

Au cours de l'évolution des vertébrés, le cerveau a continué à se complexifier. En particulier les régions associées aux organes sensoriels se sont développées et étendues chez les amphibiens et les amniotes (Northcutt, 1995). Ceci a eu pour conséquence d'augmenter les capacités de détection et d'interaction des vertébrés avec leur environnement et a donc permis leur adaptation à de multiples milieux.

1.1.3.4 Le système endocrinien

Les vertébrés présentent également un système endocrinien complexifié comprenant les glandes endocrines, les hormones et leurs tissus cibles (Campbell et al., 2004; Norris & Carr, 2020). Comparé aux céphalochordés et urochordés qui ont un système endocrinien très basique, celui des vertébrés présente de multiples glandes comme par exemple la glande surrénale, le pancréas, l'hypophyse et la thyroïde. Ces glandes, permettant la régulation de systèmes hormonaux précis, sont à l'origine de processus métaboliques et physiologiques essentiels, comme la régulation de sucres par l'insuline (produite par le pancréas) ou encore les hormones placentaires pour le développement du placenta.

1.1.3.5 Les os et cartilages

Les os et les cartilages représentent les tissus durs. Ces deux types de tissus, à l'origine embryonnaire commune, sont retrouvés dans la plupart des espèces de vertébrés. Ils sont à l'origine de la structure représentative de ce clade : les vertèbres (Hirasawa & Kuratani, 2015). Ils forment également le crâne et la mâchoire. Le crâne est une innovation essentielle des vertébrés puisqu'elle a permis la mise en place d'un support et d'une protection du cerveau (Kaucka & Adameyko, 2019). Le développement d'une mâchoire articulée est également un événement majeur ayant permis l'émergence des vertébrés à mâchoires (gnathostomes). Cette structure a été décisive dans la modification de leur régime alimentaire, mettant en place un système de prédation plus large et actif. La colonne vertébrale a, quant à elle, joué un rôle essentiel lors de la transition vers un mode de vie terrestre. Au cours de cette transition, elle a subi de considérables changements, notamment en devenant plus ossifiée, plus dure et régionalisée, pour permettre de supporter le poids des corps, parfois extrême, sur terre (Hirasawa & Kuratani, 2015).

1.1.3.6 Les membres

Enfin, les vertébrés présentent tous une paire de membres (nageoires ou pattes), ce qui a permis l'apparition de nouveaux moyens de locomotion (Freitas et al., 2006). L'évolution de ces membres a ensuite donné lieu à deux paires de membres, chez les tétrapodes, particulièrement adaptées lors de la transition vers une vie terrestre et de prédation intense.

1.1.3.7 Conclusion

Ainsi, nous avons pu nous rendre compte que l'évolution des vertébrés est associée à l'apparition d'innovations majeures, ayant fait de ces espèces un clade au succès inégalé (Zimmer, 2000). Cependant, la question de l'origine génétique de ces innovations n'est pas toujours bien connue et caractérisée encore aujourd'hui. L'apparition de nouveaux gènes ainsi que la complexification des réseaux de régulation des gènes sont des événements essentiels lors de la mise en place de nouvelles fonctions (Kaessmann, 2010). À la base des vertébrés, deux événements de duplication de génomes entiers ont permis des expansions majeures du répertoire de gènes (Dehal & Boore, 2005). En particulier, ceux-ci ont pu permettre la mise en place de nouveaux réseaux de régulation de gènes (Holland et al., 1994). Un exemple significatif est celui des gènes HOX (gènes codant pour des facteurs de transcription, permettant la mise en place des différents organes le long de l'axe antéro-postérieur) qui sont passés au nombre de 38 chez les vertébrés à mâchoires (Holland et al., 1994; Pendleton et al., 1993). Ainsi, Ohno a proposé que ces événements de duplications de génomes correspondent à des événements fondateurs dans l'apparition d'innovations (Ohno, 1999). Cependant, la formation de nouveaux gènes par d'autres mécanismes ne doit pas être sous-estimée, en témoignent les gènes *RAG*, issus de la domestication moléculaire d'éléments transposables, à l'origine du système immunitaire adaptatif (voir **section 1.3**).

1.2 Les éléments transposables : de l'ADN parasite vers des éléments qui boostent l'évolution des génomes

Le génome d'un organisme est l'ensemble de l'information génétique des cellules d'un individu, qui se trouve sous forme d'ADN (Acide Désoxyribonucléique) (ou d'Acide RiboNucléique – ARN – pour certains virus). La connaissance des génomes des organismes, par leur séquençage, est donc une mine d'informations pour la compréhension du fonctionnement et de l'évolution, c'est-à-dire l'histoire, des espèces. Ainsi, le séquençage du génome humain, qui a débuté au début des années 1990 et qui a duré plus de 10 ans, fut une avancée majeure dans le monde de la recherche scientifique (Lander et al., 2001). Au-delà de cette avancée, les résultats des projets de séquençage du génome humain furent particulièrement étonnants, car ils ont permis de mettre en évidence une composition inattendue de notre génome. En effet, ces projets permirent d'estimer à 30.000-35.000 le nombre de gènes codants pour des protéines, et les séquences codantes à environ 2 % de notre génome (Lander et al., 2001). La partie non-codante du génome apparût donc comme une part beaucoup plus importante que ce que les scientifiques pensaient jusqu'alors. Parmi ces séquences non-codantes, la majorité fut considérée pendant longtemps comme de l'ADN parasite voire de « l'ADN poubelle », car a priori inutile.

1.2.1 Découverte des éléments transposables

Longtemps considérés comme de « l'ADN poubelle », les éléments transposables furent découverts par Barbara McClintock dans les années 1950 (McClintock, 1956). Elle identifia ces éléments grâce à l'observation de variations de couleur des grains de maïs. Elle relia ces variations à des «

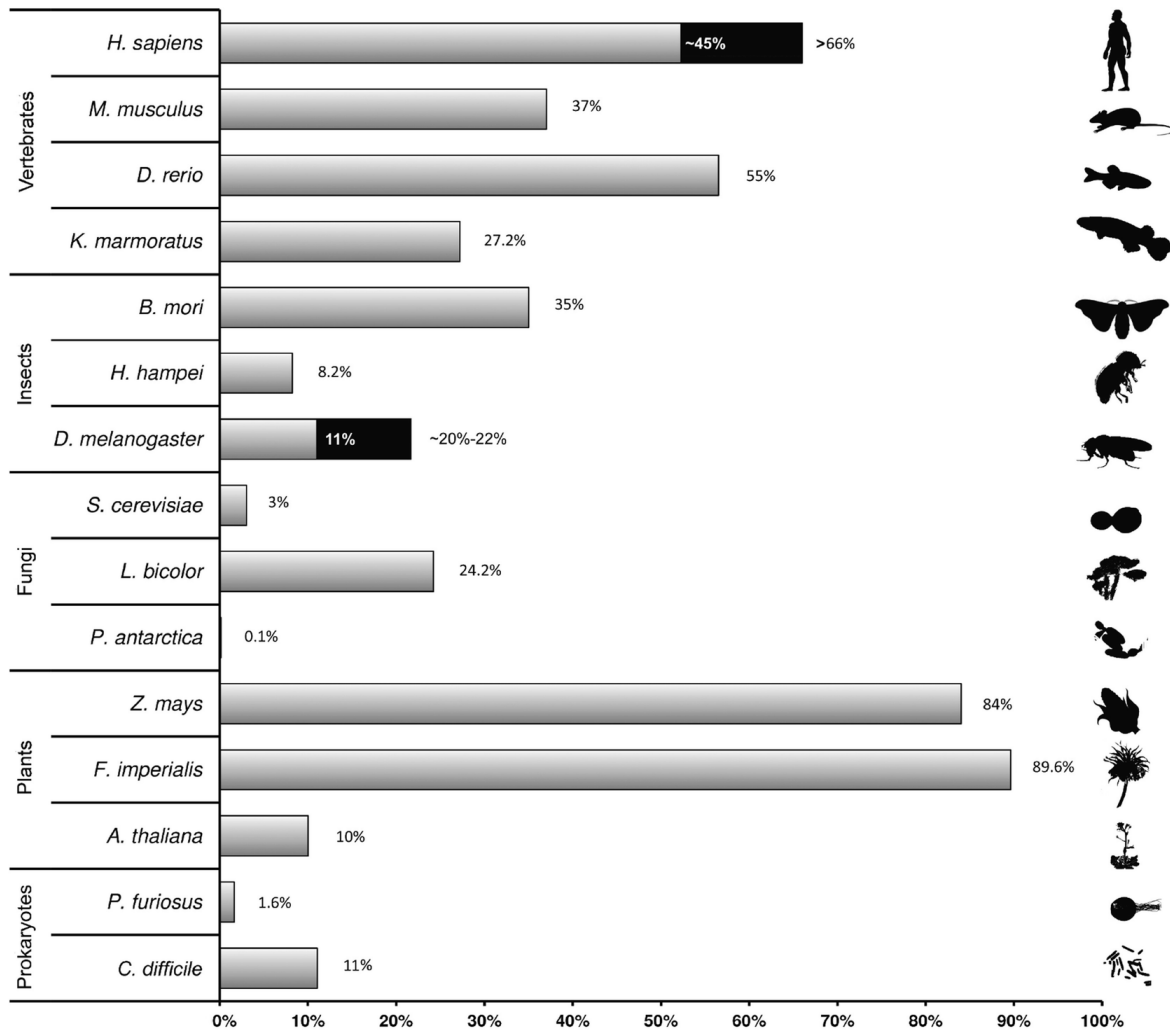


FIGURE 1.7 – **Composition en éléments transposables des génomes de différents organismes, exprimée en pourcentage du génome.** Les zones en noir représentent la variation des estimations. D'après Guio & González (2019).

controlling elements » qui pourraient influencer la régulation des gènes. Cela permis donc de remettre en cause l'idée que ces séquences puissent être totalement inutiles, comme suggéré par Doolittle & Sapienza (1980) et Orgel & Crick (1980). Ces travaux auront valu à Barbara McClintock le prix Nobel de physiologie ou médecine en 1983.

À l'heure actuelle, les éléments transposables ont été retrouvés dans tous les génomes où ils ont été recherchés, comprenant les procaryotes, les protistes, les champignons, les plantes et les animaux (Biémont & Vieira, 2006; Bourque et al., 2018; Kazazian, 2004). De plus, ils peuvent composer une partie majeure de ces génomes, ils représentent plus de 50 % du génome humain par exemple (Figure 1.7) (Guio & González, 2019; Lander et al., 2001).

1.2.2 Définition des éléments transposables

Les éléments transposables (ETs) sont des séquences d'ADN répétées, qui peuvent être insérés ou copiés à de nouvelles positions génomiques. Ils peuvent donc causer de l'instabilité génomique du fait de leur insertion ou de recombinaisons. Certains de ces éléments peuvent contenir des ORF (open reading frame ou cadre ouvert de lecture) leur permettant de coder pour leurs propres protéines, nécessaires à leur transposition autonome. D'autres éléments non-autonomes sont totalement non-codants, ils ont perdu la capacité de transposer en autonomie, et dépendent donc d'éléments autonomes. Les éléments transposables sont séparés en deux classes principales, dépendant de leur mécanisme de transposition (**Figure 1.8**) (Kapitonov & Jurka, 2008; Wicker et al., 2007). Au sein de ces deux classes, les éléments sont divisés successivement en sous-classes, superfamilles et familles. Les sous-classes se définissent en fonction des différences de mécanismes d'intégration chromosomique. Les éléments de différentes superfamilles sont généralement retrouvés dans de nombreuses espèces, ont une stratégie de réplication et une lointaine origine phylogénétique communes (Kapitonov & Jurka, 2008; Wicker et al., 2007). Les éléments d'une même famille sont reliés phylogénétiquement et correspondent aux copies qui découlent de la multiplication d'une même séquence ancestrale. Ainsi, les familles d'ETs sont définies par la conservation de la séquence ADN (Kapitonov & Jurka, 2008; Wicker et al., 2007).

1.2.2.1 Classe I : Les rétrotransposons

La classe I correspond aux rétrotransposons, ils utilisent un mécanisme de « copier-coller » pour leur transposition (Coffin, 1992; Kumar & Bennetzen, 1999; Voytas & Boeke, 1992). La séquence de ces éléments est transcrite en un ARN qui sert de matrice à la synthèse d'un ADN complémentaire (ADNc) par réverse transcription, qui sera lui intégré dans le génome (Beauregard et al., 2008; Goodier, 2016). Il s'agit donc d'une transposition répllicative, qui permet l'expansion des familles de rétroéléments dans le génome hôte.

Parmi les rétrotransposons une première sous-classe correspond aux éléments LTRs (Long Terminal Repeats), qui possèdent des séquences LTR répétées à leurs bords en orientation directe (Kapitonov & Jurka, 2008; Wicker et al., 2007; Zhang et al., 2014). Ces séquences, entre 250 et 600 paires de bases, sont nécessaires à l'expression et l'intégration des éléments dans les génomes hôtes, car ils contiennent des séquences de promoteur et amplificateur (enhancer en anglais – séquence régulatrice qui active la transcription). Les éléments LTRs autonomes contiennent deux ORFs : Gag et Pol (un troisième ORF Env est également présent dans certaines superfamilles tels que pour les rétrovirus endogènes (ERV)). Ces ORFs codent pour des polyprotéines GAG, POL (et ENV) qui sont des précurseurs clivés par la suite. Les protéines produites sont une protéine de matrice (MA), une capsid (CA), une nucléocapsid (NC), une protéase (AP), une reverse transcriptase (RT), une ribonucléase H (RNaseH), une intégrase (INT) et éventuellement une protéine d'enveloppe (ENV) dans le cas des ERVs (Beauregard et al., 2008; Curcio & Derbyshire, 2003; Goodier, 2016). Ces protéines permettent la formation d'une capsid et le repliement de l'ARNm du LTRs à l'intérieur (MA, CA, NC), la rétrotranscription de cet ARNm en ADN complémentaire (RT), l'hydrolyse de l'hybride ADN/ARN formé lors de la rétrotranscription (RNaseH) et l'intégration de l'ADN complémentaire dans le génome (INT). La protéase (AP) quant à elle permet de cliver la polyprotéine, et la

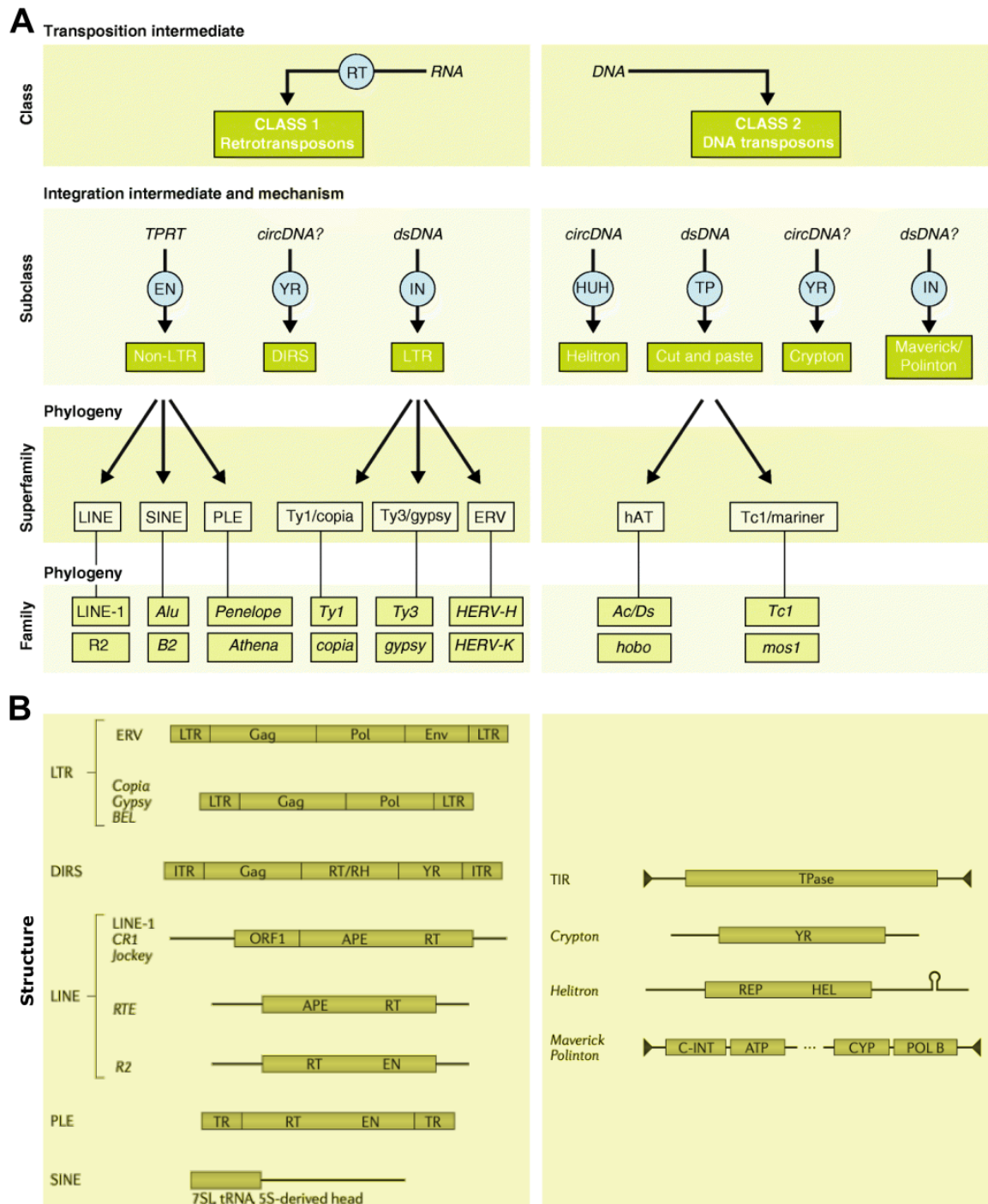


FIGURE 1.8 – **Classification et structure des éléments transposables.** (A) Classification des éléments transposables des eucaryotes. D’après Bourque et al. (2018). Représentations schématiques des caractéristiques et relations principales entre les classes, sous-classes, superfamilles et familles d’ETs. Les cercles bleus représentent les enzymes codées par un ET. circDNA ADN circulaire intermédiaire, DIRS Dictyostelium repetitive sequence, dsDNA ADN linéaire double-brin, EN endonucléase, IN intégrase, PLEs élément Penelope-like, HUH protéine Rep/Helicase avec une activité endonucléasique HUH, RT reverse transcriptase, TP transposase, TPRT target primed reverse transcription, YR tyrosine recombinase. (B) Structures générales des différentes sous-classes d’ETs. Adapté de Deniz et al. (2019).

protéine ENV permet la formation d’une enveloppe enveloppant la capsid. Il existe des éléments non-autonomes, tels que les éléments *Morgane*, *LARD* et *TRIM* qui sont mobilisés en trans par des LTRs autonomes (Kalendar et al., 2004; Sabot et al., 2006; Witte et al., 2001).

Les Dictyostelium Intermediate Repeat Sequences (DIRS), forment une autre sous-classe de rétrotransposons. Ces éléments ressemblent aux LTRs mais codent pour une tyrosine recombinase à la place de l'intégrase.

Enfin, dans une dernière sous-classe on trouve les rétrotransposons non-LTRs aussi appelés Long Interspersed Nuclear Elements (LINEs) (Kapitonov & Jurka, 2008; Wicker et al., 2007). Les LINEs autonomes codent pour une RT et une endonuclease. Il existe également des éléments non-autonomes appelés Short Interspersed Nuclear Elements (SINEs), qui ne codent pour aucune protéine et sont mobilisés en trans par les LINEs (Dewannieux et al., 2003; Richardson et al., 2015).

1.2.2.2 Classe II : Les transposons à ADN

La classe II correspond aux transposons à ADN. Leur transposition ne nécessite pas d'intermédiaire ARN servant lors d'une reverse transcription (Feschotte & Pritham, 2007). Ces éléments utilisent en général un mécanisme de « couper-coller ». La séquence ADN du transposon est excisée du locus initial et est ensuite intégrée à une autre position du génome. Il s'agit d'un type de transposition dite conservatrice, puisque le nombre de copies de l'élément reste le même. Ces éléments codent pour une transposase et possèdent des répétitions terminales inversées (TIR : Terminal Inverted Repeats) qui sont liées par la transposase pour l'excision et l'intégration du transposon (Curcio & Derbyshire, 2003; Feschotte & Pritham, 2007). Il existe aussi les Hélitrons, des transposons ADN sans TIR, qui utilisent un mécanisme « peler-coller », qui est une transposition répllicative utilisant un intermédiaire ADN circulaire (un des deux brins d'ADN est enroulé sur lui-même) (Grabundzija et al., 2016; Kapitonov & Jurka, 2007; Thomas & Pritham, 2015). Ces ETs codent pour une hélicase, permettant de dérouler la double hélice d'ADN. On retrouve également les Polintons/Mavericks qui sont des éléments avec de long TIRs et à « synthèse autonome » puisqu'ils codent pour une ADN polymérase (Kapitonov & Jurka, 2006; Krupovic & Koonin, 2015). Enfin, les éléments Cryptons, encore peu décrits actuellement, sont des éléments sans TIRs codant pour une tyrosine recombinase (Kojima & Jurka, 2011). Concernant les éléments non-autonomes de cette classe, les Miniature Inverted Repeat Transposable Elements (MITEs) possèdent des TIRs mais ne codent pour aucune protéine. Ils sont mobilisés en trans par des transposons à ADN autonomes.

1.2.2.3 Les transposons *Harbinger*

Lors de cette thèse je me suis particulièrement intéressée aux transposons à ADN nommés *Harbinger*. Les éléments *Harbinger* ont été identifiés dans de nombreuses espèces incluant des protistes, des plantes, des insectes, des vertébrés, mais sont absents des génomes de mammifères (Casola et al., 2007; Grzebelus et al., 2007; Han et al., 2015; Jiang et al., 2003; Jurka & Kapitonov, 2001; Kapitonov & Jurka, 1999, 2004; Kikuchi et al., 2003; Markova & Mason-Gamer, 2015; Pereira et al., 2013; Yuan & Wessler, 2011; Zhang et al., 2001, 2004). Les transposons *Harbinger* possèdent des TIRs de 25 à 50 paires de bases. Leur transposition produit généralement des TSDs de 3 paires de bases (TSD : Target Site Duplication, séquences identiques à chaque extrémité de l'ET, formées par l'intégration de l'élément). Ces éléments possèdent deux ORFs. Le premier code pour une transposase, contenant un motif DDE d'endonuclease (DDE fait référence aux trois acides aminés catalytiques nécessaires à l'activité enzymatique – D acide aspartique et E acide glutamique). Le

deuxième ORF code pour une protéine possédant un domaine de liaison à l'ADN, Myb/SANT-like, composé d'un motif tri-hélice. Cette protéine est capable de se lier à la transposase, ce qui induit leur transport nucléaire (Sinzelle et al., 2008). Elle peut également se lier à l'ADN au niveau des TIRs de transposons *Harbinger*, permettant l'excision de la séquence de *Harbinger* par la transposase.

1.2.3 Influence des éléments transposables sur les génomes

1.2.3.1 Les éléments transposables : des éléments aux effets délétères

Bien que la plupart des insertions des éléments transposables dans les génomes soient neutres, c'est-à-dire que l'insertion ne produit ni avantage ni désavantage, les ETs sont également associés à des effets délétères (Bourque et al., 2018). Historiquement, ils étaient d'abord considérés comme de l'ADN parasitique (Doolittle & Sapienza, 1980; Orgel & Crick, 1980). En effet, les ETs ont un fort potentiel mutagène, avec éventuellement des effets délétères. Des mutations délétères peuvent être induites directement, comme lorsque la séquence d'un ET est insérée dans une région codante ou régulatrice, conduisant à une protéine non-fonctionnelle ou dérégulée. Dû au caractère répétitif des ETs dans les génomes et aux recombinaisons que cela induit, des mutations délétères peuvent également apparaître. L'insertion d'un ET peut avoir un impact sur les régions codantes, puisqu'elle peut induire la modification de structure et d'environnement épigénétique de la chromatine alentour. Ces différents mécanismes ont été reliés à diverses maladies humaines (Chenais, 2015; Payer & Burns, 2019). De nombreux cancers sont associés aux mutations induites par les ETs, notamment lorsque l'insertion d'un ET conduit à l'altération de gènes suppresseurs de tumeur ou de proto-oncogènes (Morse et al., 1988). Un cas documenté est celui de l'insertion d'un élément *LINE1* dans le gène *APC* suppresseur de tumeur, ce qui induit un cancer colorectal (Scott et al., 2016). Les éléments *Alu*, des éléments SINEs qui composent 10 % du génome humain (Lander et al., 2001), sont à l'origine de recombinaisons et duplications qui ont pu être associées à diverses leucémies (Jeffs et al., 1998; O'Neil et al., 2007). Un autre exemple est le cas d'un *Alu* inséré dans le gène de la neurofibromine, qui induit la neurofibromatose de type I (maladie caractérisée par la survenue de tumeurs le long des nerfs) (Wallace et al., 1991). Un élément *LINE1* inséré dans le gène de la dystrophine est également responsable d'une myopathie de Duchenne, une maladie de dégénérescence progressive des muscles (Narita et al., 1993).

1.2.3.2 Les éléments transposables : des éléments qui boostent l'évolution

Au-delà de leurs effets délétères ou neutres des ETs, il est maintenant largement admis que les ETs peuvent avoir des effets bénéfiques ayant eu une importance non négligeable dans l'évolution des espèces (Biémont & Vieira, 2006; Bourque, 2009; Fedoroff, 2012). En effet, de par leur quantité dans les génomes, mais également leur diversité, ils sont une source de séquences directement disponible et ré-utilisable par l'hôte. Ainsi, lorsque ces séquences ont un effet positif sur la valeur sélective de l'hôte, elles peuvent être co-optées/exaptées, c'est-à-dire recrutées pour remplir une fonction utile à l'hôte. L'exaptation des ETs pouvant être à l'origine d'adaptation et d'innovation pour les organismes est de plus en plus documentée, du fait d'un intérêt grandissant, aidé par le développement des nouvelles technologies de séquençage et d'analyse d'expression de gènes.

Nous reviendrons plus en détails sur les mécanismes et des exemples précis d'exaptation d'éléments transposables dans la **section 1.3.2**

1.3 Les vertébrés et les éléments transposables

Dans le cadre de cette thèse je me suis particulièrement intéressée aux vertébrés, c'est pourquoi il convient de s'arrêter plus précisément sur le lien entre éléments transposables et ce clade particulier.

1.3.1 Les éléments transposables dans les génomes de vertébrés

Les éléments transposables sont des composants majeurs des génomes de vertébrés (Böhne et al., 2008; Deininger et al., 2003; Feschotte & Pritham, 2007; Kazazian, 2004; Kordis, 2009). Le contenu en éléments transposables des génomes est d'ailleurs corrélé à la taille des génomes (Chalopin et al., 2015; Naville et al., 2019). Cependant, ils y contribuent à des taux variables en fonction des lignées (**Figure 1.9**) (Carducci et al., 2020; Chalopin et al., 2015). En effet, les génomes de mammifères sont par exemple plus riches en ETs que ceux des oiseaux. On observe également des variations au sein des lignées, comme chez les poissons où ils peuvent constituer jusqu'à 55 % du génome du poisson-zèbre, mais seulement 6 % chez le tétraodon.

En plus d'une variation quantitative, on observe également une variation qualitative des éléments transposables dans les génomes de vertébrés. Par exemple, les ETs de mammifères sont principalement des rétrotransposons avec très peu de transposons ADN, alors qu'on observe plutôt l'inverse chez les poissons téléostéens. De plus, en s'intéressant aux nombres de superfamilles différentes d'ETs dans les génomes, on observe également une grande variation (**Figure 1.10**). La plus grande diversité de superfamilles d'ETs est observée chez les poissons téléostéens.

Toutes ces variations sont liées à des succès d'invasion différents, ainsi qu'à la diversité d'activité et de compétition entre différentes familles d'ETs dans ces génomes. Ceci a conduit à des génomes de vertébrés présentant un éventail important de composition en ETs, offrant de multiples environnements génomiques permettant de multiplier les possibilités de co-optation des ETs.

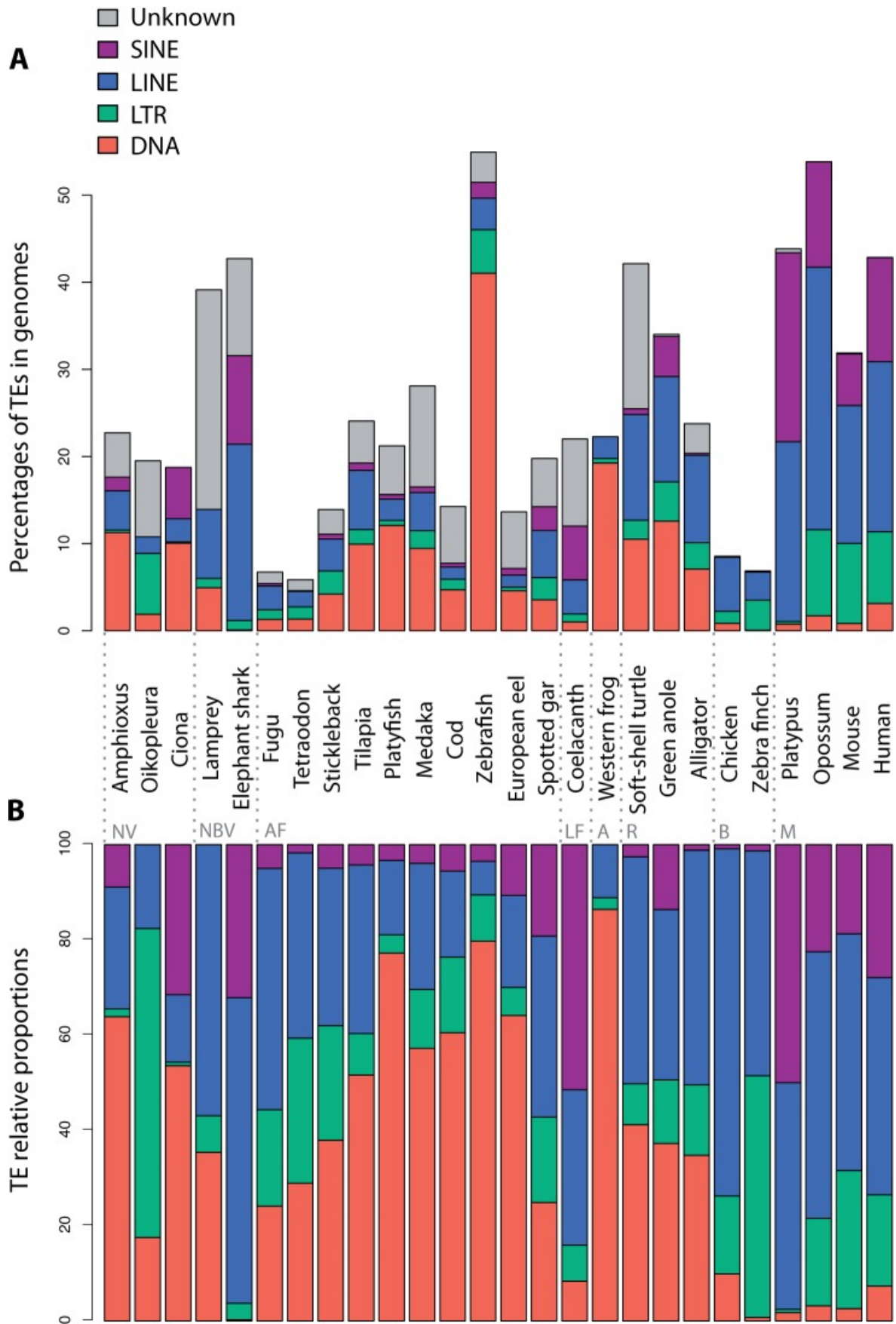


FIGURE 1.9 – **Quantité totale (A) et proportion relative (B) d'éléments transposables dans les génomes de vertébrés.** NV, invertébrés; NBV, vertébrés non-osseux; AF, poissons actinoptérygiens; LF, sarcoptérygiens; A, amphibiens; R, reptiles; B, oiseaux; M, mammifères. D'après Chalopin et al. (2015).

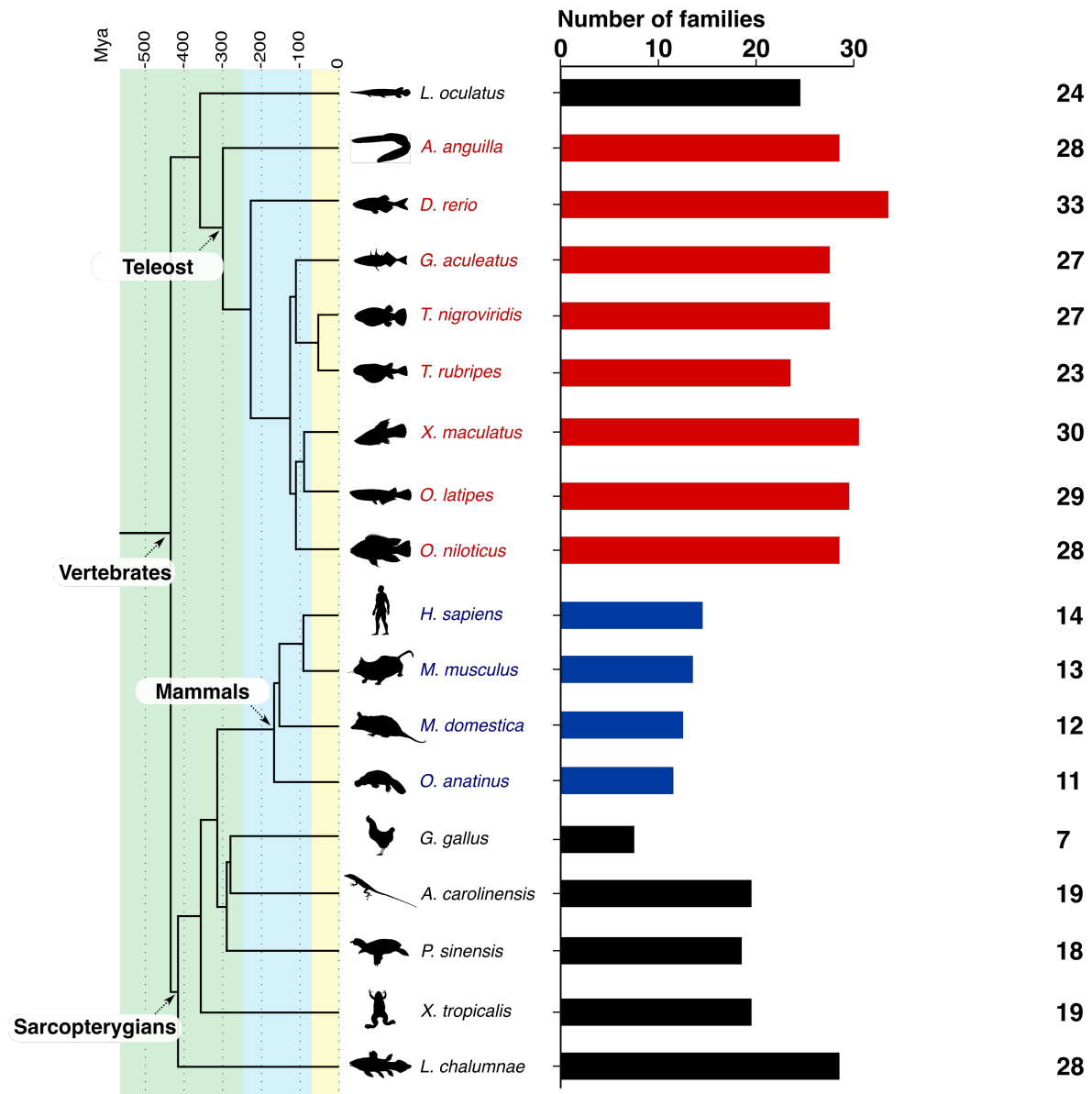


FIGURE 1.10 – **Phylogénie simplifiée des vertébrés présentant la diversité des superfamilles d'éléments transposables.** Poissons téléostéens (en rouge), mammifères (en bleu) © Milton Tan - *O. niloticus*, *G. aculeatus*, *X. maculatus* CC BY-NC-SA – © Sarah Werning – *M. domestica*, *O. anatinus*, *A. carolinensis* CC BY - © Soledad Miranda-Rottmann – *P. sinensis* CC BY - © Maija Karala – *L. chalumnae* - CC BY-NC-SA - <http://phylopic.org/>. Adapté de Chalopin et al. 2015. Adapté de Chalopin et al. (2015).

1.3.2 Les éléments transposables à l'origine d'innovations développementales chez les vertébrés


Comme évoqué précédemment, les éléments transposables au-delà de leurs effets négatifs ou neutres, peuvent avoir un impact positif sur la valeur sélective de leurs hôtes. Ainsi, depuis leur découverte il y a plus de 70 ans, les exemples prouvant l'influence bénéfique des ETs se sont multipliés, ce qui a permis de mettre en lumière leur rôle essentiel dans l'évolution et plus particulièrement dans l'innovation développementale des vertébrés. Dans ce cadre, une revue bibliographique a été rédigée lors de cette thèse et publiée dans le journal *Mobile DNA* (Etchegaray et al., 2021). Cette revue avait pour objectif de présenter de nombreux exemples illustrant le rôle des éléments transposables dans l'innovation développementale des vertébrés, au travers de différents mécanismes qui sont la formation de nouvelles séquences codantes, d'ARNs non-codants et de séquences régulatrices.

REVIEW

Open Access



Transposable element-derived sequences in vertebrate development

Ema Etchegaray^{*} , Magali Naville, Jean-Nicolas Volff and Zofia Haftek-Terreau

Abstract

Transposable elements (TEs) are major components of all vertebrate genomes that can cause deleterious insertions and genomic instability. However, depending on the specific genomic context of their insertion site, TE sequences can sometimes get positively selected, leading to what are called “exaptation” events. TE sequence exaptation constitutes an important source of novelties for gene, genome and organism evolution, giving rise to new regulatory sequences, protein-coding exons/genes and non-coding RNAs, which can play various roles beneficial to the host. In this review, we focus on the development of vertebrates, which present many derived traits such as bones, adaptive immunity and a complex brain. We illustrate how TE-derived sequences have given rise to developmental innovations in vertebrates and how they thereby contributed to the evolutionary success of this lineage.

Keywords: Transposable elements, Vertebrates, Development, Genetic innovation, Exaptation, Genome evolution

Background

Transposable elements (TEs) were discovered by Barbara McClintock in the 1940s and described as moving DNA sequences that can cause genomic instability [1]. As she was able to link TE activity with variations in maize kernel colors, she coined them “controlling elements”, underlying their apparent involvement in gene regulation. TEs are nowadays known to be major components of genomes and have been found in every species that has been looked at, including prokaryotes, protists, fungi, plants and animals [2–4].

TEs are classified into two main classes according to their transposition mechanism [5, 6]. The transposition of retrotransposons (class I TEs) occurs through the reverse transcription of an RNA intermediate into a cDNA molecule that is subsequently inserted into a new locus [7, 8]. This replicative transposition process, a “copy-and-paste” mechanism called retrotransposition, leads to

the expansion of the retroelement family in the host genome. Retrotransposons gather both Long Terminal Repeat retrotransposons (LTRs), with flanking repeated sequences in direct orientation necessary for the expression and integration of the element, and non-LTR retrotransposons, also called Long Interspersed Nuclear Elements (LINEs). Autonomous retrotransposons encode a reverse transcriptase (RT) and other proteins necessary for integration (an integrase for LTRs and an endonuclease for LINEs) and other aspects of transposition [7–9]. In contrast, non-autonomous retrotransposons, including Short Interspersed Nuclear Elements (SINEs) that are mobilized by autonomous non-LTR retrotransposons, do not encode any proteins and rely on those produced *in trans* by autonomous elements to transpose [10, 11]. DNA transposons (class II TEs) do not require the reverse transcription of an RNA intermediate for their transposition [12]. They mostly use a “cut-and-paste” mechanism, the TE copy being excised from its original locus and integrated elsewhere into the genome. Many DNA transposons, including the widespread DDE transposon family, classically encode a

* Correspondence: ema.etchegaray@ens-lyon.fr

Institut de Genomique Fonctionnelle de Lyon, Univ Lyon, CNRS UMR 5242, Ecole Normale Supérieure de Lyon, Université Claude Bernard Lyon 1, 46 allée d'Italie, F-69364 Lyon, France



© The Author(s). 2021 **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

transposase (with the DDE motif forming its active site in DDE transposons) and are flanked by Terminal Inverted Repeat (TIR) sequences that are bound by the transposase for excision and integration [9, 12]. Other types of DNA transposons include Helitrons [13, 14], which are rolling-circle DNA transposons with no TIRs encoding a helicase, and Polintons/Mavericks [15, 16], which are self-synthesizing DNA transposons with long TIRs encoding a DNA polymerase. Non-autonomous elements called Miniature Inverted Repeat Transposable Elements (MITEs) are mobilized *in trans* by related autonomous DNA transposons [12].

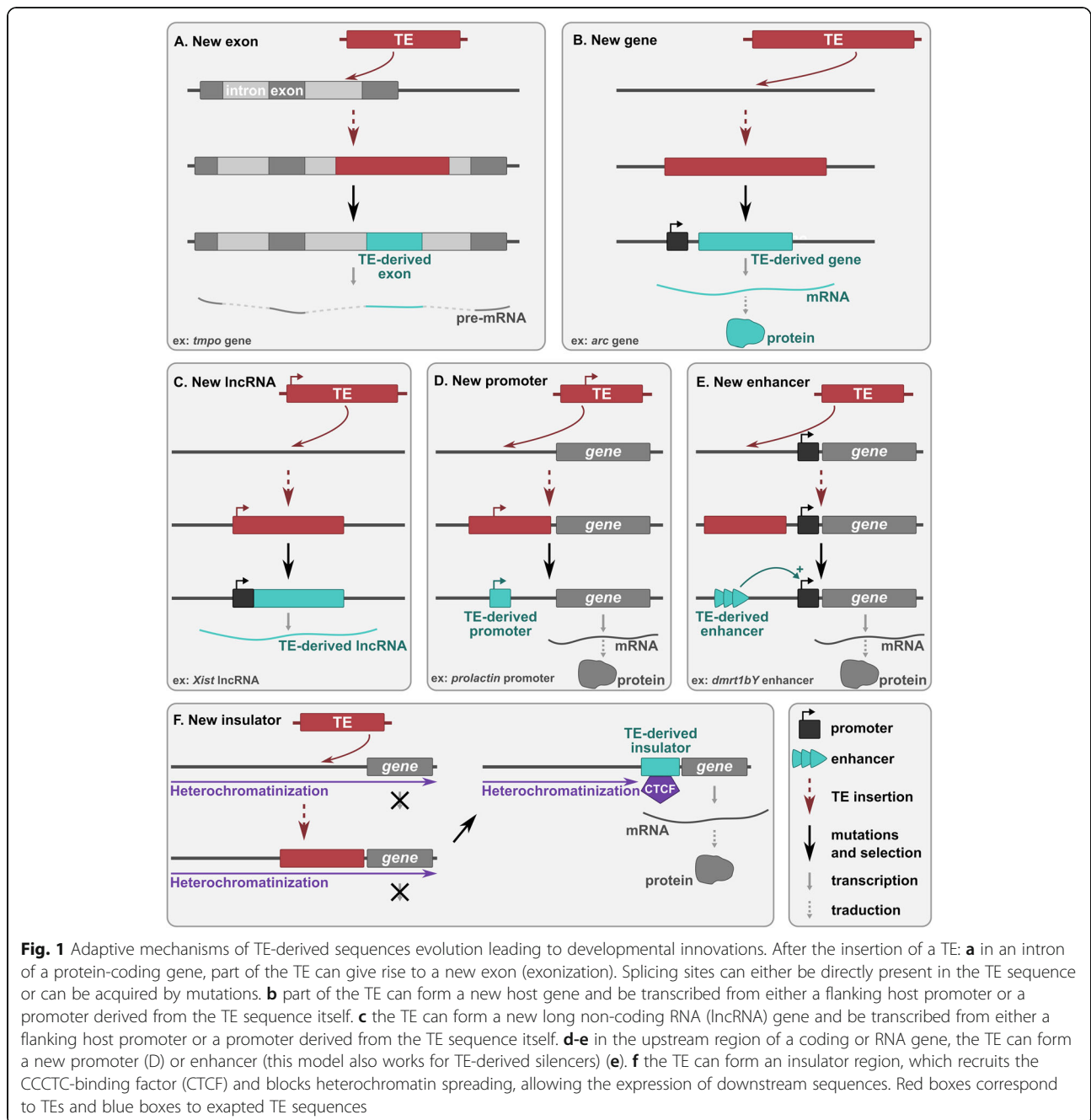
Each species genome is characterized by a specific composition in TEs, both quantitatively and qualitatively. For instance, the genome of the maize *Zea mays* is composed of nearly 85% of transposable elements [17], whereas the genome of the yeast *Saccharomyces cerevisiae* contains less than 4% of TEs [18]. In unicellular organisms, the genome of *Trichomonas vaginalis* contains almost exclusively DNA transposons, while almost only retrotransposons are found in *Entamoeba histolytica* [19, 20]. A marked variability in TE content and diversity has been also observed among vertebrates [21]. Indeed, the genomic amount of TEs ranges from 6% in the pufferfish *Tetraodon nigroviridis* up to 55% in the zebrafish *Danio rerio*. Some groups of TEs are found in most vertebrate species (LINE retrotransposons or Tc-Mariner DNA transposons for instance), whereas others are restricted to certain vertebrate sublineages and absent from others, such as the DIRS and Copia retrotransposons that are present in fish and amphibians but absent from mammals and birds [21].

Most TE insertions are thought to be either neutral or deleterious, depending on the context of the genomic region where they are inserted. TE insertions can be deleterious for instance by disrupting open reading frames (ORFs) or by altering gene transcriptional regulations. However, and despite their “selfish” characteristics, TEs are subject to the drift-selection balance and can be positively selected if they are beneficial to the host [12]. Indeed, some insertions have been shown to play a positive role in species evolution by contributing to new regulatory and coding sequences (Fig. 1) [22–28]. Such a recruitment by the host to fulfil useful functions is called exaptation or molecular domestication. The ability of TE sequences to give rise to evolutionary innovations has been more and more documented in the past years and becomes of growing interest, helped by the recent technological developments in genome sequencing and gene expression profile analysis. The structural and functional characteristics of different TE families might confer them with different potential to be exapted. TEs can contain different functional ORFs

encoding proteins with various properties such as endonucleases, integrases, transposases, reverse transcriptases and other proteins with DNA/RNA/protein-binding domains, and diverse transcriptional regulatory sequences such as promoters or enhancers. For example, LINE L1 elements contain an internal RNA polymerase II promoter and encode beside an RT an RNA-binding protein and an endonuclease; SINES in contrast do not carry any ORF and have an RNA polymerase III promoter; LTR retrotransposons present transcriptional regulatory sequences in their long terminal repeats and generally encode an integrase, a protease, a RNase H and a structural protein called GAG in addition to their RT, with an additional Envelope gene that Endogenous Retroviruses (ERVs) have occasionally kept from their infectious ancestors; DNA transposons can among others code for transposases, helicases and DNA polymerases. These functional ORFs and regulatory sequences can be reused to the host benefits. The mobilome can thus be regarded as an evolutionary toolbox, as TEs bring with them in host genomes sequences encoding proteins able to bind, replicate, cut, rearrange or degrade nucleic acids, and to associate with and modify other proteins, among other biologically relevant properties.

Vertebrates constitute a geographically widely expanded taxonomic group that appeared more than 500 million years ago and has colonized almost all ecological environments [29]. The emergence of vertebrates represents a major evolutionary transition. This group has acquired many derived traits, namely: a unique nervous system composed of a complex brain with forebrain, midbrain and hindbrain specialized regions, and cranial nerves, spinal cord and ganglia; the sensory placodes and the sensory organs they give rise to (olfactory bulbs, vestibular apparatus and otic placode for example); the neural crest, which develops into cranium, branchial skeleton and sensory ganglia; a complex endocrine system allowing the apparition of new hormones and new organs such as the placenta; bones and cartilages contributing to the skull, jaws and vertebrae; paired appendages; adaptive immunity [30–32]. These novelties, which subsequently diversified in different sublineages, have contributed to the evolutionary success of vertebrates, allowing them to improve the sense of and the move in their environment, to develop new organs and complexify them, and to turn to extensive predation.

At the origin of vertebrates, two events of whole genome duplications allowed a massive expansion of the gene repertoire [33]. However, the sole emergence of paralogous genes may not explain all the innovations that appeared, and it has been also proposed



that regulatory divergence might account for major organismal diversification [34, 35]. Accordingly, the analysis of the genome of the cephalochordate amphioxus, a sister outgroup species of vertebrates, has underlined the specialization of gene expression and the complexification of gene regulation during invertebrate to vertebrate transition, mainly due to the recruitment of new regulatory networks [36]. The precise understanding of the genetic and evolutionary mechanisms underlying this transition is of particular interest, and we propose to explore the role of TEs in

this context. Several examples of TE recruitment events crucial for vertebrate development have been documented in the last years. In this review, we discuss the different mechanisms through which TE-derived sequences have played a role in vertebrate genome evolution. We focus on selected examples illustrating the innovative potential of transposable elements as a source of new protein-coding sequences, new small and long non-coding RNA genes and new regulatory elements having driven the evolution of vertebrate development.

TE-derived sequences as new protein-coding sequences

TE exonization

Inserted TE sequences can occasionally be recruited as new exons of pre-existing genes, a process called TE exonization (Fig. 1a). Exonization is defined as the formation of a novel exon from an intronic or intergenic sequence carrying splicing sites. Such new exons can be protein-coding but might also constitute new 5' or 3' untranslated regions with possible regulatory functions.

TE exonization is not an anecdotal process and has been largely documented in mammals and other vertebrates, where it occurs more frequently than in non-vertebrate species [37–39]. In the human genome, among 233,785 exons, more than 3000 (~1%) are derived from TEs [37, 40]. Among them, about 1640 correspond to Alu SINE elements, 640 to LINES, 310 to MIRs (Mammalian-wide Interspersed Repeats, SINE elements), 300 to LTRs and 230 to DNA transposons [37]. Human exonized TEs are generally alternatively spliced, allowing protein variability [41–43]. It was also hypothesized that many TE-derived exons act as post-transcriptional gene regulators instead of being part of the protein-coding sequence itself [40]. The prevalence of Alu elements as TE-derived exons can be linked not only to their high copy number—with 1200,000 copies, they constitute as much as 10% of the human genome [44], but also to the fact that Alu sequences contain many potential splicing sites [45]. Alu elements indeed present up to ten 5' and thirteen 3' cryptic splicing sites that can be activated into functional splice sites through mutations or modifications such as adenosine-to-inosine RNA editing [38, 41]. Alu exons often modulate translational efficiency and can lead to lineage-specific regulations of gene translation [46]. Alu exonization can also cause genetic diseases in human such as the Alport syndrome, which is characterized by progressive renal failure, hearing loss and ocular abnormalities [47]. LINES and to a lesser extent LTR retroelements can be exonized too [48, 49].

Exonization of intronic insertions is influenced by multiple factors. In the human genome, exonization is promoted by large intron size, high intronic GC content, and, importantly, by the presence of young transposable elements, in particular close to transcription starting sites [50]. These factors might contribute to a decrease of RNA polymerase II elongation rate and to a reduction of spliceosomal efficiency, allowing an increase of the “window of opportunity” for spliceosomal recognition and thus for exonization. Other mechanisms inhibit Alu exonization. It has been shown in human that the RNA-binding protein hnRNP C prevents Alu exonization by avoiding the binding of splicing factor U2AF65 to Alu cryptic exons, thus blocking Alu splicing sites; this

prohibits Alu exon inclusion that would potentially lead to the formation of aberrant transcripts [51]. The binding of hnRNP C to Alu RNA is highly dependent on two poly(U) tracts present in Alu sequences inserted and transcribed in antisense orientation compared to the gene. These poly(U) arise from the antisense transcription by the gene promoter of the Alu terminal poly(A) and the internal poly(A) linker separating the two arms of Alu sequences (Alu are dimeric elements). Point mutations in these Alu poly(U) sequences are sufficient to impair the binding of hnRNP C [51]. Thus, the accumulation of mutations preventing hnRNP C binding can favor Alu exon inclusion.

Some examples illustrate well how intronic TEs can drive transcriptome and proteome diversification through the formation of lineage- and tissue-specific alternative exons. The vertebrate *lamina-associated polypeptide 2* gene (*tmpo* for *thymopoetin*) encodes several membrane protein isoforms including LAP2 β suggested to control nuclear lamina dynamics at the nuclear periphery by binding specifically to B-type lamins. Another isoform, the mammalian-specific LAP2 α protein, has a domain derived from the *gag* ORF of a DIRS1-like retrotransposon [52]. Unlike other isoforms, LAP2 α is a non-membrane protein that binds to A-type lamins in the nucleoplasm [53]. This isoform is implicated in nuclear organization dynamics during the cell cycle [54, 55]. A mutation in the TE-derived domain of LAP2 α has been associated with dilated cardiomyopathy in humans [56].

In mammals, the gene *prl3c1* belonging to the prolactin gene family encodes a cytokine expressed in uterine decidua and implicated in the establishment of pregnancy. In rodents, this gene has acquired a novel transcript variant in a common ancestor of the house mouse *Mus musculus*, *M. spretus* and *M. caroli* through the insertion of a composite TE into its first intron [57]. The inserted TE, which consists of an LTR element interrupted by a LINE, gave rise to an alternative promoter and an alternative first exon. In contrast to the “classical” transcript, the new variant is expressed in the Leydig cells of the testis. The variant protein shows a different intracellular localization and modulates the growth of testes and their capacity to produce testosterone and sperm. Such a TE co-option might contribute to the diversity of testicular development and functioning.

The *rdpoz-T1* and *rdpoz-T2* retrogenes, specifically expressed in testis and in the developing embryo in rat, and supposed to encode nuclear scaffold proteins functioning as transcription regulators, have multiple exons deriving from TE sequences [58, 59]. For example, *rdpoz-T1* has 5 out of 8 exons and an alternative polyadenylation signal that are derived from various TEs, mainly L1 and ERVs. These TE-derived exons may be

implicated in the translational regulation of these transcripts, notably through the formation of upstream ORFs [59].

The vertebrate insulin-like growth factor 1 (IGF-1) is a hormone involved in the development and growth of many tissues. IGF-1 plays a role for instance in synapse maturation and skeletal muscle development. Three isoforms of IGF-1 are known, IGF-1Ea, IGF-1Eb and IGF-1Ec [60]. The IGF-1Ea isoform is conserved among vertebrates, whereas the two others are mammal-specific and coincide with the insertion of a MIR-b SINE element that allows the formation of a fifth exon [61]. This fifth exon adds a disordered tail to IGF-1, which is highly suspected to be the source of post-translational modifications and regulatory functions. This allows a lineage-specific regulation of IGF-1.

Finally, the exonization of an Alu-J SINE element has been linked to the evolution of hemochorial placentation in anthropoid primates [62]. Hemochorial placentation is a placental implantation specific to rodents and higher order primates. In this type of placenta, the maternal blood is separated from the fetal blood by only one barrier, the chorion. This may optimize nutrient and gas exchange but makes the immune tolerance more challenging. The chorionic gonadotropin (CG) is a heterodimeric glycoprotein hormone formed by an alpha subunit, the glycoprotein hormone alpha (GPHA), and a beta subunit CGB [63]. CG is involved in the regulation of ovarian, testicular and placental functions. An Alu-J is inserted in the *gpha* gene in anthropoid primates, and its alternative exonization induces the formation of a GPHA isoform called Alu-GPHA that contains an additional N-terminus [62]. This isoform is only expressed in chorionic villus tissues and placenta, while the GPHA isoform without the Alu is expressed in other tissues. In human, the heterodimer Alu-hCG formed with the subunit Alu-GPHA shows a longer serum half-life and has a better trophoblast invasion activity compared to hCG, allowing the improvement of placenta implantation and invasion.

TE molecular domestication to form new protein-coding genes

TEs can give rise to new functional host genes, a process known as molecular domestication (Fig. 1b). In the human genome, more than hundred protein-coding genes are thought to be derived from TEs [64, 65], representing about 0.5% of the complete set of human protein-coding genes. For example, the mammalian centromere protein B (CENP-B) is derived from the transposase of a pogo-like DNA transposon [66, 67]. Like its transposase ancestor, this protein is able to bind DNA. CENP-B is involved in centromere formation during both interphase and mitosis, and directs kinetochore assembly.

Ty3/gypsy LTR retrotransposons have given rise to several multigenic gene families including the Paraneoplastic (PNMA, also called *Ma* genes, 15 genes), MART (12 genes) and SCAN families (56 genes) [68–71]. Overall, at least 103 genes derived from GAG proteins of Gypsy LTR retrotransposons have been identified in mammalian genomes, 85 being present in the human genome.

TE domestication and lymphocyte development

Two important TE-derived proteins in jawed vertebrates are RAG1 and RAG2 (Recombination Activating Gene 1 and 2) that together catalyze the V(D)J somatic recombination, a mechanism essential for the establishment of the vertebrate immune repertoire [72]. This genetic recombination, which takes place in developing lymphocytes, is at the basis of the adaptive immune system, since it allows the formation of diverse antibodies and T-cell receptors capable of specifically recognizing a great variety of pathogens. Pathogen recognition is ensured by the antigen-binding domain, which is encoded after assembling gene segments called variable (V), diversity (D) and joining (J). The joining of different V, D and J segments generates, in association with additional mutational processes, the great diversity of antibodies that can be produced by a jawed vertebrate.

RAG1 and RAG2 lymphoid-specific endonucleases are key enzymes for this somatic recombination. Both proteins associate as a recombinase to introduce double-strand breaks in DNA at recombination signal sequences (RSSs) that frame each V, D and J gene segment. This DNA cleavage resembles the transposition mechanism of DNA transposons in early steps. Indeed, the *rag1* and *rag2* genes have been derived from a *RAG* transposon related to *Transib* DNA transposons approx. 500–600 million years ago [73–75]. The RSSs recognized by RAG1/RAG2 might be derived from the TIRs of the ancestral transposon. The hypothesis is that, at the basis of deuterostomes, a *Transib* element originally containing only a *rag1* transposase might have captured an additional *rag2* ORF, leading to a *RAG* transposon with increased transposition activity [76]. By comparing vertebrate RAG proteins to a *RAG* transposon from the amphioxus genome that carries both *rag1*- and *rag2*-like genes [76, 77], putative key mutations in the domestication process, that impaired the transposition ability of the *rag* genes in the post-cleavage steps, have been identified [78]. This example of molecular domestication illustrates well how a specific genomic context may favor the selection and domestication of a transposable element. Indeed, for the emergence of the V(D)J recombination, the insertion of a TE with its RSS sequences into a gene encoding an immunoglobulin-domain receptor protein was probably a prerequisite to the formation of the ancestral fragmented antigen receptor gene [78].

TE domestication and brain development

Several retrotransposon-derived genes are implicated in vertebrate brain development, such as members of the PNMA, MART, SCAN and ARC gene families, that are all derived from *gag* genes of Ty3/gypsy LTR retrotransposons [68–71].

The *pnma10* gene (aka *sizn1/zcchc12/pnma7a*) from the PNMA gene family is involved in mouse forebrain development and mutations are associated with X-linked mental retardation in human [79]. The *pnma5* gene shows a neocortex-specific expression in primate adult brain particularly in the association areas [80]. Higher order association areas are primate-specific areas responsible for the integration of multiple inputs such as somatosensory, visuospatial, auditory and memory processes; they contribute to perception, cognition and behavior [81]. The *pnma5* gene is also present in mice but its neocortex-specific expression is not conserved. Thus, *pnma5* is thought to be one of the major genes involved in the expansion and specialization of association areas in the primate brain [80].

The protein encoded by the eutherian gene *sirh11* (aka *mart4/rtl4*), which belongs to the MART gene family, has conserved the *gag* zinc finger domain necessary for its binding to nucleic acids [70]. *Sirh11* is of crucial function for cognition [82]. Indeed, mice *sirh11* knockout mutants show impulsivity, attention and working memory defects as well as hyperactivity, suggesting a critical role in behavior. As this gene is present in eutherians only and could have conferred an essential advantage for competition by developing cognitive functions, it has been suggested to have played an important role in eutherian evolution [82].

The placental mammal gene *peg3* (*zscan24*) from the SCAN gene family has been also shown to be involved in mouse behavior [70]. This gene is paternally expressed during embryonic development and in adult brain. Its inactivation leads to growth retardation and abnormal maternal behavior for nest building, pup retrieval and crouching over pups, which can cause offspring death [83]. Moreover, mutant mothers present milk ejection defects. This phenotype has been related to a reduced number of oxytocin neurons. Growth retardation and abnormal maternal behavior are suggested to be due to impaired neuronal connectivity [83].

Finally, the *arc* tetrapod gene was shown in mice to be essential for synapse maturation and synaptic plasticity, and is involved in major neuronal processes of learning [70, 84]. *Arc* mutations have also been linked to several human disorders such as Alzheimer's disease, Angelman neurodevelopmental disease, schizophrenia and autism among others, highlighting the crucial role of the *arc* gene in brain development and functioning [85–92]. The ARC protein has conserved structural properties similar

to those of GAG proteins. Particularly, it forms capsid-like structures that transport RNA molecules across synapses and thus mediate intercellular communication between neurons [93]. Interestingly, *arc*-like genes called *darc* have been identified as duplicated copies in the genome of *Drosophila melanogaster*. Although tetrapod *arc* and *Drosophila darc* genes have been formed from Ty3/gypsy retrotransposons by independent molecular domestication events, they present similar properties of mRNA trafficking, suggesting evolutionary convergence [93, 94].

TE domestication and placenta development

TE molecular domestication probably played crucial roles in the appearance and diversification of placenta development during mammalian evolution (Fig. 2). For instance, the MART genes *peg10* (aka *mart2/rtl2*) and *peg11* (aka *mart1/rtl1*) are placental genes derived from *gag* and partial *pol* sequences of Sushi Ty3/gypsy LTR retrotransposons [95, 96]. *Peg10* influences the development of the spongiotrophoblast and labyrinth layers, which are the cell layers separating the embryo from the maternal tissues of the placenta, and *peg11* maintains the fetal capillary endothelial cells. Mutation of the *sirh7* (aka *mart7/rtl7/doc1*) gene leads to dysregulation of placental cell differentiation and maturation linked to placental hormone overproduction [97].

Syncytin genes also play a central role in placenta development. They are derived from endogenous retrovirus envelope (*env*) sequences, which encode membrane proteins that allow viral fusion with the target cells necessary for infection. The SYNCYTIN proteins have kept some properties of the ancestral ENV proteins. They are able to promote cell-cell fusion, allowing trophoblast differentiation and the formation of the syncytiotrophoblast tissue, which triggers the exchange of nutrients and gases between mother and child [98–100]. Moreover, some SYNCYTIN proteins play a role in maternal immune tolerance, this being probably linked to the capacity of parental retroviruses to target and repress immune cells thanks to the immunosuppressive activity of the ENV protein [101–103]. Indeed, at least one human (SYNCYTIN-2) and one mouse SYNCYTIN (SYNCYTIN-B) show immunosuppressive activity in vivo in mouse [104].

Among placental mammals, 14 different *syncytin* genes have been identified in different lineages presenting various placenta structures characterized by different invasion levels of the uterus by trophoblast cells. The different *syncytin* genes, their expression and their properties may play a role in the placental morphological diversity observed among mammals. In sheep, the *env* gene of a very recently endogenized Jaagsiekte Sheep Retrovirus (JSRV), present at ca. 20 copies in the

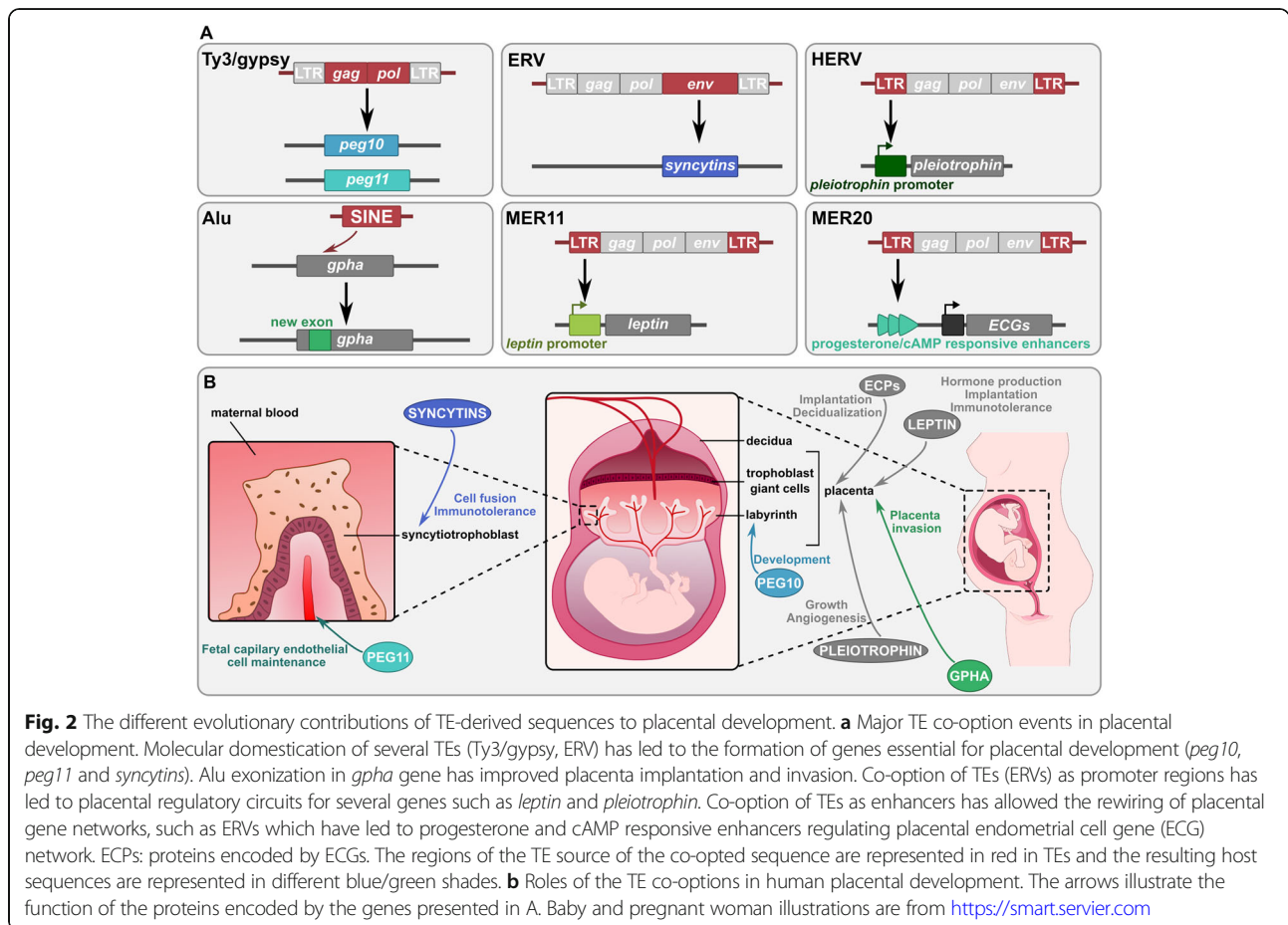


Fig. 2 The different evolutionary contributions of TE-derived sequences to placental development. **a** Major TE co-option events in placental development. Molecular domestication of several TEs (Ty3/gypsy, ERV) has led to the formation of genes essential for placental development (*peg10*, *peg11* and *syncytins*). Alu exonization in *gpha* gene has improved placenta implantation and invasion. Co-option of TEs (ERVs) as promoter regions has led to placental regulatory circuits for several genes such as *leptin* and *pleiotrophin*. Co-option of TEs as enhancers has allowed the rewiring of placental gene networks, such as ERVs which have led to progesterone and cAMP responsive enhancers regulating placental endometrial cell gene (ECG) network. ECPs: proteins encoded by ECGs. The regions of the TE source of the co-opted sequence are represented in red in TEs and the resulting host sequences are represented in different blue/green shades. **b** Roles of the TE co-options in human placental development. The arrows illustrate the function of the proteins encoded by the genes presented in A. Baby and pregnant woman illustrations are from <https://smart.servier.com>

genome, has functions similar to those of *syncytin* domesticated genes [105]. This *env* gene indeed contributes to trophoblast (first epithelium of the mammalian embryo) development and leads to pregnancy loss when downregulated. This might represent an example of a retrovirus gene being on the way of molecular domestication. Additionally, the human gene *suppressyn* has also been identified as an ERV *env*-derived gene [106]. Its protein product acts as a regulator of SYNCYTIN by binding to SYNCYTIN-1 receptor, thus inhibiting SYNCYTIN-1-mediated cell fusion.

Interestingly, *syncytin* genes in different lineages are not orthologous and have been formed by independent events of molecular domestication of ERV envelope genes, testifying for a fascinating case of convergent evolution. This underlines how TEs can represent (almost) ready-to-use molecular material that can be repurposed independently several times during the evolution of different lineages. In addition, it has been recently demonstrated that ERV *env* sequence captures are not specific of eutherian mammals, since other *syncytin* genes of independent origins have been found in marsupials and even in some viviparous lizards [107, 108].

Mammalian placenta evolution through the molecular domestication of several different retrotransposon and retrovirus genes has been proposed to follow a “baton pass” mechanism [109]. First, the early birth and high conservation of the three LTR retrotransposon-derived genes *peg10*, *peg11* and *sirh7* among mammals suggest that they could be at the origin of the primitive placenta at the base of placental mammals. Subsequently, an ancestral gene responsible for cell fusion may have been substituted by *syncytin* gene(s), which might have then replaced one another, ensuring or even improving the function and the performance of the previous *syncytin* gene, and allowing placenta morphological innovations [109, 110].

Placenta appears thus to be the place of multiple events of TE co-option. Some studies suggest that these domestications may have been facilitated by the hypomethylation of DNA in placenta compared to other tissues, allowing higher TE expression and subsequent easier TE recruitment [111, 112].

TE domestication and the diverse roles of the ZBED family

The ZBED gene family derives from *hAT* DNA transposons, and more precisely from the BED zinc finger

domain of their transposase, which is involved in DNA binding [113]. This gene family is implicated in various aspects of tissue or organ development in vertebrates. For example, the mammalian ZBED3 binds to the AXIN protein to form a complex that regulates the Wnt/ β -catenin signaling pathway, which is essential for embryogenesis and carcinogenesis [114]. In addition to the BED domain, *zbed1*, *zbed4* and *zbed6* also kept the DDE catalytic domain of the ancestral TE transposase, which contains an α -helical domain and a dimerization domain. Present in bony vertebrates, *zbed4* is proposed to be involved in retinal morphogenesis and in the functioning of Müller retinal glial cells by activating the transcription of genes expressed in Müller cells or by regulating their nuclear hormone receptors [115]. The placental mammal gene *zbed6* encodes a transcription factor essential for muscle development. A single nucleotide (nt) mutation in an *igf2* intronic sequence prevents the repression of this gene by ZBED6, leading to an increase in muscle growth and heart size and to a decrease in fat deposition [116]. ChIP-sequencing experiments have revealed about 1200 additional putative genes targeted by ZBED6, with

particular enrichment in genes involved in development, cell differentiation, morphogenesis, neurogenesis, cell-cell signaling and muscle development. Finally, the vertebrate gene *zbed1* is implicated in cell proliferation by regulating several ribosomal protein genes [117, 118].

TEs as a source of new non-coding RNA genes

TE-derived small non-coding RNAs

TE sequences can be a source of small non-coding RNAs (sncRNAs) (Fig. 1c). Several studies have shown that some sncRNAs can derive from TEs, such as microRNAs (miRNAs) [119] and Piwi-interacting RNAs (piRNAs) [120]. These sncRNAs generally constitute TE silencing factors, but they have also shown abilities to regulate host gene expression by sequence complementarity through mRNA degradation and translation inhibition (Fig. 3a). sncRNAs can also induce DNA methylation of the loci close to the nascent mRNA their target. This can induce heterochromatinization, which can spread in the targeted genomic region and thus can potentially lead to the transcriptional repression of neighboring genes (Fig. 3a) [121].

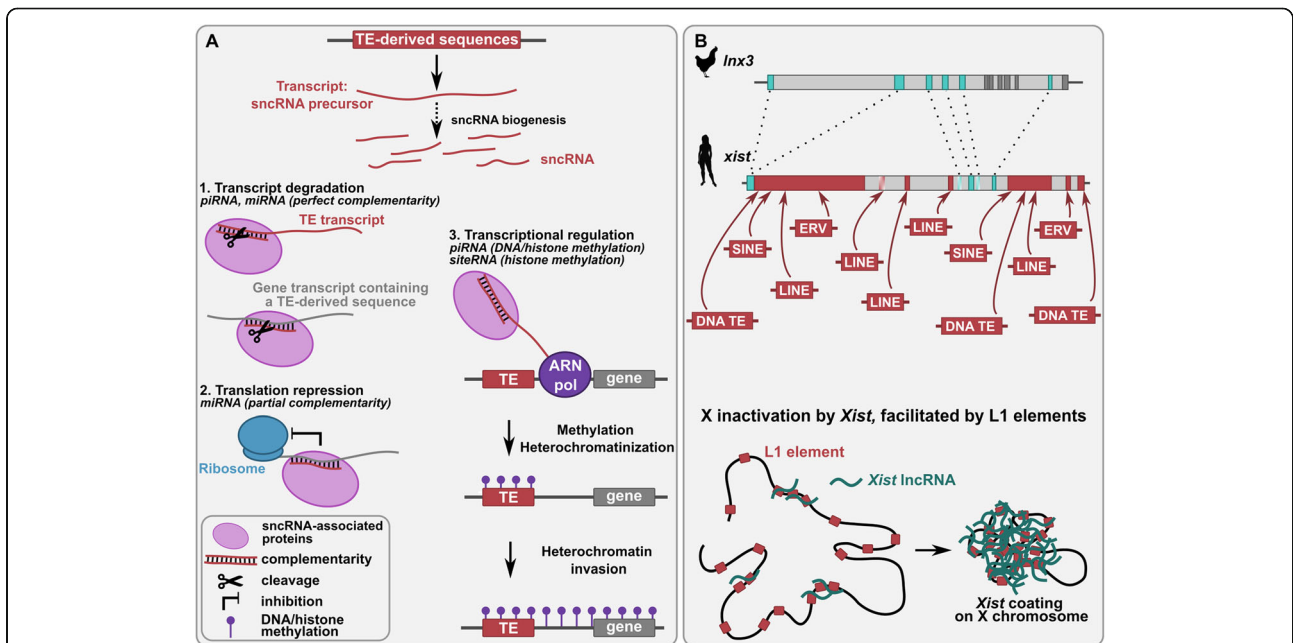


Fig. 3 Functions of TE-derived non-coding RNAs. **a** Mechanisms of action of TE-derived small non-coding RNAs (sncRNAs) through sequence complementarity. TE-derived sncRNAs are formed by fragmentation of TE-derived transcripts [122, 294], siRNAs being generated through the cleavage of the successive precursors pri-miRNAs and pre-miRNAs [122]. TE-derived sncRNAs, associated to proteins (RNA-induced silencing complex for miRNAs [122], PIWI proteins for piRNAs [150]) form double-stranded RNAs with complementarity to some RNAs of the host transcriptome, this leading to the cleavage of RNAs (1) and to the inhibition of translation (2). sncRNAs also mediates the heterochromatinization of TEs to silence them after the recruitment of DNA and histone methyltransferases (3). This heterochromatinization can spread to neighboring regions, altering their expression. **b** Evolution and function of the *xist* gene. Top: the human *xist* lncRNA gene has been formed after ancient insertions of several TEs (red boxes) into the ancestral protein-coding *Inx3* gene, which is still present in chicken. *Inx3* blue boxes represent the exons homologous to *xist* exons and dark grey boxes other exons. *Xist* shaded boxes represent human pseudo-exons (intronic regions in human but exonic in other species). Red arrows indicate TE and *xist* exon homology. Bottom: *Xist* lncRNAs coat the X chromosome, leading to X chromosome inactivation, which is facilitated by LINE-1 elements present on the chromosome [190, 191]. Silhouette images from <http://phylopic.org>

TE-derived miRNAs

TEs have contributed to the formation of miRNAs that play important roles in vertebrate developmental processes such as cell differentiation, maternal mRNA clearance and brain development [122–128]. miRNAs are sncRNAs with an average of 22 nt in length that are generated after the cleavage of 70–90 nt precursor miRNAs (pre-miRNAs), which are themselves produced by the cleavage of primary miRNA (pri-miRNA) transcripts [122]. Through complementary binding, miRNAs regulate mRNA degradation and translation. In the case of perfect sequence complementarity between miRNA and mRNA, the mRNA molecule will undergo endonucleolytic cleavage. Partial complementarity will lead to translational repression.

About 20% of human miRNAs are derived from TEs [119]. This proportion seems to be lower in other vertebrates, from 0% in the Western clawed frog to 15% in rhesus macaque and mouse [119]. In human and globally in other vertebrate species, DNA transposons make the highest contribution to miRNAs, followed by non-LTRs (LINEs and SINEs) and LTR elements; proportions that generally do not reflect the relative amount of the different types of TEs in species genomes [124, 126].

TE-derived miRNAs appear to be less conserved than non-TE-derived miRNAs, suggesting that they could constitute more lineage-specific regulators allowing the emergence of potential new phenotypes [124]. TE sequences present in the untranslated regions (UTRs) of genes constitute main targets for TE-derived miRNAs, in particular LINE1-, Alu- and MIR-derived sequences in mammals [128, 129]. The expansion of TE families such as Alu elements in primates or B1 SINEs in rodents has led to lineage-specific miRNA target sites and thus to lineage-specific regulatory potential [128].

Among the TE-derived miRNAs with a role in processes linked to development in vertebrates, miR-587, a miRNA derived from a MER element (Medium Reiteration frequency, non-autonomous DNA transposon), has been shown to be implicated in cell cycle progression in human by regulating the *tgfbr2* and *smad4* genes [130]. Another miRNA, miR-122, is involved in liver metabolic functions and is essential for the differentiation of hepatoblasts, the fetal precursor of liver cells, in zebrafish [131, 132].

Several miRNAs are involved in myeloid regulation in mouse and human. As an example, miR-652, which is derived from a MER element, is specific of myeloid lineage cells and is supposed to regulate cell identity by targeting cell type-specific regulatory proteins [133–136]. miR-935, miR-720, miR-422 and miR-378, which have been formed from different types of TEs, are all specific of one particular myeloid cell type: mucosal mast cells for miR-935, neutrophils for miR-720 and

monocytes for miR-422 and miR-378. However, their precise roles remain to be elucidated. miR-378 has also been shown to be involved in myoblast differentiation and has a pro-angiogenic and possible anti-inflammatory effect during skeletal vascularization in mice [137].

The mammalian miR-340 and miR-374, respectively derived from a *Mariner* DNA transposon and a L2 non-LTR retrotransposon, are regulators of the microtubule-associated MIDI protein, an E3 ubiquitin ligase that is an activator of the mammalian Target Of Rapamycin (mTOR) in a signaling pathway essential for cell proliferation, growth and mobility, and protein biosynthesis among others [138–140]. MIDI mutations cause the Opitz BBB/G syndrome, characterized by ventral midline malformations, with defects in heart, palate and brain structure, and hypertelorism and hypospadias [141]. In rodents, miR-374 has been shown to regulate the differentiation of myoblasts [142] and chondrocytes [143], and plays a role in retinal ganglion cell development [144]. This miRNA is also involved in primary porcine adipocyte differentiation [145] and in the production of goat hair [146].

The miR-513 subfamily, derived from a MER element, is composed of several miRNAs resulting from successive duplications in primates [147]. miR-513b regulates at both mRNA and protein levels the DR1 (down-regulator of transcription 1) protein, which is a phosphoprotein associated with TBP (TATA box-binding protein) that represses transcription. As TBP is important for spermatogenesis in mammals, miR-513b might participate in male sexual maturation by regulating DR1 [148].

TE-derived piRNAs

piRNAs are 24–31 nt long sncRNAs that together with PIWI proteins (such as MILI, MIWI and MIWI2) form complexes implicated in TE repression in the germ line and in gene regulation [149–152]. piRNA/protein complexes recognize mRNAs by complementarity with the piRNA sequence. The target mRNA is then cleaved, leading to its degradation and to the formation of secondary piRNAs that can in turn target additional complementary mRNAs. These complexes also induce DNA methylation of the regulatory regions of the mRNA they target [149, 153]. piRNA targeting is not restricted to identical sequences, this relaxed specificity increasing the number of possible targets [154]. piRNAs are major actors in TE inactivation and can thus prevent the deleterious transposition of TEs in germ cells [155]. Several studies have demonstrated the evolutionary conservation of the piRNA pathway, suggesting important functions particularly during development [156].

The origin of piRNAs is not always well characterized. piRNAs can either derive from remnant TE sequences (i.e. ancient insertions of TEs in genomic piRNA

clusters) or from single insertions of active TEs [120]. TE insertion into genes can therefore represent a way to regulate genes through their targeting by TE-derived piRNAs [157]. piRNAs might also be formed from non-TE sequences, but a very ancient TE origin not detectable at the sequence level due to divergence can often not be excluded. piRNA clusters can evolve rapidly, allowing interesting adaptation ability [158].

In mammals two populations of piRNAs are of particular importance during spermatogenesis: pre-pachytene and pachytene piRNAs, which correspond to piRNAs expressed at two distinct stages of male germ cell development [151, 159, 160]. Pre-pachytene piRNAs are expressed during early stages of spermatogenesis and in fetal and perinatal male germ cells, and are associated with the MILI and MIWI2 proteins [149, 161]. Pachytene piRNAs are produced in pachytene spermatocytes and post-meiotic spermatids, and form complexes with the MILI and MIWI proteins [160, 162]. Knockout of the proteins associated with both types of piRNAs causes male infertility [151, 159].

Most pre-pachytene piRNAs have been shown to derive from TE sequences, with SINEs (49%), LINEs (16%) and LTR elements (34%) being the main contributors in mouse [149]. They are directly involved in the de novo DNA methylation of TE sequences but also of genes and other non-TE sequences, probably through their binding to genomic DNA or nascent transcripts [153, 160, 161, 163]. Pachytene piRNAs are essential for the degradation of complementary mRNA in spermatids and maternal mRNA in early embryos, regulations that contribute to correct germ cell and embryo development. Mouse pachytene piRNAs are formed from about 3000 genomic clusters [164]; most of them target retrotransposon sequences, and more particularly SINE elements [160]. Pachytene piRNAs, some of them derived from TEs, have also been identified in bovine, macaque and human female germline and have been suggested to be involved in oogenesis and early embryogenesis [165].

TE-derived siteRNAs

A new class of sncRNAs called siteRNAs (for small intronic transposable element RNAs) has been defined in the frog *Xenopus tropicalis* [166]. These sncRNAs are 23–29 nt in length and derived from TE sequences inserted in introns of protein-coding genes. They have the ability to participate in the transcriptional silencing of the genes from which they originate by recruiting repressive histone marks (Fig. 3a). Thus, by targeting TE sequences, this TE silencing mechanism acts on regions flanking TE insertions.

TE-derived long non-coding RNAs

Long non-coding RNAs (lncRNAs) are non-coding RNAs longer than 200 nt in length. They include long

intergenic non-coding RNAs (lincRNAs) that do not overlap with protein coding-genes and make up more than half of lncRNAs in human [167]. LncRNAs can act as chromatin, transcription and post-transcription regulators through the recruitment of transcription factors and chromatin-remodeling complexes, as well as through interactions with the RNA polymerase machinery, splicing factors and mRNAs by sequence complementarity [168]. LncRNAs and more particularly lincRNAs have been shown to be implicated in many cellular [169, 170], epigenetic [171–174] and developmental processes [175], such as transcriptional silencing, cellular reprogramming and X chromosome inactivation. LncRNAs are also involved in erythroid, myeloid and lymphoid development (reviewed in [176]). They are highly expressed during central nervous system development and more particularly during neuronal and retinal differentiation, in a very time- and region-specific manner (reviewed in [177]). They are often associated to nervous system disorders.

In vertebrates, most lncRNAs in each species are lineage-specific, indicating their rapid evolutionary turnover [178, 179]. The majority of lncRNAs are thus young, and new lncRNAs are formed at a very high rate compared to protein-coding genes (ca. 100 new genes per million years in primates and rodents) [178]. lncRNA expression also seems to evolve faster than that of protein-coding genes [178, 180–182]. However, a thousand human lncRNAs are likely to have conserved functions across mammals, and hundreds beyond mammals [179].

A major part of vertebrate lncRNAs and lincRNAs contains TE-derived sequences (Fig. 1c), the estimations ranging from 50 to over 80% depending on the study and the species considered [183–186]. Within lincRNAs, which experience the same maturation steps as pre-mRNAs of protein-coding genes but are frequently poorly spliced [187], TE-derived sequences are preferentially found in introns and then in exons and promoters in mammals [185]. In a study focusing on human and mouse, the contribution of the different TE families to lncRNAs was found to reflect globally the amount of each family in the genome, except for a depletion of LINEs in lncRNA exons and promoters [185]. Within a species, the contribution of TE-derived sequences in terms of coverage can be very variable depending on the lncRNA considered. In human, TE coverage between different lncRNAs ranges from 0 to 95%, with half of lncRNAs being covered by more than 20% of TE-derived sequences [184]. Some TE-derived sequences are of functional importance by allowing notably the formation of RNA-, DNA- or protein-binding domains [188]. In human, LINE2 and MIR elements drive the nuclear enrichment of lncRNAs that allows them to modulate gene expression [186].

Even in conserved lncRNAs, sequence conservation is generally unequal along the lncRNA molecules, with small patches of high conservation separated by less constrained sequences [179]. This is consistent with a high rate of exon gain/loss and exon/intron structure modification [172]. Such a pattern might be indicative of a tolerance for sequence evolution by TE acquisition in lncRNA genes. TEs are therefore likely to be major actors of the rapid evolutionary turnover of the lncRNA repertoire in species, since they can be source of novel transcription initiation, splicing, polyadenylation and regulatory sites, as well as of new exonic sequences.

TE-derived lncRNAs in X chromosome inactivation

One best studied example of TE-containing lncRNA is *Xist*, which is involved in X-chromosome inactivation in females of eutherian mammals [189]. Inactivation of one X chromosome is essential for the dosage compensation of X-linked genes in females (XX) compared to males (XY), which have only one X chromosome. Six of the ten exons of the *Xist* lncRNA show similarities to SINES, LINES or DNA transposons [172] (Fig. 3b). Some of these TEs, particularly LINES, are essential for *Xist* addressing and for inactivation of the X chromosome in mouse [190, 191]. *Xist* lncRNA colocalizes with LINE elements and probably binds to these sequences, which cover a large part of the X chromosome [192]. These interactions are thought to be essential for the establishment of X chromosome inactivation.

The primate-specific *Xact* lncRNA is rich in repetitive elements, particularly in LTR-derived sequences [193]. *Xact* coats the active X chromosome and has been proposed to act as a transient *Xist* antagonist inhibiting inactivation. A *Xact* enhancer is derived from an ERV and is responsible for *Xact* expression in human pluripotent cells [193].

TE-derived lncRNAs in embryonic stem cells

Some TE-derived lncRNAs present a conserved expression in induced pluripotent stem cells of different primate species, suggesting an important function that remains to be uncovered [194]. Several lncRNAs are involved in maintaining embryonic stem cell pluripotency, with a particular influence of LTR-derived sequences [195–197]. For example, a human ERV-lncRNA has a domain that can recruit RNA-binding proteins, pluripotency factors and histone modifiers [197]. Human ERVs can form a hundred of lncRNAs that are specific for human pluripotent stem cells and ensure their cell identity and pluripotency [169, 183, 196, 198]. LINE1 RNAs can act as lncRNAs and chromatin regulators, and are involved in mouse embryonic stem cell self-renewal and preimplantation embryo development. These effects occur via the activation of rRNA expression and the

repression, through the recruitment of Nucleolin and Kap1/Trim28, of the *dux* developmental gene, which encodes a transcription factor activating a program specific to 2-cell embryos [199, 200].

TE-derived lncRNAs in brain development

A recently described class of lncRNAs, called SINEUPs, up-regulates translation through an embedded inverted SINE element that forms a short hairpin [201, 202]. This hairpin has been shown to be essential for the up-regulation function of SINEUP lncRNAs and serves as a recognition motif for the RNA-binding protein ILF3 (IL enhancer-binding Factor 3) [203]. The first representative member of this family, which was described in mouse, is responsible for the translational regulation of the ubiquitin carboxy-terminal hydrolase L1 (*uchl1/PARK5*), which is essential for brain function and particularly for neuron maintenance [201, 204, 205]. This SINEUP lncRNA, which carries a SINEB2 element, is antisense to *uchl1*. Another antisense SINEUP lncRNA, isolated from human brain, contains a free right Alu monomer element and increases the translation of the gene expressing the phosphatase 1 regulatory subunit 12A (PPP1R12A) [206]. PPP1R12A presents human pathogenic variants that have been associated with a congenital malformation syndrome affecting brain embryogenesis [207] and is involved in the development of the central nervous system in zebrafish [208]. More than 100 potential additional antisense SINEUP lncRNAs expressed in human brain have been identified [206], revealing other candidates for SINEUP-regulated genes involved in brain development and functioning. Interestingly, analysis of these genes indicates that different SINE elements can potentially function as effector domains in SINEUP lncRNAs [206].

Non-SINEUP examples of lncRNAs involved in brain development include the vertebrate lincRNA *cyrano*, the polyA signals of which are embedded in different TEs (LTR, SINE or LINE) depending on the transcript [184]. *Cyrano* has been shown to be essential for proper embryonic development and neurodevelopment in zebrafish [184, 209, 210]. The lincRNA *megamind* is implicated in brain morphogenesis and eye development in vertebrates. Its transcription starting site is located in a L3 LINE element in mammals, but it is not known if *megamind* uses the original promoter of the retrotransposon for its transcription [184, 209].

TE-derived sequences as a source of new regulatory elements

TE-derived sequences as new developmental cis-regulatory elements

Many studies have established the capacity of TEs to be bound by transcription factors, a property that has been

repeatedly used in host genomes to form new gene regulatory sequences and networks [27, 211] (Fig. 1d/e). For example, the ESR1, TP53, POU5F1, SOX2 or CTCF (CCCT C-binding factor) proteins are able to bind to TE sequences [211]. This ability has been shown to be essential for mammalian evolution since it can occasionally mediate the rapid expansion of transcription factor (TF) binding sites carried by the TEs and consequently the evolution of regulatory networks. As assessed by ChIP-seq technology, as much as 20% of transcription factor binding sites (TFBS) in human and mouse genomes are embedded in TEs, and this can range from 2 to 40% depending on the TF [212]. TE-derived regulatory sequences are often associated with active chromatin regions that are species-specific, suggesting their major involvement in the evolution of species-specific regulations [212]. A recent genome-wide analysis characterized human molecular pathways associated with retrotransposon-derived TFBS [213]. Olfaction, color vision, fertilization, cellular immune response, amino and fatty acids metabolism and detoxification were found to be particularly enriched for retrotransposon-derived gene regulation, i.e. mainly pathways with strong lineage/species specificity. The analysis of the association between TEs and active/repressed chromatin marks across 24 human tissues showed that SINES and DNA transposons are enriched in globally active regions, while LTRs show a more tissue-specific enrichment [214]. Moreover, TEs enriched in tissue-specific regulatory regions present binding sites for tissue-specific TFs, and their expression correlates with the tissue-specific expression of neighboring genes. This indicates that TEs can serve as a major source for regulatory sequence turnover in a tissue-specific manner, as observed in human and mouse [214, 215].

In addition to enhancers and silencers, TEs can form new gene promoters. As much as 11 and 16% of RNA polymerase II binding sites have been estimated to be derived from TEs in mouse and human genomes respectively [212]. In mouse and primates, multiple RNA polymerase II promoters have been formed from SINES, which are different from the polymerase III promoters that are classically used by these elements [216, 217]. LTR elements are also a source of new gene promoters [218], for instance in embryonic developmental genes (see below).

The *wnt5a* enhancer illustrates well the potential of TE-derived sequences in the evolution of developmental programs [219]. The *wnt5a* gene is a secreted signaling protein important for vertebrate embryogenesis [220]. This enhancer, which is essential for the morphological evolution of the mammalian secondary palate, has been formed by a combination of different TE sequences (AmnSINE1, X6b_DNA and MER117). Each TE sequence contributed to different tissue-specific enhancer

activities, cooperatively allowing an expression pattern compatible with the formation of the whole secondary palate. This example illustrates how a combination of TE-derived enhancers can generate the fine-tuned and complex diversification of developmental enhancers during evolution.

TE-derived regulatory sequences in early embryogenesis

Many TEs are involved in the expression landscape of early mouse embryos [221]. In particular, LTR elements have a strong impact on the expression of neighboring genes at earliest stages, probably through the recruitment of homeobox factors. SINE elements also induce the expression of neighboring genes during zygotic genome activation and in embryonic stem cells [221]. TEs and particularly ERVs have given rise to hundreds of thousands of primate-specific regulatory elements, and among these sequences thousands are activated specifically in embryonic cells concomitantly with neighboring genes [222]. TEs can be major actors in the formation and evolution of specific developmental regulatory networks, as demonstrated for OCT4 and NANOG, two transcription factors essential for early embryogenesis and embryonic stem cell pluripotency in mammals. A high proportion of the binding sites of these proteins are indeed derived from TEs, in particular ERV elements (21% in human and 7% in mouse for OCT4, 17% in both human and mouse for NANOG) [223].

The evolvability that TEs can confer to vertebrate developmental regulatory networks is well illustrated by mammalian embryonic stem cells. The regulatory networks of these cells are plastic, and this plasticity is at least partially due to the species-specific co-option of TEs as enhancers and promoters [223]. The potency of mouse embryonic stem cell depends on the promoter activity of MERV (murine ERV) LTRs [224]. MERV LTRs can act as promoters for two-cell stage (2C) genes, i.e. genes normally expressed in early developmental stages and repressed thereafter, this modifying cell fate. Similar results were obtained for human ERVs (HERV) [225]. HERV/LTRs can be grouped depending on the TFBS they carry. Four main patterns of TFBS were identified: binding sites for pluripotent TFs (such as SOX2, POU5F1 and NANOG), for embryonic endoderm/mesendoderm TFs (such as GATA4/6, SOX17 and FOXA1/2), for hematopoietic TFs (such as SPI1/PU1, GATA1/2 and TAL1) and for CTCF.

In vertebrates, TE-derived sequences can be targeted by Kruppel-associated box zinc finger proteins (KRAB-ZFPs) [226]. KRAB-ZFPs are early embryonic controllers that mediate the methylation of histones and DNA, inducing the repression of targeted TEs and TE-derived sequences. This can impact the expression of

neighboring genes and control regulatory networks acting during early development. Consequently, it has been proposed that the expansion of the KRAB-ZFP family results not only from the necessity of controlling TEs but could be an innovative way to build new regulatory networks through TE exaptation and controlling [226].

TE-derived regulatory sequences in brain development

SINEs are of particular importance for mammalian brain development. For instance, two SINE insertions recruited as enhancers in a mammalian common ancestor are involved in brain development [227]. The fibroblast growth factor 8 (*fgf8*) gene encodes a factor required for embryonic development, morphogenesis and particularly for normal brain, eye, ear and limb development. The first SINE insertion controls the expression of the *fgf8* gene in the diencephalon and the hypothalamus. This allows the mammalian-specific patterning of the forebrain, which is the most complex region of the vertebrate central nervous system, implicated in diverse functions such as body temperature homeostasis, sleeping, eating and reproductive function regulation, as well as in the display of emotions. The second SINE insertion regulates the *satb2* gene, which is a DNA binding protein involved in chromatin remodeling and essential for telencephalon functioning [228, 229].

An insertion of the MER130 SINE is involved in the development of the neocortex, a mammalian-specific structure responsible for the implementation of cognitive, emotive and perceptive functions [230]. This TE works as an enhancer of critical neocortical genes. A tetrapod LF-SINE-derived enhancer controls the *islet-1* (*isl1*) gene, which encodes a transcription factor essential for tetrapod brain development, particularly for motor and sensory neuron differentiation [231, 232].

Interestingly, a new regulatory function has been identified for SINEs in mouse neurons [233]. In neurons, synaptic activity influences gene expression through epigenetic modifications and the recruitment of regulatory proteins. SINE sequences located close to activity-regulated genes act as regulators for their expression. In response to neuron depolarization, these SINE sequences are acetylated, inducing the binding of the transcription factor TFIIIC. TFIIIC recruitment allows activity-dependent transcription, the relocation of inducible genes to transcription factories (i.e. specific nuclear foci where stimulation-responsive genes are expressed), as well as dendritogenesis [233]. In this context, the binding of TFIIIC to SINEs mediates the coordination of the nuclear architecture, allowing activity-dependent gene expression.

Finally, TE-derived sequences can be involved in neural gene *cis*-regulation through epigenetic modifications [234]. Indeed, TEs can be silenced by DNA methylation, which prevents transposition. This silencing can

affect surrounding sequences, altering neighboring gene expression. Hypomethylated TE-derived sequences are associated with active tissue-specific enhancer marks. This allows these sequences to gain active functions in tissue-specific gene expression [234]. This mechanism appears to be essential for the development of brain and specifically of neurons in human. For instance, the hypomethylation of the *UCON29* DNA transposon and the LF-SINE retroelement, which occurs only in fetal brain, allows the transcriptional activation of several neuron and telencephalon developmental genes specific to human [234].

TE-derived regulatory sequences in liver development

Liver developmental evolution is also linked to TE exaptation. A recent analysis of liver *cis*-regulatory elements evolution within primates distinguished two types of sequences: those conserved within primates, which represent 63% of liver *cis*-regulatory elements, and those that are not conserved, which correspond to newly evolved regulatory sequences mostly derived from TEs [235]. The majority of these sequences arose from TEs having recently transposed, particularly LTR retroelements and SINEs. Moreover, newly evolved *cis*-regulatory elements are species-specific and are associated with the species-specific binding of transcription factors involved in liver functions. They are also associated with immune- and neuro-developmental functions.

TE-derived regulatory sequences in sexual development and gametogenesis

Several examples illustrate how TEs can be involved in the control and evolution of sexual development in vertebrates. In the medaka fish *Oryzias latipes*, a DNA transposon called *Izanagi* controls the expression of the master gene regulator of male development *dmrt1bY* [236]. *dmrt1bY*, located on the medaka Y chromosome, appeared through the duplication of the autosomal *dmrt1* gene, a male gene acting downstream in the sex determination cascade. The co-option of the *Izanagi* TE-derived sequence allowed *dmrt1bY*, by inducing a new regulation, to take the lead of the sex-determining cascade of the medaka.

Estrogen receptor α , FoxA1, GATA3 and AP2 are crucial regulators of mammary gland development. The expansion of retrotransposons in mammals has given rise to thousands of binding sites for these regulators [237]. Such a spreading particularly resulted from the expansion in two phases of L2/MIR elements in a eutherian ancestor, and of ERV1 elements in simians and rodents. These retrotransposon-derived sequences act as enhancers and their recruitment allowed the establishment of the gene network of the

mammary gland regulators, allowing its morphological innovation.

LTR elements are involved in oogenesis in mammals [238]. They can form enhancers, promoters and first exon sequences of host genes and thus lead to a synchronized and developmentally regulated expression of genes. More than 800 LTR elements, mainly from the ORR1, MT, MT2 and MLT families, gave rise to promoters and first exons in mouse genes expressed in oocytes and early embryos [239]. These elements can activate the transcription of their neighboring genes during the oocyte-to-embryo transition. For example, an MTC LTR element is at the origin of the oocyte-specific high-activity isoform of Dicer (protein involved in sncRNAs biogenesis) in mouse. The deletion of this MTC element causes meiosis spindle defects and an increase of endo-siRNA target levels, and finally leads to female sterility [240]. LTR sequences are also involved in vertebrate spermatogenesis by acting as tissue-specific promoters of protein-coding and lncRNA genes [241].

TE-derived regulatory sequences in placenta development

TE sequences have been repeatedly selected, often in a lineage-specific manner, as new regulatory elements for mammalian placental development, sometimes in association with new TE-derived genes (Fig. 2). It has been shown for example that the ERV-derived *syncytin-1* is regulated by a TE-related sequence in human. Indeed, an LTR promoter combined to an adjacent cellular enhancer is responsible for the high expression of *syncytin-1* in placenta [242].

Ancient TEs have been key actors of the establishment of the decidualization, i.e. the differentiation of endometrial stromal fibroblasts into decidual stromal cells in response to different signals such as progesterone [243]. Decidualization is a key step of pregnancy establishment and maintenance, because it allows maternal-fetal communication and maternal immunotolerance. Strikingly, the exaptation of thousands of TEs has allowed the endometrial expression of numerous genes that were ancestrally expressed in other tissues [243]. Rewiring of these genes was responsible for the apparition of new functions such as immune response regulation and maternal-fetal signaling. The rewiring capacity of TEs, considered to be a major mechanism at the origin of pregnancy, was explained by the fact that they bring enhancers responsive to progesterone and cAMP, as well as TFBSs for master transcriptional regulators responsible for endometrial stromal cell-type identity [243, 244]. This was particularly suggested for the eutherian-specific MER20 DNA transposon, which has played a major role in the rewiring of the placental endometrial cell gene network [244].

More specifically, LTR promoters allow the trophoblast-specific expression of placental genes such as *pleiotrophin* and *leptin* in human [245, 246]. Pleiotrophin is a growth factor with mitogenic, growth promoting and angiogenic activities [247]. Leptin is a hormone essential for reproductive function. It is necessary for gonadotrophin hormone production, placentation and embryo implantation, and acts as an immunomodulator [248]. Another ERV (MER21A) gave rise to a placenta-specific promoter for the *cyp19* gene in primates [249, 250]. *Cyp19* encodes the aromatase P450 essential for estrogen synthesis; mutations and expression alterations of this gene are associated with reproduction abnormalities such as infertility and ovulation failure [251]. Thus, this ERV co-option is assumed to be of major importance for estrogen regulation during primate pregnancy. Finally, the promoter sequence of a LINE family is used to drive the placenta-specific expression of lncRNAs in human [252].

TE-derived enhancers are of peculiar importance for the regulation of the *prolactin (prl)* gene [253, 254]. PRL is a hormone involved in lactation as well as in the regulation of immune system, metabolism, pancreatic development and placental implantation during eutherian pregnancy. Its expression is promoted by MER20/MER39 ERV, MER77 ERV and LINE-1-derived enhancers in human, mice and elephant respectively, these regulatory sequences being progesterone- and cAMP-responsive [255]. TEs are also main contributors of the trophoblast stem cell (TSC) regulatory network, ERV retroelements forming hundreds of mouse-specific enhancers that can recruit TSC-determining factors such as CDX2, EOMES and ELF5 [256].

A two-step model has been proposed to explain the role of TEs in the evolution of mammalian placenta [112]. The first step consists in an ancestral acquisition of ERV-derived regulatory sequences responsible for the recruitment of genes to build a new network controlling placenta development, this allowing the rise of an ancestral form of placenta. Then, a relaxed repression of ERVs in trophoblast cells and the capture and replacement of *syncytin* genes facilitated the lineage-specific divergence of this network, allowing the developmental diversification of mammalian placentas that we observe today. The transient state of the placenta during life cycle may have favored its evolution and multiple TE co-options, by limiting harmful TE mutagenic activity [112].

TE-derived sequences involved in chromosomal architecture and chromatin organization

Chromosome 3D organization is essential for multiple processes such as replication, chromosome segregation during meiosis and mitosis, transcription and long-distance gene regulation, which are indispensable to ensure proper organism development [257]. Alterations in this genome

organization can lead to developmental disorders such as limb syndromes and neurodevelopmental disorders (ex. Hutchinson–Gilford progeria and Warsaw Breakage syndromes), as well as to psychiatric disorders [258–260].

It has been demonstrated that TE-derived sequences can be involved in chromosome architecture (Fig. 1f). They can provide insulator regions, which can partition the genome into topologically associated domains (TADs) and smaller chromosomal loops, and can hinder interactions between adjacent enhancers and promoters [261, 262]. CTCF, a zinc finger protein that is the only insulator protein identified so far in vertebrates, is responsible for the proper separation of different chromatin domains [263]. TEs such as SINE B2, HERV and MER20 DNA transposons can be bound by CTCF [225, 244]. Strikingly, 40% of CTCF binding sites are located in TEs in mouse genome [212]. Accordingly, it has been shown that 12–18% of human loops and 15–27% of mouse loops are indeed associated with repetitive element-derived CTCF anchor sites, the great majority of them being TEs [264].

Looking at multiple mammalian genomes, several conserved ancient retrotransposon sequences surround CTCF-binding sites, suggesting that TE expansion tens of million years ago may have given rise to mammalian and probably vertebrate conserved CTCF insulator regions [265]. On the other hand, CTCF-binding TEs have mainly enabled the species-specific expansion and diversification of CTCF binding regions in vertebrates, which are otherwise generally very constrained [265, 266]. This is likely to promote gene expression diversification between cells and between species [267], as proposed for SINE invasion in dog, rodent and opossum genomes [265]. Accordingly, multiple TEs can form chromatin loop anchors in a species-specific manner: in human, LTR, LINE and DNA transposons mostly contribute to CTCF anchors, while in the mouse SINEs, and particularly the B2 SINE family, are the main contributors [264]. Interestingly, the ChAHP complex (a protein complex constituted by the chromatin remodeler CHD4, the transcription factor ADNP and heterochromatin-binding protein HP1) binds at younger, less divergent SINE B2 elements and competes with CTCF for binding, buffering the genome architecture rewiring, associated with SINE B2 expansion in mice [268]. Most TE-derived CTCF anchors are cell-type specific, showing the potential of TEs to influence cell-type specific expression programs. TE-derived anchors are also hypomethylated, consistent with the fact that CTCF only binds unmethylated DNA.

In hominid pluripotent stem cells, HERV-H elements have been shown to be able to form TADs [269]. Deletion of HERV-H sequences induces the loss of their corresponding TADs and leads to a reduction of transcription

of upstream genes. Conversely, the insertion of novel HERV-H copies is able to form new TADs. Repression of HERV-H transcription induces TAD loss, suggesting an importance of HERV-H expression in TAD formation [269]. In the human genome, insulators can also arise from MIR retrotransposons, but in a CTCF-independent manner [270]. They are characterized by an RNA Pol III transcription and various histone modifications that can directly impact chromosomal organization.

In mouse, the SINE B2 repeat has been linked to organogenesis through its dynamic insulator activity [271]. Bidirectional transcripts of a SINE B2-derived sequence located upstream of the murine *growth hormone* gene (*gh*) are synthesized using both Pol II and Pol III promoters. These transcripts act as boundary elements by perturbing chromatin structure and inducing chromatin modifications, resulting in a change from heterochromatin to a permissive euchromatic state in this region. This transcription is both tissue- and time-specific and is responsible for the developmentally controlled expression of the *gh* gene, which promotes pituitary gland development [271]. SINE B1 elements also have insulator properties and can form heterochromatic barriers [272, 273]. It has been shown that B1 transcripts influence the chromatin state of proximal genes between embryonic stem cells and fibroblast cells, suggesting a primordial role of B1 elements in cell differentiation.

In addition to insulators, local chromatin structure is influenced by so called super-enhancers, which correspond to clusters of enhancers associated with Mediator complexes (transcriptional coactivators) that trigger the tissue-specific expression of genes [274]. A novel group of lncRNAs has recently been shown to interact with super-enhancers. These “super-lncRNAs” are able to form RNA:DNA:DNA triplex structures at specific sites within super-enhancers. Interestingly, approx. 40% of super-lncRNA binding sites in super-enhancers overlap with TEs, with SINEs and particularly Alu elements being the major contributors [274]. Moreover, it has been demonstrated that some lncRNAs can act as platforms interacting with several proteins and DNA [275]. For example, *Xist* lncRNAs can recruit Polycomb repression complex 2 [276] and also possess regions necessary for binding to DNA and transcriptional silencing [277, 278]. Thus, super-lncRNAs can possibly transport major regulators such as transcription factors and Mediator complexes to super-enhancers, influencing chromatin organization and driving surrounding tissue-specific gene expression.

Conclusions

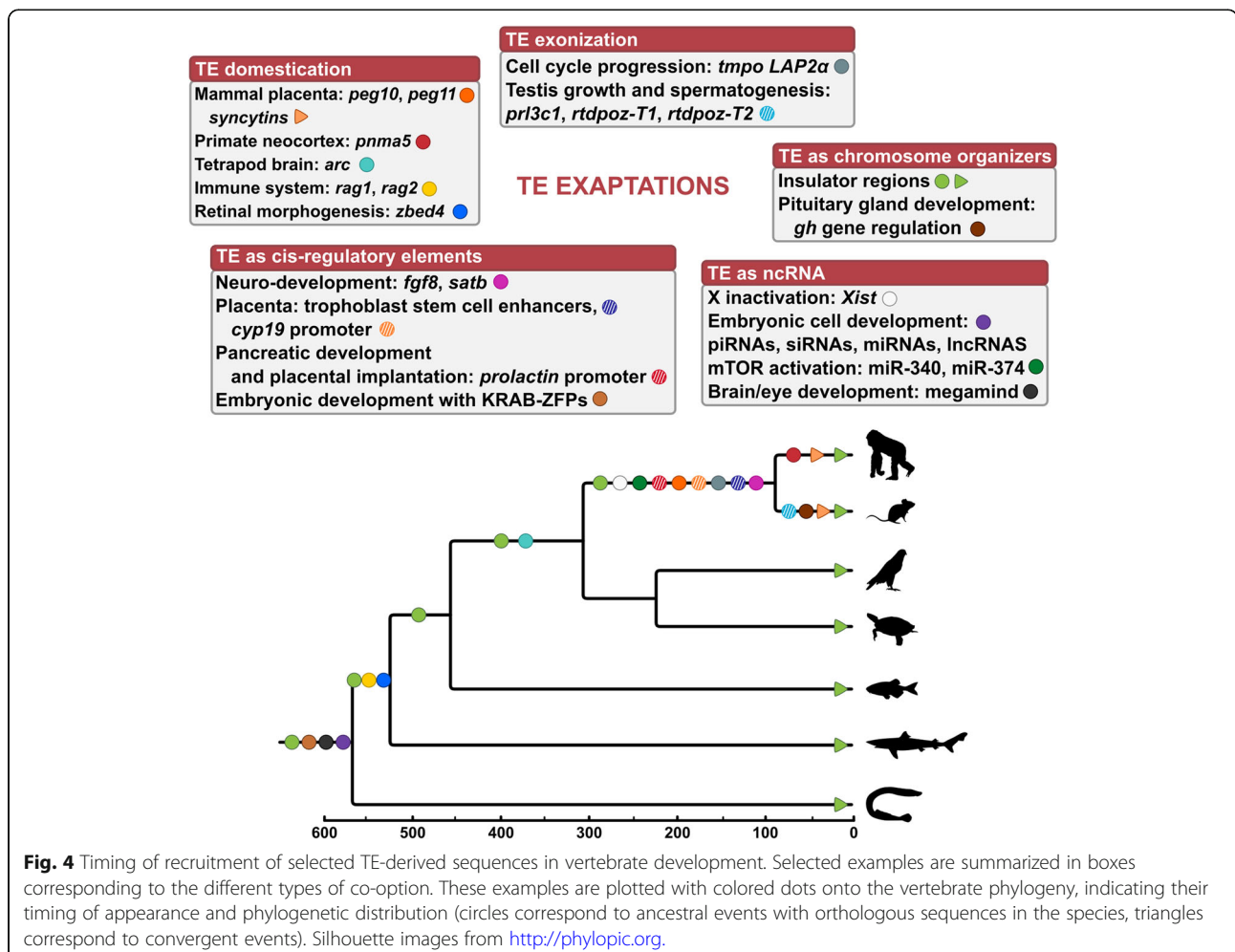
In this review, we present an overview of the multiple TE resources and functionalities that can be co-opted by host genomes (Fig. 4). TEs can be the source of

developmental innovations through their recruitment as new coding sequences and new ncRNAs, and by acting as regulatory sequences, even if TEs are probably less active in gene regulation than expected from their abundance in vertebrate genomes [215]. Particularly, TEs have been instrumental to the evolution of brain, placenta, immunity and embryonic development in vertebrates. The pace of TE recruitment in vertebrate developmental program remains to be investigated. According to the developmental gene hypothesis for punctuated equilibrium, developmental regulatory genes essential for organism morphogenesis are extremely conserved and intolerant to mutations, maintaining an equilibrium state [279]. Changes might not be progressive but rather punctuated, this being often due to transposable elements accumulation and co-option as regulatory sequences to give rise to bursts of morphological innovations and species divergence.

Concerning the formation of new genes, Ohno proposed in 1999 that gene duplication is the main mechanism shaping evolutionary transitions [33]. New genes

can also be formed from scratch, but this mechanism is very rare. We show here that TEs are a major source of material for the birth of novel protein-coding and RNA genes. In the absence of events of whole genome duplications, it has been estimated in primates that 53% of new genes originate at least partially from TE exaptation (mostly in primate-specific regions) compared to 24% from gene duplication and 5.5% de novo from non-coding sequences (the origin of the last 17.5% is still unclear) [280]. The contribution of TEs in this process is thus quantitatively important, in addition to the new functions they provide to the genome.

Several characteristics could modulate the propensity of TEs to be exapted. First, the different characteristics of each TE, such as the presence/absence of internal promoters, protein-binding motifs and ORFs encoding proteins with various properties, might favor the domestication of certain families depending on the needs of the host. For instance, ERVs have greater capacities to become gene regulatory drivers than most other TE families [215]. This has been proposed to be linked to the



frequent loss of functional internal genes in ERVs, which abolish their transposition ability but leaves LTRs in genomes that can be readily repurposed. ERVs are frequently non-repressed in hypomethylated tissues, this also possibly facilitates their recruitment. Second, the age of the TE sequences might also be of importance. Repressive silencing being relaxed in old TEs, the repression of younger elements in the genome might limit their chance to be recruited by the host. Third, the activity, copy number and diversity of a TE family probably influence its evolutionary potential for the host. Even if low copy number elements can also lead to important innovations, as shown for the *Izanagi* transposon in the sex determination cascade of the medaka fish [236], high copy number and diversity of TEs might increase the probability of generating an element advantageous for the host at both sequence and localization levels. On the other hand, maintenance of transposition activity and recombination opportunity with other TE copies might hinder the fixation of a beneficial TE-derived sequence at a specific position in the genome. Fourth, the insertion preferences of TEs or the strength of the selection pressure against their maintenance certainly impact their possible recruitment. While TEs inserting or better tolerated in gene-poor regions will probably undergo less counter-selection, they might be often silenced in heterochromatin. On the other hand, TE preferential insertion or tolerance in gene-rich regions might be more frequently deleterious but could also increase the chance of generating a beneficial combination between TE and host sequences [27]. This might for example be the case for Alu elements in primates, which are probably better tolerated than LINES in gene-rich regions due to their smaller size and therefore more frequently recruited in exaptation processes. The major factor influencing the co-option of a TE is probably the context of its insertion, as proposed for the domestication of the *Transib*-like DNA transposon at the origin of the V(D)J recombination [281]. A significant part (36.5% in the human genome) of TE-derived genes are positioned head-to-head to a host gene and share with him a bidirectional promoter containing a CpG island [282]. Since CpG islands correspond to open and actively transcribed chromatin regions, these promoters could be targeted by TE insertions and would provide them with a permissive transcriptional context for their expression, favoring the TE recruitment by the host as new transcribed sequences. TE domestication might also be facilitated by an insertion close to a promoter, or when the insertion results in a fusion with a host gene, with the TE possibly benefiting from the regulatory elements of the linked host gene if this gene is expressed in the germ line [64, 283, 284]. Fifth, if a novel TE is acquired by horizontal transfer, it will transiently escape the repression mechanisms of the

host, bringing new evolutionary potentialities and recruitment opportunities.

Developmental pathways are closely linked to those causing cancer. Illustrating this, several examples of TE-derived developmental innovations have also been associated to cancer formation. The human *syncytin-1* gene, involved in immunomodulation and cell-cell fusion in placenta, is expressed in several cancers such as colorectal and breast cancers, and endometrial carcinoma [285–287]. Several genes of the PNMA family have also been implicated in cancers, such as *pnma5* or *pnma7a*, which acts as an oncogene in thyroid cancers [288, 289]. Finally, the RAG1/RAG2 recombinase, which catalyzes the V(D)J recombination, is a driver of the genetic instability linked to lymphoblastic leukemia [290].

To conclude, Barbara McClintock's initial model [1] is now widely illustrated. In addition to form “controlling elements”, TEs are also a rich source of new host coding and RNA sequences. Most current examples illustrating the role of TE-derived sequences in vertebrate developmental innovation stems from mammals, but it is reasonable to think that TEs play also a major role in the evolution of other vertebrate species, which generally present even a higher diversity of transposable elements compared to mammals [21]. More studies in other vertebrate sub-lineages are therefore needed. For instance, an accumulation of TE sequences in the *Hox* gene clusters has been recently reported in four species of squamates (green anole lizard, slow-worm, corn snake and gecko), which contrasts with the extremely conserved structure of *Hox* clusters in other vertebrates [291, 292]. It has been suggested that these TEs may provide new coding and non-coding regions or novel regulations of transcription to the cluster genes. The emergence of such elements inside the *Hox* clusters may explain the observed morphological diversity of squamates, but this hypothesis must now be tested at the functional level [292, 293]. The accurate characterization of the whole mobilome of multiple and divergent vertebrate species, i.e. the accurate and complete genome-wide identification and annotation of TEs and TE-derived sequences in genomes along with their evolutionary and functional characteristics, is an ongoing challenge that will allow to better assess the impact of TEs on vertebrate evolution.

Abbreviations

2C: Two-Cell stage; ERV: Endogenous Retrovirus; HERV: Human Endogenous RetroVirus; JSRV: Jaagsiekte Sheep Retrovirus; KRAB-ZFP: Kruppel-associated box zinc finger proteins; lincRNA: long intergenic non-coding RNAs; LINE: Long Interspersed Nuclear Elements; lncRNA: long non-coding RNAs; LTR: Long Terminal Repeat; MER: Medium Reiteration frequency; MIR: Mammalian-wide Interspersed Repeat; miRNA: microRNA; MITE: Miniature Inverted Repeat Transposable Element; nt: nucleotide; ORF: Open Reading Frame; piRNA: PIWI-interacting RNAs; RSS: Recombination

Signal Sequence; RT: Reverse Transcriptase; SINE: Short Interspersed Nuclear Elements; siteRNA: small intronic transposable element RNA; snRNA: small non-coding RNA; TAD: Topologically Associated Domains; TBP: TATA box-binding protein; TE: Transposable Element; TF: Transcription Factor; TFBS: Transcription Factor Binding Site; TIR: Terminal Inverted Repeat; TSC: Trophoblast Stem Cell; UTR: Untranslated Region

Acknowledgements

Not applicable.

Authors' contributions

EE has drafted the initial version of the review and designed the figures; MN, JNV and ZH have contributed to the writing of the manuscript. All authors have approved the final version.

Funding

Our work is supported by grants from the French National Research Agency ANR (EVOBOOSTER project) and the Ecole Normale Supérieure de Lyon (emerging project grant) (to JNV). EE is the recipient of a competitive PhD fellowship from the French Ministry of Higher Education, Research and Innovation.

Availability of data and materials

Not applicable.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Received: 22 July 2020 Accepted: 15 December 2020

Published online: 06 January 2021

References

- McClintock B. Controlling elements and the gene. *Cold Spring Harb Symp Quant Biol.* 1956;21:197–216.
- Kazazian HH. Mobile elements: drivers of genome evolution. *Science.* 2004; 303(5664):1626–32.
- Biémont C, Vieira C. Junk DNA as an evolutionary force. *Nature.* 2006; 443(7111):521–4.
- Bourque G, Burns KH, Gehring M, Gorbunova V, Seluanov A, Hammell M, et al. Ten things you should know about transposable elements. *Genome Biol.* 2018;19(1):199.
- Wicker T, Sabot F, Hua-Van A, Bennetzen JL, Capy P, Chalhoub B, et al. A unified classification system for eukaryotic transposable elements. *Nat Rev Genet.* 2007;8(12):973–82.
- Kapitonov VV, Jurka J. A universal classification of eukaryotic transposable elements implemented in Repbase. *Nat Rev Genet.* 2008;9(5):411–2.
- Beauregard A, Curcio MJ, Belfort M. The take and give between retrotransposable elements and their hosts. *Annu Rev Genet.* 2008;42(1): 587–617.
- Goodier JL. Restricting retrotransposons: a review. *Mobile DNA.* 2016;7(1):16.
- Curcio MJ, Derbyshire KM. The outs and ins of transposition: from mu to kangaroo. *Nat Rev Mol Cell Biol.* 2003;4(11):865–77.
- Dewannieux M, Esnault C, Heidmann T. LINE-mediated retrotransposition of marked Alu sequences. *Nat Genet.* 2003;35(1):41–8.
- Richardson SR, Doucet AJ, Kopera HC, Moldovan JB, Garcia-Perez JL, Moran JV. The influence of LINE-1 and SINE retrotransposons on mammalian genomes. *Microbiol Spectr.* 2015;3(2):MDNA3–0061–2014.
- Feschotte C, Pritham EJ. DNA transposons and the evolution of eukaryotic genomes. *Annu Rev Genet.* 2007;41:331–68.
- Kapitonov VV, Jurka J. Helitrons on a roll: eukaryotic rolling-circle transposons. *Trends Genet.* 2007;23(10):521–9.
- Thomas J, Pritham EJ. Helitrons, the eukaryotic rolling-circle transposable elements. *Microbiol Spectr.* 2015;3(4):893–926.
- Kapitonov VV, Jurka J. Self-synthesizing DNA transposons in eukaryotes. *Proc Natl Acad Sci U S A.* 2006;103(12):4540–5.
- Krupovic M, Koonin EV. Polintons: a hotbed of eukaryotic virus, transposon and plasmid evolution. *Nat Rev Microbiol.* 2015;13(2):105–15.
- Schnable PS, Ware D, Fulton RS, Stein JC, Wei F, Pasternak S, et al. The B73 maize genome: complexity, diversity, and dynamics. *Science.* 2009; 326(5956):1112–5.
- Carr M, Bensasson D, Bergman CM. Evolutionary genomics of transposable elements in *Saccharomyces cerevisiae*. *PLoS ONE.* 2012;7(11):e50978.
- Pritham EJ, Feschotte C, Wessler SR. Unexpected diversity and differential success of DNA transposons in four species of Entamoeba protozoans. *Mol Biol Evol.* 2005;22(9):1751–63.
- Carlton JM, Hirt RP, Silva JC, Delcher AL, Schatz M, Zhao Q, et al. Draft genome sequence of the sexually transmitted pathogen *Trichomonas vaginalis*. *Science.* 2007;315(5809):207–12.
- Chalopin D, Naville M, Plard F, Galiana D, Volf J-N. Comparative analysis of transposable elements highlights mobilome diversity and evolution in vertebrates. *Genome Biol Evol.* 2015;7(2):567–80.
- Kidwell MG, Lisch DR. Transposable elements and host genome evolution. *Trends Ecol Evol.* 2000;15(3):95–9.
- Warren IA, Naville M, Chalopin D, Levin P, Berger CS, Galiana D, et al. Evolutionary impact of transposable elements on genomic diversity and lineage-specific innovation in vertebrates. *Chromosome Res.* 2015;23(3):505–31.
- Lee H-E, Ayarpadikannan S, Kim H-S. Role of transposable elements in genomic rearrangement, evolution, gene regulation and epigenetics in primates. *Genes Genet Syst.* 2015;90(5):245–57.
- Garcia-Perez JL, Widmann TJ, Adams IR. The impact of transposable elements on mammalian development. *Development.* 2016;143(22):4101–14.
- Chuong EB, Elde NC, Feschotte C. Regulatory evolution of innate immunity through co-option of endogenous retroviruses. *Science.* 2016;351(6277):1083–7.
- Chuong EB, Elde NC, Feschotte C. Regulatory activities of transposable elements: from conflicts to benefits. *Nat Rev Genet.* 2017;18(2):71–86.
- Jangam D, Feschotte C, Betrán E. Transposable element domestication as an adaptation to evolutionary conflicts. *Trends Genet.* 2017;33(11):817–31.
- Kumar S, Hedges SB. A molecular timescale for vertebrate evolution. *Nature.* 1998;392(6679):917–20.
- Shimeld SM, Holland PWH. Vertebrate innovations. *Proc Natl Acad Sci U S A.* 2000;97(9):4449–52.
- Khaner O. Evolutionary innovations of the vertebrates. *Integr Zool.* 2007;2(2):60–7.
- Sugahara F, Murakami Y, Pascual-Anaya J, Kuratani S. Reconstructing the ancestral vertebrate brain. *Develop Growth Differ.* 2017;59(4):163–74.
- Ohno S. Gene duplication and the uniqueness of vertebrate genomes circa 1970–1999. *Semin Cell Dev Biol.* 1999;10(5):517–22.
- King M, Wilson A. Evolution at two levels in humans and chimpanzees. *Science.* 1975;188(4184):107–16.
- Carroll SB, Grenier JK, Weatherbee SD. From DNA to diversity: molecular genetics and the evolution of animal design. 2nd ed. Malden: Blackwell Pub; 2005. p. 258.
- Marlétaz F, Firbas PN, Maeso I, Tena JJ, Bogdanovic O, Perry M, et al. Amphioxus functional genomics and the origins of vertebrate gene regulation. *Nature.* 2018;564(7734):64–70.
- Sela N, Mersch B, Gal-Mark N, Lev-Maor N, Hotz-Wagenblatt A, Ast G. Comparative analysis of transposed element insertion within human and mouse genomes reveals Alu's unique role in shaping the human transcriptome. *Genome Biol.* 2007;8(6):R127.
- Sela N, Mersch B, Hotz-Wagenblatt A, Ast G. Characteristics of transposable element exonization within human and mouse. *PLoS ONE.* 2010;5(6):e10907.
- Sela N, Kim E, Ast G. The role of transposable elements in the evolution of non-mammalian vertebrates and invertebrates. *Genome Biol.* 2010;11(6):R59.
- Piriyapongsa J, Rutledge MT, Patel S, Borodovsky M, Jordan IK. Evaluating the protein coding potential of exonized transposable element sequences. *Biol Direct.* 2007;2(1):31.
- Sorek R, Ast G, Graur D. Alu-containing exons are alternatively spliced. *Genome Res.* 2002;12(7):1060–7.
- Modrek B, Lee CJ. Alternative splicing in the human, mouse and rat genomes is associated with an increased frequency of exon creation and/or loss. *Nat Genet.* 2003;34(2):177–80.
- Alekseyenko AV, Kim N, Lee CJ. Global analysis of exon creation versus loss and the role of alternative splicing in 17 vertebrate genomes. *RNA.* 2007; 13(5):661–70.
- International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature.* 2001;409(6822):860–921.

45. Krull M, Brosius J, Schmitz J. Alu-SINE exonization: En route to protein-coding function. *Mol Biol Evol.* 2005;22(8):1702–11.
46. Shen S, Lin L, Cai JJ, Jiang P, Kenkel EJ, Stroik MR, et al. Widespread establishment and regulatory impact of Alu exons in human genes. *Proc Natl Acad Sci U S A.* 2011;108(7):2837–42.
47. Nozu K, Iijima K, Ohtsuka Y, Fu XJ, Kaito H, Nakanishi K, et al. Alport syndrome caused by a COL4A5 deletion and exonization of an adjacent AluY. *Mol Genet Genomic Med.* 2014;2(5):451–3.
48. Piriyaopongsa J, Polavarapu N, Borodovsky M, McDonald J. Exonization of the LTR transposable elements in human genome. *BMC Genomics.* 2007;8:291.
49. Attig J, Agostini F, Gooding C, Chakrabarti AM, Singh A, Haberman N, et al. Heteromeric RNP assembly at LINES controls lineage-specific RNA processing. *Cell.* 2018;174(5):1067–1081.e17.
50. Avgan N, Wang JI, Fernandez-Chamorro J, Weatheritt RJ. Multilayered control of exon acquisition permits the emergence of novel forms of regulatory control. *Genome Biol.* 2019;20(1):141.
51. Zarnack K, König J, Tajnik M, Martincorena I, Eustermann S, Stévant I, et al. Direct competition between hnRNP C and U2AF65 protects the transcriptome from the exonization of Alu elements. *Cell.* 2013;152(3):453–66.
52. Abascal F, Tress ML, Valencia A. Alternative splicing and co-option of transposable elements: the case of TMPO/LAP2 α and ZNF451 in mammals. *Bioinformatics.* 2015;31(14):2257–61.
53. Dechat T, Korbei B, Vaughan OA, Vlcek S, Hutchison CJ, Foisner R. Lamina-associated polypeptide 2 α binds intranuclear A-type lamins. *J Cell Sci.* 2000;113(Pt 19):3473–84.
54. Dechat T. Detergent-salt resistance of LAP2 α in interphase nuclei and phosphorylation-dependent association with chromosomes early in nuclear assembly implies functions in nuclear structure dynamics. *EMBO J.* 1998;17(16):4887–902.
55. Vlcek S, Just H, Dechat T, Foisner R. Functional diversity of LAP2 α and LAP2 β in postmitotic chromosome association is caused by an α -specific nuclear targeting domain. *EMBO J.* 1999;18(22):6370–84.
56. Taylor MRG, Slavov D, Gajewski A, Vlcek S, Ku L, Fain PR, et al. Thymopoietin (lamina-associated polypeptide 2) gene mutation associated with dilated cardiomyopathy. *Hum Mutat.* 2005;26(6):566–74.
57. Bu P, Yagi S, Shiota K, Alam SMK, Vivian JL, Wolfe MW, et al. Origin of a rapidly evolving homeostatic control system programming testis function. *J Endocrinol.* 2017;234(2):217–32.
58. Huang C-J, Chen C-Y, Chen H-H, Tsai S-F, Choo K-BTDPOZ. a family of bipartite animal and plant proteins that contain the TRAF (TD) and POZ/BTB domains. *Gene.* 2004;324:117–27.
59. Huang C-J, Lin W-Y, Chang C-M, Choo K-B. Transcription of the rat testis-specific Rtdpoz-T1 and -T2 retrogenes during embryo development: co-transcription and frequent exonization of transposable element sequences. *BMC Mol Biol.* 2009;10(1):74.
60. Barton ER. The ABCs of IGF-I isoforms: impact on muscle hypertrophy and implications for repair. *Appl Physiol Nutr Metab.* 2006;31(6):791–7.
61. Annibalini G, Bielli P, De Santi M, Agostini D, Guescini M, Sisti D, et al. MIR retroposon exonization promotes evolutionary variability and generates species-specific expression of IGF-1 splice variants. *Biochim Biophys Acta.* 2016;1859(5):757–68.
62. Chen H, Chen L, Wu Y, Shen H, Yang G, Deng C. The exonization and functionalization of an Alu-J element in the protein coding region of glycoprotein hormone alpha gene represent a novel mechanism to the evolution of hemochorial placentation in primates. *Mol Biol Evol.* 2017;34(12):3216–31.
63. Fournier T, Guibourdenche J, Review E-BD. hCGs: Different sources of production, different glycoforms and functions. *Placenta.* 2015;36:S60–5.
64. Volff J-N. Turning junk into gold: domestication of transposable elements and the creation of new genes in eukaryotes. *Bioessays.* 2006;28(9):913–22.
65. Alzohairy AM, Gyulai G, Jansen RK, Bahieldin A. Transposable elements domesticated and neofunctionalized by eukaryotic genomes. *Plasmid.* 2013;69(1):1–15.
66. Tudor M, Lobočka M, Goodell M, Pettitt J, O'Hare K. The pogo transposable element family of *Drosophila melanogaster*. *Mol Gen Genet.* 1992;232(1):126–34.
67. Smit AF, Riggs AD. Tiggers and DNA transposon fossils in the human genome. *Proc Natl Acad Sci U S A.* 1996;93(4):1443–8.
68. Volff J-N, Körting C, Scharl M. Ty3/Gypsy retrotransposon fossils in mammalian genomes: Did they evolve into new cellular functions? *Mol Biol Evol.* 2001;18(2):266–70.
69. Brandt J, Veith AM, Volff J-N. A family of neofunctionalized Ty3/gypsy retrotransposon genes in mammalian genomes. *Cytogenet Genome Res.* 2005;110(1–4):307–17.
70. Campillos M, Doerks T, Shah PK, Bork P. Computational characterization of multiple Gag-like human proteins. *Trends Genet.* 2006;22(11):585–9.
71. Chalopin D, Galiana D, Volff J-N. Genetic innovation in vertebrates: gypsy integrase genes and other genes derived from transposable elements. *Int J Evol Biol.* 2012;2012:1–11.
72. Thompson CB. New insights into V(D) J recombination and its role in the evolution of the immune system. *Immunity.* 1995;3(5):531–9.
73. Kapitonov W, Jurka J. RAG1 core and V(D) J recombination signal sequences were derived from Transib transposons. *PLoS Biol.* 2005;3(6):e181.
74. Kapitonov W, Koonin EV. Evolution of the RAG1-RAG2 locus: both proteins came from the same transposon. *Biol Direct.* 2015;10(1):20.
75. Carmona LM, Schatz DG. New insights into the evolutionary origins of the recombination-activating gene proteins and V(D) J recombination. *FEBS J.* 2017;284(11):1590–605.
76. Carmona LM, Fugmann SD, Schatz DG. Collaboration of RAG2 with RAG1-like proteins during the evolution of V(D) J recombination. *Genes Dev.* 2016;30(8):909–17.
77. Huang S, Tao X, Yuan S, Zhang Y, Li P, Beilinson HA, et al. Discovery of an active RAG transposon illuminates the origins of V(D) J recombination. *Cell.* 2016;166(1):102–14.
78. Zhang Y, Cheng TC, Huang G, Lu Q, Surleac MD, Mandell JD, et al. Transposon molecular domestication and the evolution of the RAG recombinase. *Nature.* 2019;569(7754):79–84.
79. Cho G, Lim Y, Golden JA. XLMR candidate mouse gene, Zcchc12 (Sizn1) is a novel marker of Cajal–Retzius cells. *Gene Expr Patterns.* 2011;11(3–4):216–20.
80. Takaji M, Komatsu Y, Watakabe A, Hashikawa T, Yamamori T. Paraneoplastic antigen-like 5 gene (PNMA5) is preferentially expressed in the association areas in a primate specific manner. *Cereb Cortex.* 2009;19(12):2865–79.
81. Yamamori T. Selective gene expression in regions of primate neocortex: Implications for cortical specialization. *Prog Neurobiol.* 2011;94(3):201–22.
82. Irie M, Yoshikawa M, Ono R, Iwafune H, Furuse T, Yamada I, et al. Cognitive function related to the Sirh11/Zcchc16 gene acquired from an LTR retrotransposon in eutherians. *PLoS Genet.* 2015;11(9):e1005521.
83. Li L, Keverne EB, Aparicio SA, Ishino F, Barton SC, Surani MA. Regulation of maternal behavior and offspring growth by paternally expressed Peg3. *Science.* 1999;284(5412):330–3.
84. Plath N, Ohana O, Dammermann B, Errington ML, Schmitz D, Gross C, et al. Arc/Arg3.1 is essential for the consolidation of synaptic plasticity and memories. *Neuron.* 2006;52(3):437–44.
85. Park S, Park JM, Kim S, Kim J-A, Shepherd JD, Smith-Hicks CL, et al. Elongation factor 2 and fragile X mental retardation protein control the dynamic translation of Arc/Arg3.1 essential for mGluR-LTD. *Neuron.* 2008;59(1):70–83.
86. Greer PL, Hanayama R, Bloodgood BL, Mardinly AR, Lipton DM, Flavell SW, et al. The Angelman Syndrome protein Ube3A regulates synapse development by ubiquitinating Arc. *Cell.* 2010;140(5):704–16.
87. Wu J, Petralia RS, Kurushima H, Patel H, Jung M, Volk L, et al. Arc/Arg3.1 regulates an endosomal pathway essential for activity-dependent β -amyloid generation. *Cell.* 2011;147(3):615–28.
88. Fromer M, Pocklington AJ, Kavanagh DH, Williams HJ, Dwyer S, Gormley P, et al. De novo mutations in schizophrenia implicate synaptic networks. *Nature.* 2014;506(7487):179–84.
89. Purcell SM, Moran JL, Fromer M, Ruderfer D, Solovieff N, Roussos P, et al. A polygenic burden of rare disruptive mutations in schizophrenia. *Nature.* 2014;506(7487):185–90.
90. Alhowikan AM. Activity-regulated cytoskeleton-associated protein dysfunction may contribute to memory disorder and earlier detection of autism spectrum disorders. *Med Princ Pract.* 2016;25(4):350–4.
91. Managò F, Mereu M, Mastwal S, Mastrogiacomo R, Scheggia D, Emanuele M, et al. Genetic disruption of Arc/Arg3.1 in mice causes alterations in dopamine and neurobehavioral phenotypes related to schizophrenia. *Cell Rep.* 2016;16(8):2116–28.
92. Pastuzyn ED, Shepherd JD. Activity-dependent Arc expression and homeostatic synaptic plasticity are altered in neurons from a mouse model of Angelman syndrome. *Front Mol Neurosci.* 2017;10:234.

93. Pastuzyn ED, Day CE, Kearns RB, Kyrke-Smith M, Taibi AV, McCormick J, et al. The neuronal gene *Arc* encodes a repurposed retrotransposon gag protein that mediates intercellular RNA transfer. *Cell*. 2018;172(1–2):275–288.e18.
94. Ashley J, Cordy B, Lucia D, Fradkin LG, Budnik V, Thomson T. Retrovirus-like gag protein *Arc1* binds RNA and traffics across synaptic boutons. *Cell*. 2018;172(1–2):262–274.e11.
95. Ono R, Nakamura K, Inoue K, Naruse M, Usami T, Wakisaka-Saito N, et al. Deletion of *Peg10*, an imprinted gene acquired from a retrotransposon, causes early embryonic lethality. *Nat Genet*. 2006;38(1):101–6.
96. Sekita Y, Wagatsuma H, Nakamura K, Ono R, Kagami M, Wakisaka N, et al. Role of retrotransposon-derived imprinted gene, *Rtl1*, in the fetomaternal interface of mouse placenta. *Nat Genet*. 2008;40(2):243–8.
97. Naruse M, Ono R, Irie M, Nakamura K, Furuse T, Hino T, et al. *Sirh7/Ldoc1* knockout mice exhibit placental P4 overproduction and delayed parturition. *Development*. 2014;141(24):4763–71.
98. Frendo J-L, Olivier D, Cheynet V, Blond J-L, Bouton O, Vidaud M, et al. Direct involvement of HERV-W Env glycoprotein in human trophoblast cell fusion and differentiation. *Mol Cell Biol*. 2003;23(10):3566–74.
99. Mallet F, Bouton O, Prudhomme S, Cheynet V, Oriol G, Bonnaud B, et al. The endogenous retroviral locus *ERVWE1* is a bona fide gene involved in hominoid placental physiology. *Proc Natl Acad Sci U S A*. 2004;101(6):1731–6.
100. Dupressoir A, Vernochet C, Harper F, Guegan J, Dessen P, Pierron G, et al. 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*. 2011;108(46):E1164–73.
101. Cianciolo G, Copeland T, Oroszlan S, Snyderman R. Inhibition of lymphocyte proliferation by a synthetic peptide homologous to retroviral envelope proteins. *Science*. 1985;230(4724):453–5.
102. Haraguchi S, Good RA, James-Yarish M, Cianciolo GJ, Day NK. Differential modulation of Th1- and Th2-related cytokine mRNA expression by a synthetic peptide homologous to a conserved domain within retroviral envelope protein. *Proc Natl Acad Sci U S A*. 1995;92(8):3611–5.
103. Schlecht-Louf G, Renard M, Mangeney M, Letzelter C, Richaud A, Ducos B, et al. Retroviral infection in vivo requires an immune escape virulence factor encrypted in the envelope protein of oncoretroviruses. *Proc Natl Acad Sci U S A*. 2010;107(8):3782–7.
104. Mangeney M, Renard M, Schlecht-Louf G, Bouallaga I, Heidmann O, Letzelter C, et al. Placental syncytins: Genetic disjunction between the fusogenic and immunosuppressive activity of retroviral envelope proteins. *Proc Natl Acad Sci U S A*. 2007;104(51):20534–9.
105. Dunlap KA, Palmarini M, Varela M, Burghardt RC, Hayashi K, Farmer JL, et al. Endogenous retroviruses regulate periimplantation placental growth and differentiation. *Proc Natl Acad Sci U S A*. 2006;103(39):14390–5.
106. Sugimoto J, Sugimoto M, Bernstein H, Jinno Y, Schust D. A novel human endogenous retroviral protein inhibits cell-cell fusion. *Sci Rep*. 2013;3(1):1462.
107. Cornelis G, Vernochet C, Carradec Q, Souquere S, Mulot B, Catzeflis F, et al. Retroviral envelope gene captures and syncytin exaptation for placentation in marsupials. *Proc Natl Acad Sci U S A*. 2015;112(5):E487–96.
108. Cornelis G, Funk M, Vernochet C, Leal F, Tarazona OA, Meurice G, et al. An endogenous retroviral envelope syncytin and its cognate receptor identified in the viviparous placental *Mabuya* lizard. *Proc Natl Acad Sci U S A*. 2017;114(51):E10991–1000.
109. Imakawa K, Nakagawa S, Miyazawa T. Baton pass hypothesis: successive incorporation of unconserved endogenous retroviral genes for placentation during mammalian evolution. *Genes Cells*. 2015;20(10):771–88.
110. Lavalie C, Cornelis G, Dupressoir A, Esnault C, Heidmann O, Vernochet C, et al. Paleovirology of 'syncytins', retroviral env genes exapted for a role in placentation. *Philos Trans R Soc Lond B Biol Sci*. 2013;368(1626):20120507.
111. Chapman V, Forrester L, Sanford J, Hastie N, Rossant J. Cell lineage-specific undermethylation of mouse repetitive DNA. *Nature*. 1984;307(5948):284–6.
112. Chuong EB. Retroviruses facilitate the rapid evolution of the mammalian placenta: Insights & Perspectives. *BioEssays*. 2013;35(10):853–61.
113. Hayward A, Ghazal A, Andersson G, Andersson L, Jern P. ZBED evolution: Repeated utilization of DNA transposons as regulators of diverse host functions. *PLoS ONE*. 2013;8(3):e59940.
114. Chen T, Li M, Ding Y, Zhang L, Xi Y, Pan W, et al. Identification of zinc-finger BED domain-containing 3 (*Zbed3*) as a novel Axin-interacting protein that activates Wnt/ β -catenin signaling. *J Biol Chem*. 2009;284(11):6683–9.
115. Saghizadeh M, Gribanova Y, Akhmedov NB, Farber DB. ZBED4, a cone and Müller cell protein in human retina, has a different cellular expression in mouse. *Mol Vis*. 2011;17:2011–8.
116. Markljung E, Jiang L, Jaffe JD, Mikkelsen TS, Wallerman O, Larhammar M, et al. ZBED6, a novel transcription factor derived from a domesticated DNA transposon regulates IGF2 expression and muscle growth. *PLoS Biol*. 2009;7(12):e1000256.
117. Ohshima N, Takahashi M, Hirose F. Identification of a human homologue of the DREF transcription factor with a potential role in regulation of the histone H1 gene. *J Biol Chem*. 2003;278(25):22928–38.
118. Yamashita D, Sano Y, Adachi Y, Okamoto Y, Osada H, Takahashi T, et al. hDREF regulates cell proliferation and expression of ribosomal protein genes. *Mol Cell Biol*. 2007;27(6):2003–13.
119. Qin S, Jin P, Zhou X, Chen L, Ma F. The role of transposable elements in the origin and evolution of microRNAs in human. *PLoS ONE*. 2015;10(6):e0131365.
120. Betel D, Sheridan R, Marks DS, Sander C. Computational analysis of mouse piRNA sequence and biogenesis. *PLoS Comput Biol*. 2007;3(11):e222.
121. Rebollo R, Karimi MM, Bilenky M, Gagnier L, Miceli-Royer K, Zhang Y, et al. Retrotransposon-induced heterochromatin spreading in the mouse revealed by insertional polymorphisms. *PLoS Genet*. 2011;7(9):e1002301.
122. Bartel DP. MicroRNAs: Genomics, biogenesis, mechanism, and function. *Cell*. 2004;116(2):281–97.
123. Smalheiser N, Torvik V. Mammalian microRNAs derived from genomic repeats. *Trends Genet*. 2005;21(6):322–6.
124. Piriyaopongsa J, Mariño-Ramírez L, Jordan IK. Origin and evolution of human microRNAs from transposable elements. *Genetics*. 2007;176(2):1323–37.
125. Piriyaopongsa J, Jordan IK. A family of human microRNA genes from miniature inverted-repeat transposable elements. *PLoS ONE*. 2007;2(2):e203.
126. Borchert GM, Holton NW, Williams JD, Hernan WL, Bishop IP, Dembosky JA, et al. Comprehensive analysis of microRNA genomic loci identifies pervasive repetitive-element origins. *Mob Genet Elements*. 2011;1(1):8–17.
127. Roberts JT, Cooper EA, Favreau CJ, Howell JS, Lane LG, Mills JE, et al. Continuing analysis of microRNA origins: Formation from transposable element insertions and noncoding RNA mutations. *Mob Genet Elements*. 2013;3(6):e27755.
128. Spengler RM, Oakley CK, Davidson BL. Functional microRNAs and target sites are created by lineage-specific transposition. *Hum Mol Genet*. 2014;23(7):1783–93.
129. Smalheiser N, Torvik V. Alu elements within human mRNAs are probable microRNA targets. *Trends Genet*. 2006;22(10):532–6.
130. Jahangirimoez M, Medlej A, Tavallaie M, Soltani B. Hsa-miR-587 regulates TGF β /SMAD signaling and promotes cell cycle progression. *Cell J*. 2019;22(2):158–64.
131. Esau C, Davis S, Murray SF, Yu XX, Pandey SK, Pear M, et al. miR-122 regulation of lipid metabolism revealed by in vivo antisense targeting. *Cell Metab*. 2006;3(2):87–98.
132. Xu R-R, Zhang C-W, Cao Y, Wang Q. miR122 deficiency inhibits differentiation of zebrafish hepatoblast into hepatocyte. *Hereditas (Beijing)*. 2013;35(4):488–94.
133. Ward JR, Heath PR, Catto JW, Whyte MKB, Milo M, Renshaw SA. Regulation of neutrophil senescence by microRNAs. *PLoS ONE*. 2011;6(1):e15810.
134. Allantaz F, Cheng DT, Bergauer T, Ravindran P, Rossier MF, Ebeling M, et al. Expression profiling of human immune cell subsets identifies miRNA-mRNA regulatory relationships correlated with cell type specific expression. *PLoS ONE*. 2012;7(1):e29979.
135. Molnár V, Érsek B, Wiener Z, Tömböl Z, Szabó PM, Igaz P, et al. MicroRNA-132 targets HB-EGF upon IgE-mediated activation in murine and human mast cells. *Cell Mol Life Sci*. 2012;69(5):793–808.
136. Gillicze AB, Wiener Z, Tóth S, Buzás E, Pállinger É, Falcone FH, et al. Myeloid-derived microRNAs, miR-223, miR27a, and miR-652, are dominant players in myeloid regulation. *BioMed Res Int*. 2014;2014:1–9.
137. Krist B, Podkalicka P, Mucha O, Mendel M, Sepioł A, Rusiecka OM, et al. miR-378a influences vascularization in skeletal muscles. *Cardiovasc Res*. 2020;116(7):1386–97.
138. Trockenbacher A, Suckow V, Foerster J, Winter J, Krauß S, Ropers H-H, et al. MID1, mutated in Opitz syndrome, encodes an ubiquitin ligase that targets phosphatase 2A for degradation. *Nat Genet*. 2001;29(3):287–94.
139. Liu E, Knutzen CA, Krauss S, Schweiger S, Chiang GG. Control of mTORC1 signaling by the Opitz syndrome protein MID1. *Proc Natl Acad Sci U S A*. 2011;108(21):8680–5.
140. Unterbruner K, Matthes F, Schilling J, Navalade R, Weber S, Winter J, et al. MicroRNAs miR-19, miR-340, miR-374 and miR-542 regulate MID1 protein expression. *PLoS ONE*. 2018;13(1):e0190437.

141. Quaderi NA, Schweiger S, Gaudenz K, Franco B, Rugarli EJ, Berger W, et al. Opitz G/BBB syndrome, a defect of midline development, is due to mutations in a new RING finger gene on Xp22. *Nat Genet.* 1997;17(3):285–91.
142. Ma Z, Sun X, Xu D, Xiong Y, Zuo B. MicroRNA, miR-374b, directly targets Myf6 and negatively regulates C2C12 myoblasts differentiation. *Biochem Biophys Res Commun.* 2015;467(4):670–5.
143. Jee YH, Wang J, Yue S, Jennings M, Clokie SJ, Nilsson O, et al. mir-374-5p, mir-379-5p, and mir-503-5p regulate proliferation and hypertrophic differentiation of growth plate chondrocytes in male rats. *Endocrinology.* 2018;159(3):1469–78.
144. Rasheed VA, Sreekanth S, Dhanesh SB, Divya MS, Divya TS, Akhila PK, et al. Developmental wave of Brn3b expression leading to RGC fate specification is synergistically maintained by miR-23a and miR-374: miR-23a and 374 in RGC differentiation. *Dev Neurobiol.* 2014;74(12):1155–71.
145. Pan S, Zheng Y, Zhao R, Yang X. miRNA-374 regulates dexamethasone-induced differentiation of primary cultures of porcine adipocytes. *Horm Metab Res.* 2013;45(07):518–25.
146. Su R, Fu S, Zhang Y, Wang R, Zhou Y, Li J, et al. Comparative genomic approach reveals novel conserved microRNAs in Inner Mongolia cashmere goat skin and longissimus dorsi. *Mol Biol Rep.* 2015;42(5):989–95.
147. Sun Z, Zhang Y, Zhang R, Qi X, Su B. Functional divergence of the rapidly evolving miR-513 subfamily in primates. *BMC Evol Biol.* 2013;13(1):255.
148. Schmidt EE, Ohbayashi T, Makino Y, Tamura T, Schibler U. Spermatid-specific overexpression of the TATA-binding protein gene involves recruitment of two potent testis-specific promoters. *J Biol Chem.* 1997;272(8):5326–34.
149. Aravin AA, Sachidanandam R, Girard A, Fejes-Toth K, Hannon GJ. Developmentally regulated piRNA clusters implicate MILI in transposon control. *Science.* 2007;316(5825):744–7.
150. Vourekas A, Zheng Q, Alexiou P, Maragkakis M, Kirino Y, Gregory BD, et al. Mili and Miwi target RNA repertoire reveals piRNA biogenesis and function of Miwi in spermiogenesis. *Nat Struct Mol Biol.* 2012;19(8):773–81.
151. Gou L-T, Dai P, Yang J-H, Xue Y, Hu Y-P, Zhou Y, et al. Pachytene piRNAs instruct massive mRNA elimination during late spermiogenesis. *Cell Res.* 2014;24(6):680–700.
152. Grivna ST, Pyhtila B, Lin H. MIWI associates with translational machinery and PIWI-interacting RNAs (piRNAs) in regulating spermatogenesis. *Proc Natl Acad Sci U S A.* 2006;103(36):13415–20.
153. Aravin AA, Sachidanandam R, Bourc'his D, Schaefer C, Pezic D, Toth KF, et al. A piRNA pathway primed by individual transposons is linked to de novo DNA methylation in mice. *Molecular Cell.* 2008;31(6):785–99.
154. Zhang P, Kang J-Y, Gou L-T, Wang J, Xue Y, Skogerboe G, et al. MIWI and piRNA-mediated cleavage of messenger RNAs in mouse testes. *Cell Res.* 2015;25(2):193–207.
155. Ernst C, Odom DT, Kutter C. The emergence of piRNAs against transposon invasion to preserve mammalian genome integrity. *Nat Commun.* 2017;8(1):1411.
156. Grimson A, Srivastava M, Fahey B, Woodcroft BJ, Chiang HR, King N, et al. Early origins and evolution of microRNAs and Piwi-interacting RNAs in animals. *Nature.* 2008;455(7217):1193–7.
157. Sarkar A, Volff J-N, Vaury C. piRNAs and their diverse roles: a transposable element-driven tactic for gene regulation? *FASEB J.* 2017;31(2):436–46.
158. Assis R, Kondrashov AS. Rapid repetitive element-mediated expansion of piRNA clusters in mammalian evolution. *Proc Natl Acad Sci U S A.* 2009;106(17):7079–82.
159. Zheng K, Wang PJ. Blockade of pachytene piRNA biogenesis reveals a novel requirement for maintaining post-meiotic germline genome integrity. *PLoS Genet.* 2012;8(11):e1003038.
160. Watanabe T, Cheng E, Zhong M, Lin H. Retrotransposons and pseudogenes regulate mRNAs and lncRNAs via the piRNA pathway in the germline. *Genome Res.* 2015;25(3):368–80.
161. Kuramochi-Miyagawa S, Watanabe T, Gotoh K, Totoki Y, Toyoda A, Ikawa M, et al. DNA methylation of retrotransposon genes is regulated by Piwi family members MIL1 and MIWI2 in murine fetal testes. *Genes Dev.* 2008;22(7):908–17.
162. Aravin A, Gaidatzis D, Pfeffer S, Lagos-Quintana M, Landgraf P, Iovino N, et al. A novel class of small RNAs bind to MIL1 protein in mouse testes. *Nature.* 2006;442(7099):203–7.
163. Fu A, Jacobs DI, Zhu Y. Epigenome-wide analysis of piRNAs in gene-specific DNA methylation. *RNA Biology.* 2014;11(10):1301–12.
164. Gan H, Lin X, Zhang Z, Zhang W, Liao S, Wang L, et al. piRNA profiling during specific stages of mouse spermatogenesis. *RNA.* 2011;17(7):1191–203.
165. Roovers EF, Rosenkranz D, Mahdipour M, Han C-T, He N, Chuva de Sousa Lopes SM, et al. Piwi proteins and piRNAs in mammalian oocytes and early embryos. *Cell Rep.* 2015;10(12):2069–82.
166. Harding JL, Horswell S, Heliot C, Armissen J, Zimmerman LB, Luscombe NM, et al. Small RNA profiling of *Xenopus* embryos reveals novel miRNAs and a new class of small RNAs derived from intronic transposable elements. *Genome Res.* 2014;24(1):96–106.
167. Ransohoff JD, Wei Y, Khavari PA. The functions and unique features of long intergenic non-coding RNA. *Nat Rev Mol Cell Biol.* 2018;19(3):143–57.
168. Bhat SA, Ahmad SM, Mumtaz PT, Malik AA, Dar MA, Urwat U, et al. Long non-coding RNAs: Mechanism of action and functional utility. *Noncoding RNA Res.* 2016;1(1):43–50.
169. Loewer S, Cabili MN, Guttman M, Loh Y-H, Thomas K, Park IH, et al. Large intergenic non-coding RNA-RoR modulates reprogramming of human induced pluripotent stem cells. *Nat Genet.* 2010;42(12):1113–7.
170. Rinn JL, Chang HY. Genome regulation by long noncoding RNAs. *Annu Rev Biochem.* 2012;81(1):145–66.
171. Brockdorff N, Ashworth A, Kay GF, McCabe VM, Norris DP, Cooper PJ, et al. The product of the mouse *Xist* gene is a 15 kb inactive X-specific transcript containing no conserved ORF and located in the nucleus. *Cell.* 1992;71(3):515–26.
172. Elisaphenko EA, Kolesnikov NN, Shevchenko AI, Rogozin IB, Nesterova TB, Brockdorff N, et al. A dual origin of the *Xist* gene from a protein-coding gene and a set of transposable elements. *PLoS ONE.* 2008;3(6):e2521.
173. Pandey RR, Mondal T, Mohammad F, Enroth S, Redrup L, Komorowski J, et al. *Kcnq1ot1* antisense noncoding RNA mediates lineage-specific transcriptional silencing through chromatin-level regulation. *Mol Cell.* 2008;32(2):232–46.
174. Nagano T, Mitchell JA, Sanz LA, Pauler FM, Ferguson-Smith AC, Feil R, et al. The Air noncoding RNA epigenetically silences transcription by targeting G9a to chromatin. *Science.* 2008;322(5908):1717–20.
175. Delás MJ, Hannon GJ. lncRNAs in development and disease: from functions to mechanisms. *Open Biol.* 2017;7(7):170121.
176. Wilkes MC, Repellin CE, Sakamoto KM. Beyond mRNA: The role of non-coding RNAs in normal and aberrant hematopoiesis. *Mol Genet Metab.* 2017;122(3):28–38.
177. Ng S-Y, Lin L, Soh BS, Stanton LW. Long noncoding RNAs in development and disease of the central nervous system. *Trends Genet.* 2013;29(8):461–8.
178. Necsulea A, Soumillon M, Warnefors M, Liechti A, Daish T, Zeller U, et al. The evolution of lncRNA repertoires and expression patterns in tetrapods. *Nature.* 2014;505(7485):635–40.
179. Hezroni H, Koppstein D, Schwartz MG, Avrutin A, Bartel DP, Ulitsky I. Principles of long noncoding RNA evolution derived from direct comparison of transcriptomes in 17 species. *Cell Rep.* 2015;11(7):1110–22.
180. Kutter C, Watt S, Stefflova K, Wilson MD, Goncalves A, Ponting CP, et al. Rapid turnover of long noncoding RNAs and the evolution of gene expression. *PLoS Genet.* 2012;8(7):e1002841.
181. Popadin K, Gutierrez-Arcelus M, Dermitzakis ET, Antonarakis SE. Genetic and epigenetic regulation of human lincRNA gene expression. *Am J Hum Genet.* 2013;93(6):1015–26.
182. Washietl S, Kellis M, Garber M. Evolutionary dynamics and tissue specificity of human long noncoding RNAs in six mammals. *Genome Res.* 2014;24(4):616–28.
183. Kelley D, Rinn J. Transposable elements reveal a stem cell-specific class of long noncoding RNAs. *Genome Biol.* 2012;13(11):R107.
184. Kapusta A, Kronenberg Z, Lynch VJ, Zhuo X, Ramsay L, Bourque G, et al. Transposable elements are major contributors to the origin, diversification, and regulation of vertebrate long noncoding RNAs. *PLoS Genet.* 2013;9(4):e1003470.
185. Kannan S, Chernikova D, Rogozin IB, Poliakov E, Managadze D, Koonin EV, et al. Transposable element insertions in long intergenic non-coding RNA genes. *Front Bioeng Biotechnol.* 2015;3:71.
186. Carlevaro-Fita J, Polidori T, Das M, Navarro C, Zoller TI, Johnson R. Ancient exapted transposable elements promote nuclear enrichment of human long noncoding RNAs. *Genome Res.* 2019;29(2):208–22.
187. Krchňáková Z, Thakur PK, Krausová M, Bieberstein N, Haberman N, Müller-McNicoll M, et al. Splicing of long non-coding RNAs primarily depends on polypyrimidine tract and 5' splice-site sequences due to weak interactions with SR proteins. *Nucleic Acids Res.* 2019;47(2):911–28.

188. Johnson R, Guigo R. The RIDL hypothesis: transposable elements as functional domains of long noncoding RNAs. *RNA*. 2014;20(7):959–76.
189. Loda A, Heard E. Xist RNA in action: Past, present, and future. *PLoS Genet*. 2019;15(9):e1008333.
190. Lyon MF. The Lyon and the LINE hypothesis. *Semin Cell Dev Biol*. 2003;14(6):313–8.
191. Tang YA, Huntley D, Montana G, Cerase A, Nesterova TB, Brockdorff N. Efficiency of Xist-mediated silencing on autosomes is linked to chromosomal domain organisation. *Epigenetics Chromatin*. 2010;3(1):10.
192. Chow JC, Ciaudo C, Fazzari MJ, Mise N, Servant N, Glass JL, et al. LINE-1 activity in facultative heterochromatin formation during X chromosome inactivation. *Cell*. 2010;141(6):956–69.
193. Casanova M, Moscatelli M, Chauvière LÉ, Huret C, Samson J, Liyakat Ali TM, et al. A primate-specific retroviral enhancer wires the XACT lncRNA into the core pluripotency network in humans. *Nat Commun*. 2019;10(1):5652.
194. Ramsay L, Marchetto MC, Caron M, Chen S-H, Busche S, Kwan T, et al. Conserved expression of transposon-derived non-coding transcripts in primate stem cells. *BMC Genomics*. 2017;18(1):214.
195. The FANTOM Consortium, Fort A, Hashimoto K, Yamada D, Salimullah M, Keya CA, et al. Deep transcriptome profiling of mammalian stem cells supports a regulatory role for retrotransposons in pluripotency maintenance. *Nat Genet*. 2014;46(6):558–66.
196. Lu X, Sachs F, Ramsay L, Jacques P-É, Góke J, Bourque G, et al. The retrovirus HERVH is a long noncoding RNA required for human embryonic stem cell identity. *Nat Struct Mol Biol*. 2014;21(4):423–5.
197. Wang J, Xie G, Singh M, Ghanbarian AT, Raskó T, Szvetnik A, et al. Primate-specific endogenous retrovirus-driven transcription defines naive-like stem cells. *Nature*. 2014;516(7531):405–9.
198. Durruthy-Durruthy J, Sebastiano V, Wossidlo M, Cepeda D, Cui J, Grow EJ, et al. The primate-specific noncoding RNA HPAT5 regulates pluripotency during human preimplantation development and nuclear reprogramming. *Nat Genet*. 2016;48(1):44–52.
199. Jachowicz JW, Bing X, Pontabry J, Bošković A, Rando OJ, Torres-Padilla M-E. LINE-1 activation after fertilization regulates global chromatin accessibility in the early mouse embryo. *Nat Genet*. 2017;49(10):1502–10.
200. Percharde M, Lin C-J, Yin Y, Guan J, Peixoto GA, Bulut-Karslioglu A, et al. A LINE1-Nucleolin partnership regulates early development and ESC identity. *Cell*. 2018;174(2):391–405.e19.
201. Zucchelli S, Fasolo F, Russo R, Cimatti L, Patrucco L, Takahashi H, et al. SINEUPs are modular antisense long non-coding RNAs that increase synthesis of target proteins in cells. *Front Cell Neurosci*. 2015;9:174.
202. Podbevšek P, Fasolo F, Bon C, Cimatti L, Reißer S, Carninci P, et al. Structural determinants of the SINE B2 element embedded in the long non-coding RNA activator of translation AS Uchl1. *Sci Rep*. 2018;8(1):3189.
203. Fasolo F, Patrucco L, Volpe M, Bon C, Peano C, Mignone F, et al. The RNA-binding protein ILF3 binds to transposable element sequences in SINEUP lncRNAs. *FASEB J*. 2019;33(12):13572–89.
204. Liu Y, Fallon L, Lashuel HA, Liu Z, Lansbury PT. The UCH-L1 gene encodes two opposing enzymatic activities that affect α -synuclein degradation and Parkinson's disease susceptibility. *Cell*. 2002;111(2):209–18.
205. Carrieri C, Cimatti L, Biagioli M, Beugnet A, Zucchelli S, Fedele S, et al. Long non-coding antisense RNA controls Uchl1 translation through an embedded SINEB2 repeat. *Nature*. 2012;491(7424):454–7.
206. Schein A, Zucchelli S, Kauppinen S, Gustinich S, Carninci P. Identification of antisense long noncoding RNAs that function as SINEUPs in human cells. *Sci Rep*. 2016;6(1):33605.
207. Hughes JJ, Alkhunaizi E, Kruszka P, Pyle LC, Grange DK, Berger SI, et al. Loss-of-function variants in PPP1R12A: from isolated sex reversal to holoprosencephaly spectrum and urogenital malformations. *Am J Hum Genet*. 2020;106(1):121–8.
208. Barresi MJF, Burton S, Dipietrantonio K, Amsterdam A, Hopkins N, Karlstrom RO. Essential genes for astroglial development and axon pathfinding during zebrafish embryogenesis. *Dev Dyn*. 2010;239(10):2603–18.
209. Ulitsky I, Shkumatava A, Jan CH, Sive H, Bartel DP. Conserved function of lincRNAs in vertebrate embryonic development despite rapid sequence evolution. *Cell*. 2011;147(7):1537–50.
210. Sarangdhar MA, Chaubey D, Srikakulam N, Pillai B. Parentally inherited long non-coding RNA Cyrano is involved in zebrafish neurodevelopment. *Nucleic Acids Res*. 2018;46(18):9726–35.
211. Bourque G, Leong B, Vega VB, Chen X, Lee YL, Srinivasan KG, et al. Evolution of the mammalian transcription factor binding repertoire via transposable elements. *Genome Res*. 2008;18(11):1752–62.
212. Sundaram V, Cheng Y, Ma Z, Li D, Xing X, Edge P, et al. Widespread contribution of transposable elements to the innovation of gene regulatory networks. *Genome Res*. 2014;24(12):1963–76.
213. Nikitin D, Garazha A, Sorokin M, Penzar D, Tkachev V, Markov A, et al. Retroelement—linked transcription factor binding patterns point to quickly developing molecular pathways in human evolution. *Cells*. 2019;8(2):130.
214. Trizzino M, Kapusta A, Brown CD. Transposable elements generate regulatory novelty in a tissue-specific fashion. *BMC Genomics*. 2018;19(1):468.
215. Simonti CN, Pavličev M, Capra JA. Transposable element exaptation into regulatory regions is rare, influenced by evolutionary age, and subject to pleiotropic constraints. *Mol Biol Evol*. 2017;34(11):2856–69.
216. Ferrigno O, Virolle T, Djabari Z, Ortonne J-P, White RJ, Aberdam D. Transposable B2 SINE elements can provide mobile RNA polymerase II promoters. *Nat Genet*. 2001;28(1):77–81.
217. Shankar R, Grover D, Brahmachari SK, Mukerji M. Evolution and distribution of RNA polymerase II regulatory sites from RNA polymerase III dependant mobile Alu elements. *BMC Evol Biol*. 2004;4(1):37.
218. Cohen CJ, Lock WM, Mager DL. Endogenous retroviral LTRs as promoters for human genes: A critical assessment. *Gene*. 2009;448(2):105–14.
219. Nishihara H, Kobayashi N, Kimura-Yoshida C, Yan K, Bormuth O, Ding Q, et al. Coordinately co-opted multiple transposable elements constitute an enhancer for *wnt5a* expression in the mammalian secondary palate. *PLoS Genet*. 2016;12(10):e1006380.
220. Yamaguchi TP, Bradley A, McMahon AP, Jones S. A *Wnt5a* pathway underlies outgrowth of multiple structures in the vertebrate embryo. *Development*. 1999;126(6):1211–23.
221. Ge SX. Exploratory bioinformatics investigation reveals importance of “junk” DNA in early embryo development. *BMC Genomics*. 2017;18(1):200.
222. Jacques P-É, Jeyakani J, Bourque G. The majority of primate-specific regulatory sequences are derived from transposable elements. *PLoS Genet*. 2013;9(5):e1003504.
223. Kunarso G, Chia N-Y, Jeyakani J, Hwang C, Lu X, Chan Y-S, et al. Transposable elements have rewired the core regulatory network of human embryonic stem cells. *Nat Genet*. 2010;42(7):631–4.
224. Macfarlan TS, Gifford WD, Driscoll S, Lettieri K, Rowe HM, Bonanomi D, et al. Embryonic stem cell potency fluctuates with endogenous retrovirus activity. *Nature*. 2012;487(7405):57–63.
225. Ito J, Sugimoto R, Nakaoka H, Yamada S, Kimura T, Hayano T, et al. Systematic identification and characterization of regulatory elements derived from human endogenous retroviruses. *PLoS Genet*. 2017;13(7):e1006883.
226. Ecco G, Cassano M, Kauzlaric A, Duc J, Coluccio A, Offner S, et al. Transposable elements and their KRAB-ZFP controllers regulate gene expression in adult tissues. *Dev Cell*. 2016;36(6):611–23.
227. Sasaki T, Nishihara H, Hirakawa M, Fujimura K, Tanaka M, Kokubo N, et al. Possible involvement of SINEs in mammalian-specific brain formation. *Proc Natl Acad Sci U S A*. 2008;105(11):4220–5.
228. Alcamo EA, Chirivella L, Dautzenberg M, Dobрева G, Fariñas I, Grosschedl R, et al. *Satb2* regulates callosal projection neuron identity in the developing cerebral cortex. *Neuron*. 2008;57(3):364–77.
229. Britanova O, de Juan Romero C, Cheung A, Kwan KY, Schwark M, Gyorgy A, et al. *Satb2* is a postmitotic determinant for upper-layer neuron specification in the neocortex. *Neuron*. 2008;57(3):378–92.
230. Notwell JH, Chung T, Heavner W, Bejerano G. A family of transposable elements co-opted into developmental enhancers in the mouse neocortex. *Nat Commun*. 2015;6(1):6644.
231. Uemura O, Okada Y, Ando H, Guedj M, Higashijima S, Shimazaki T, et al. Comparative functional genomics revealed conservation and diversification of three enhancers of the *isl1* gene for motor and sensory neuron-specific expression. *Dev Biol*. 2005;278(2):587–606.
232. Bejerano G, Lowe CB, Ahituv N, King B, Siepel A, Salama SR, et al. A distal enhancer and an ultraconserved exon are derived from a novel retroposon. *Nature*. 2006;441(7089):87–90.
233. Crepaldi L, Policarpi C, Coatti A, Sherlock WT, Jongbloets BC, Down TA, et al. Binding of TFIIC to SINE elements controls the relocation of activity-dependent neuronal genes to transcription factories. *PLoS Genet*. 2013;9(8):e1003699.

234. Xie M, Hong C, Zhang B, Lowdon RF, Xing X, Li D, et al. DNA hypomethylation within specific transposable element families associates with tissue-specific enhancer landscape. *Nat Genet.* 2013;45(7):836–41.
235. Trizzino M, Park Y, Holsbach-Beltrame M, Aracena K, Mika K, Caliskan M, et al. Transposable elements are the primary source of novelty in primate gene regulation. *Genome Res.* 2017;27(10):1623–33.
236. Herpin A, Braasch I, Kraeussling M, Schmidt C, Thoma EC, Nakamura S, et al. Transcriptional rewiring of the sex determining *dmrt1* gene duplicate by transposable elements. *PLoS Genet.* 2010;6(2):e1000844.
237. Nishihara H. Retrotransposons spread potential cis-regulatory elements during mammary gland evolution. *Nucleic Acids Res.* 2019;47(22):11551–62.
238. Peaston AE, Evsikov AV, Graber JH, de Vries WN, Holbrook AE, Solter D, et al. Retrotransposons regulate host genes in mouse oocytes and preimplantation embryos. *Dev Cell.* 2004;7(4):597–606.
239. Franke V, Ganesh S, Karlic R, Malik R, Pasulka J, Horvat F, et al. Long terminal repeats power evolution of genes and gene expression programs in mammalian oocytes and zygotes. *Genome Res.* 2017;27(8):1384–94.
240. Flemr M, Malik R, Franke V, Nejezinska J, Sedlacek R, Vlahovicek K, et al. A retrotransposon-driven dicer isoform directs endogenous small interfering RNA production in mouse oocytes. *Cell.* 2013;155(4):807–16.
241. Davis MP, Carrieri C, Saini HK, Dongen S, Leonardi T, Bussotti G, et al. Transposon-driven transcription is a conserved feature of vertebrate spermatogenesis and transcript evolution. *EMBO Rep.* 2017;18(7):1231–47.
242. Prudhomme S, Oriol G, Mallet F. A retroviral promoter and a cellular enhancer define a bipartite element which controls *env* ERWE1 placental expression. *J Virol.* 2004;78(22):12157–68.
243. Lynch VJ, Nnamani MC, Kapusta A, Brayer K, Plaza SL, Mazur EC, et al. Ancient transposable elements transformed the uterine regulatory landscape and transcriptome during the evolution of mammalian pregnancy. *Cell Rep.* 2015;10(4):551–61.
244. Lynch VJ, Leclerc RD, May G, Wagner GP. Transposon-mediated rewiring of gene regulatory networks contributed to the evolution of pregnancy in mammals. *Nat Genet.* 2011;43(11):1154–9.
245. Schulte AM, Lai S, Kurtz A, Czubyko F, Riegel AT, Wellstein A. Human trophoblast and choriocarcinoma expression of the growth factor pleiotrophin attributable to germ-line insertion of an endogenous retrovirus. *Proc Natl Acad Sci.* 1996;93(25):14759–64.
246. Bi S, Gavrilova O, Gong D-W, Mason MM, Reitman M. Identification of a placental enhancer for the human leptin gene. *J Biol Chem.* 1997;272(48):30583–8.
247. Ball M, Carmody M, Wynne F, Dockery P, Aigner A, Cameron I, et al. Expression of pleiotrophin and its receptors in human placenta suggests roles in trophoblast life cycle and angiogenesis. *Placenta.* 2009;30(7):649–53.
248. Pérez-Pérez A, Toro A, Vilariño-García T, Maymó J, Guadix P, Dueñas JL, et al. Leptin action in normal and pathological pregnancies. *J Cell Mol Med.* 2017;22(2):716–27.
249. Kamat A, Hinshelwood MM, Murry BA, Mendelson CR. Mechanisms in tissue-specific regulation of estrogen biosynthesis in humans. *Trends Endocrinol Metab.* 2002;13(3):122–8.
250. van de Lagemaat LN, Landry J-R, Mager DL, Medstrand P. Transposable elements in mammals promote regulatory variation and diversification of genes with specialized functions. *Trends Genet.* 2003;19(10):530–6.
251. Stocco C. Tissue physiology and pathology of aromatase. *Steroids.* 2012;77(1–2):27–35.
252. Chishima T, Iwakiri J, Hamada M. Identification of transposable elements contributing to tissue-specific expression of long non-coding RNAs. *Genes.* 2018;9(1):23.
253. Gerlo S, Davis JRE, Mager DL, Kooijman R. Prolactin in man: a tale of two promoters. *Bioessays.* 2006;28(10):1051–5.
254. Jabbour H, Critchley H. Potential roles of decidual prolactin in early pregnancy. *Reproduction.* 2001;121(2):197–205.
255. Emera D, Casola C, Lynch VJ, Wildman DE, Agnew D, Wagner GP. Convergent evolution of endometrial prolactin expression in primates, mice, and elephants through the independent recruitment of transposable elements. *Mol Biol Evol.* 2012;29(1):239–47.
256. Chuong EB, Rumi MAK, Soares MJ, Baker JC. Endogenous retroviruses function as species-specific enhancer elements in the placenta. *Nat Genet.* 2013;45(3):325–9.
257. Zheng H, Xie W. The role of 3D genome organization in development and cell differentiation. *Nat Rev Mol Cell Biol.* 2019;20(9):535–50.
258. Lupiáñez DG, Kraft K, Heinrich V, Krawitz P, Brancati F, Klopocki E, et al. Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions. *Cell.* 2015;161(5):1012–25.
259. Medrano-Fernández A, Barco A. Nuclear organization and 3D chromatin architecture in cognition and neuropsychiatric disorders. *Mol Brain.* 2016;9(1):83.
260. Davis L, Onn I, Elliott E. The emerging roles for the chromatin structure regulators CTCF and cohesin in neurodevelopment and behavior. *Cell Mol Life Sci.* 2018;75(7):1205–14.
261. Udvardy A. Dividing the empire: boundary chromatin elements delimit the territory of enhancers. *EMBO J.* 1999;18(1):1–8.
262. Dixon JR, Selvaraj S, Yue F, Kim A, Li Y, Shen Y, et al. Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature.* 2012;485(7398):376–80.
263. Bell AC, West AG, Felsenfeld G. The protein CTCF is required for the enhancer blocking activity of vertebrate insulators. *Cell.* 1999;98(3):387–96.
264. Choudhary MN, Friedman RZ, Wang JT, Jang HS, Zhuo X, Wang T. Co-opted transposons help perpetuate conserved higher-order chromosomal structures. *Genome Biol.* 2020;21(1):16.
265. Schmidt D, Schwalie PC, Wilson MD, Ballester B, Gonçalves Â, Kutter C, et al. Waves of retrotransposon expansion remodel genome organization and CTCF binding in multiple mammalian lineages. *Cell.* 2012;148(1–2):335–48.
266. Thybert D, Roller M, FCP N, Fiddes I, Streeter I, Feig C, et al. Repeat associated mechanisms of genome evolution and function revealed by the *Mus caroli* and *Mus pahari* genomes. *Genome Res.* 2018;28(4):448–59.
267. Diehl AG, Ouyang N, Boyle AP. Transposable elements contribute to cell and species-specific chromatin looping and gene regulation in mammalian genomes. *Nat Commun.* 2020;11(1):1796.
268. Kaaij LJT, Mohn F, van der Weide RH, de Wit E, Bühler M. The ChAHP Complex Counteracts Chromatin Looping at CTCF Sites that Emerged from SINE Expansions in Mouse. *Cell.* 2019;178(6):1437–1451.e14.
269. Zhang Y, Li T, Preissl S, Amaral ML, Grinstein JD, Farah EN, et al. Transcriptionally active HERV-H retrotransposons demarcate topologically associating domains in human pluripotent stem cells. *Nat Genet.* 2019;51(9):1380–8.
270. Wang J, Vicente-García C, Seruggia D, Moltó E, Fernandez-Miñán A, Neto A, et al. MIR retrotransposon sequences provide insulators to the human genome. *Proc Natl Acad Sci U S A.* 2015;112(32):E4428–37.
271. Lunyak VV, Prefontaine GG, Núñez E, Cramer T, Ju B-G, Ohgi KA, et al. Developmentally regulated activation of a SINE B2 repeat as a domain boundary in organogenesis. *Science.* 2007;317(5835):248–51.
272. Roman AC, Benitez DA, Carvajal-Gonzalez JM, Fernandez-Salguero PM. Genome-wide B1 retrotransposon binds the transcription factors dioxin receptor and Slug and regulates gene expression in vivo. *Proc Natl Acad Sci U S A.* 2008;105(5):1632–7.
273. Roman AC, Gonzalez-Rico FJ, Molto E, Hernando H, Neto A, Vicente-García C, et al. Dioxin receptor and SLUG transcription factors regulate the insulator activity of B1 SINE retrotransposons via an RNA polymerase switch. *Genome Res.* 2011;21(3):422–32.
274. Soibam B. Super-lncRNAs: identification of lncRNAs that target super-enhancers via RNA:DNA:DNA triplex formation. *RNA.* 2017;23(11):1729–42.
275. Engreitz JM, Ollikainen N, Guttman M. Long non-coding RNAs: spatial amplifiers that control nuclear structure and gene expression. *Nat Rev Mol Cell Biol.* 2016;17(12):756–70.
276. da Rocha ST, Boeva V, Escamilla-Del-Arenal M, Ancelin K, Granier C, Matias NR, et al. *Jarid2* is implicated in the initial Xist-induced targeting of PRC2 to the inactive X chromosome. *Molecular Cell.* 2014;53(2):301–16.
277. Beletskii A, Hong Y-K, Pehrson J, Egholm M, Strauss WM. PNA interference mapping demonstrates functional domains in the noncoding RNA Xist. *Proc Natl Acad Sci U S A.* 2001;98(16):9215–20.
278. Wutz A, Rasmussen TP, Jaenisch R. Chromosomal silencing and localization are mediated by different domains of Xist RNA. *Nat Genet.* 2002;30(2):167–74.
279. Casanova EL, Konkel MK. The developmental gene hypothesis for punctuated equilibrium: combined roles of developmental regulatory genes and transposable elements. *Bioessays.* 2020;42(2):1900173.
280. Toll-Riera M, Bosch N, Bellora N, Castelo R, Armengol L, Estivill X, et al. Origin of primate orphan genes: A comparative genomics approach. *Mol Biol Evol.* 2009;26(3):603–12.
281. Sniezewski L, Janik S, Laszkiewicz A, Majkowski M, Kisielow P, Cebrat M. The evolutionary conservation of the bidirectional activity of the NWC gene

- promoter in jawed vertebrates and the domestication of the RAG transposon. *Dev Comp Immunol.* 2018;81:105–15.
282. Kalitsis P, Saffery R. Inherent promoter bidirectionality facilitates maintenance of sequence integrity and transcription of parasitic DNA in mammalian genomes. *BMC Genomics.* 2009;10:498.
283. Long M, Betrán E, Thornton K, Wang W. The origin of new genes: glimpses from the young and old. *Nat Rev Genet.* 2003;4(11):865–75.
284. Gotea V, Makalowski W. Do transposable elements really contribute to proteomes? *Trends Genet.* 2006;22(5):260–7.
285. Bjerregaard B, Holck S, Christensen IJ, Larsson L-I. Syncytin is involved in breast cancer-endothelial cell fusions. *Cell Mol Life Sci.* 2006;63(16):1906–11.
286. Larsen JM, Christensen IJ, Nielsen HJ, Hansen U, Bjerregaard B, Talts JF, et al. Syncytin immunoreactivity in colorectal cancer: Potential prognostic impact. *Cancer Lett.* 2009;280(1):44–9.
287. Strick R, Ackermann S, Langbein M, Swiatek J, Schubert SW, Hashemolhosseini S, et al. Proliferation and cell–cell fusion of endometrial carcinoma are induced by the human endogenous retroviral Syncytin-1 and regulated by TGF- β . *J Mol Med.* 2006;85(1):23–38.
288. Wang O, Zheng Z, Wang Q, Jin Y, Jin W, Wang Y, et al. ZCCHC12, a novel oncogene in papillary thyroid cancer. *J Cancer Res Clin Oncol.* 2017;143(9):1679–86.
289. Pang SW, Lahiri C, Poh CL, Tan KO. PNMA family: Protein interaction network and cell signalling pathways implicated in cancer and apoptosis. *Cell Signal.* 2018;45:54–62.
290. Papaemmanuil E, Rapado I, Li Y, Potter NE, Wedge DC, Tubio J, et al. RAG-mediated recombination is the predominant driver of oncogenic rearrangement in ETV6-RUNX1 acute lymphoblastic leukemia. *Nat Genet.* 2014;46(2):116–25.
291. Di-Poi N, Montoya-Burgos JI, Duboule D. Atypical relaxation of structural constraints in Hox gene clusters of the green Anole lizard. *Genome Res.* 2009;19(4):602–10.
292. Di-Poi N, Montoya-Burgos JI, Miller H, Pourquié O, Milinkovitch MC, Duboule D. Changes in Hox genes' structure and function during the evolution of the squamate body plan. *Nature.* 2010;464(7285):99–103.
293. Boissinot S, Bourgeois Y, Manthey JD, Ruggiero RP. The mobilome of reptiles: evolution, structure, and function. *Cytogenet Genome Res.* 2019;157(1–2):21–33.
294. Siomi MC, Sato K, Pezic D, Aravin AA. PIWI-interacting small RNAs: the vanguard of genome defence. *Nat Rev Mol Cell Biol.* 2011;12(4):246–58.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more [biomedcentral.com/submissions](https://www.biomedcentral.com/submissions)



1.4 Le poisson-zèbre comme modèle d'étude fonctionnelle des gènes de vertébrés

1.4.1 Le poisson-zèbre...

Le poisson-zèbre, au nom latin *Danio rerio*, est un poisson téléostéen de la famille des cyprinidés. Son nom commun trouve son origine dans sa pigmentation, constituée de rayures horizontales bleues. Il s'agit d'un poisson tropical, originaire du sud-ouest de l'Asie. Plus particulièrement, son habitat naturel se trouve en Inde, Pakistan, Népal, Bangladesh et Birmanie (Parichy, 2015). Ce poisson d'eau douce vie dans des eaux peu profondes telles que les rizières, les eaux stagnantes et les rivières à faible courant. On le retrouve dans des eaux dont les températures varient entre 18°C et 38°C même si sa température optimale de croissance se situe entre 24°C et 29°C. Sa taille se situe généralement entre 2,5 et 4 cm. Il est transparent à l'état embryonnaire et larvaire, jusqu'à l'apparition de sa pigmentation et de ses rayures à l'état adulte. Il présente également un dimorphisme sexuel puisque les mâles sont fins avec de légères teintes roses ou jaunes alors que les femelles sont moins roses et plus trapues avec un plus gros abdomen, du fait des œufs qu'elles transportent (**Figure 1.11**). Ainsi, le poisson-zèbre est un poisson ovipare. Ses œufs sont libérés dans le milieu extérieur pour être fécondés selon le cycle jour/nuit, puisque la lumière du jour active la ponte.

Il s'agit également d'un animal hautement social, car vivant en bancs. La formation et la dynamique de ces bancs est complexe et peut dépendre de nombreux facteurs, tels que la localisation



FIGURE 1.11 – Poissons-zèbres adultes femelle (en haut) et mâle (en bas). © Tohru Murakami

géographique, la végétation, la température, l'âge des poissons ou encore la prédation et l'accès à la nourriture. La nage en banc leur confère des avantages adaptatifs, comme la recherche de nourriture et une protection contre les prédateurs. Malgré ce comportement social, les poissons-zèbres mâles et femelles présentent aussi des comportements agressifs caractérisés par des attaques, des morsures, des chasses et l'établissement de dominance.

C'est dans les années 1960, grâce à George Streisinger, passionné de poissons tropicaux, et ses collègues de l'université d'Oregon qui cherchaient un modèle d'étude vertébré moins complexe que la souris, que le poisson-zèbre a fait son apparition dans la recherche (Streisinger et al., 1981). Mais c'est seulement plus récemment, à partir des années 1990, grâce aux travaux de criblages génétiques dirigés par Christiane Nüsslein-Volhard et Wolfgang Driever, que le poisson-zèbre est devenu plus populaire comme modèle de laboratoire.

1.4.2 ... est un organisme modèle de choix.

Qu'est-ce qu'un organisme modèle? Il s'agit d'un organisme généralement facile à maintenir et à élever dans des conditions de laboratoire, avec des avantages et facilités expérimentales, faisant de lui une espèce largement étudiée et répandue dans les laboratoires. Ils permettent d'étudier des mécanismes et phénomènes biologiques complexes dans des représentations simplifiées. Ils sont choisis selon des critères scientifiques (liés aux particularités de l'espèce, maîtrise de divers facteurs génétiques, sanitaires etc.), techniques (équipements, réactifs etc.), éthiques/réglementaires (restriction d'usage de certaines espèces comme les primates, utilisation de l'espèce la moins sensible etc.) et pragmatiques (disponibilité, temps de génération, coûts, taille etc.). Le poisson-zèbre présente de nombreux avantages selon ces critères (Nüsslein-Volhard et al., 2002). En effet, d'un point de vue pragmatique, c'est un poisson de petite taille, robuste, dont les conditions de vie naturelles sont facilement reproductibles en laboratoire. L'élevage du poisson-zèbre en laboratoire est moins coûteux que la plupart des modèles de vertébrés, tels que la souris (Brand et al., 2002).

Du point de vue de sa reproduction, la maturité sexuelle des individus est atteinte au bout de 3-4 mois, ce qui n'est pas particulièrement court comparé aux autres vertébrés (chez la souris il faut 6-8 semaines) (Drickamer, 1981; Singleman & Holtzman, 2014). Cependant, son mode de reproduction présente d'autres nombreux avantages. La production d'œufs est régulière et très conséquente puisqu'une seule femelle peut pondre entre 200 et 300 œufs par ponte. Les embryons ont un développement extrêmement rapide, l'organogenèse se fait en 24 heures, alors qu'il faut en comparaison 11 jours chez la souris (Dahm, 2002; Kaufman, 1992; Kimmel et al., 1995). De plus, la fertilisation puis le développement des œufs se déroulent en dehors du corps de la femelle, ce qui en fait un modèle idéal pour l'étude du développement précoce, ainsi que la mise en place de techniques d'édition du génome. Ces œufs sont également transparents, ce qui facilite leur utilisation pour l'observation des événements développementaux.

D'un point de vue génétique, le poisson-zèbre montre également de nombreux atouts. En effet, le projet de séquençage du génome, lancé en 2001, a permis de séquencer les plus de 1,4 milliards de paires de bases de son génome (Howe et al., 2013). Le poisson-zèbre possède plus de 26.000 gènes codant pour des protéines. De plus, nous partageons plus de 70 % de nos gènes avec le poisson-zèbre, et même 82 % des gènes associés à des maladies humaines sont retrouvés

chez ce poisson. Ainsi, il permet d'étudier fonctionnellement des gènes communs aux vertébrés, notamment ceux impliqués dans les maladies humaines; mais il permet également d'étudier ces gènes pour comprendre l'évolution des vertébrés.

Le développement et l'optimisation de techniques de génétique moléculaire utilisables chez le poisson-zèbre en font une espèce d'autant plus précieuse. Par exemple, la technique des Morpholinos (ARN antisens synthétique utilisé pour inhiber transitoirement la transcription d'un ARN messager cible) (Nasevicius & Ekker, 2000), ou plus récemment celle d'édition du génome par CRISPR-Cas9 (Chang et al., 2013; Hwang et al., 2013a,b) sont particulièrement adaptées pour l'étude de la fonction des gènes chez le poisson-zèbre.

Tous ces avantages font du poisson-zèbre un modèle de plus en plus répandu, et même l'un des principaux, dans la recherche scientifique de ces dernières années (**Figure 1.12**).

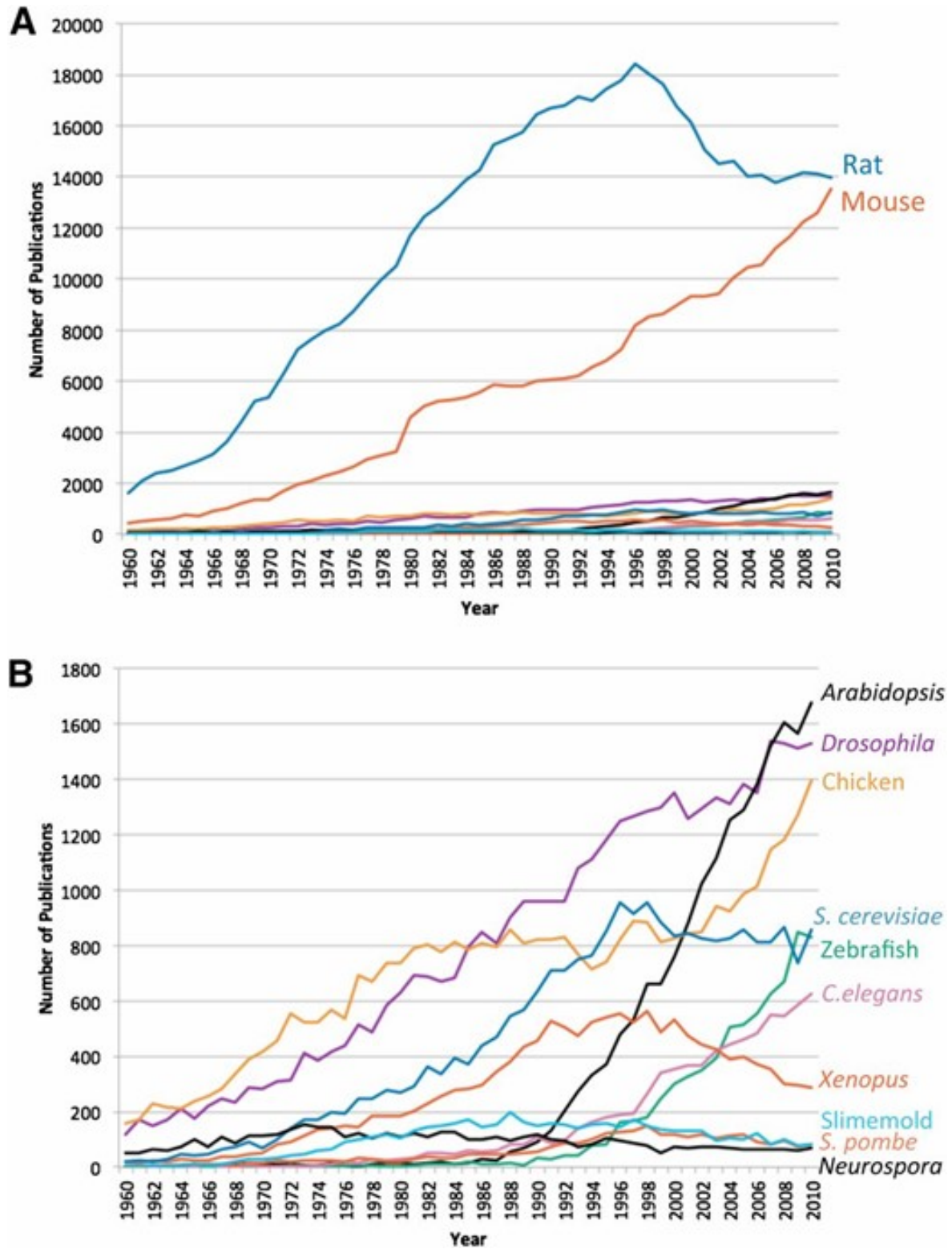


FIGURE 1.12 – Nombre de publications par un utilisant différents organismes modèles. De 1960 à 2010 (A); (B) sans les organismes modèles rat et souris. D'après Dietrich et al. (2014)

1.4.3 Développement du poisson-zèbre

1.4.3.1 Développement embryonnaire du poisson-zèbre

Le développement embryonnaire du poisson-zèbre commence après la fécondation (**Figure 1.13**) (Dahm, 2002; Kimmel et al., 1995). À ce stade, l'embryon est sous forme d'une cellule, au-dessus du sac vitellin. Environ trente minutes après la fertilisation, la période de clivage commence. Celle-ci consiste en des divisions cellulaires synchrones toutes les quinze minutes, pour produire un blastomère à 64 cellules. Ensuite, lors de la période de blastula, les divisions cellulaires se poursuivent pour passer de 128 à 1000 cellules. La blastula change alors de forme en s'allongeant pour former une sphère. Au stade gastrula, les mouvements cellulaires d'involution et d'extension vont permettre d'aboutir au stade « bud », où les axes antéro-postérieur et dorso-ventral sont déterminés. On commence également à voir à ce stade les prémices de la queue, du tube neural et du cerveau. À partir de 10 heures après la fertilisation, la période de segmentation commence, menant aux premiers mouvements de l'embryon. La queue se détache petit à petit du vitellus. 24 heures après la fertilisation, on entre dans la période pharyngienne. Les organes majeurs du poisson-zèbre sont établis et visibles. Apparaît alors la pigmentation au niveau de l'œil et de la peau. La circulation sanguine se met également en place avec les premiers battements de cœur, ainsi que des contractions musculaires spontanées. Enfin, pendant la période d'éclosion, l'embryon continue de grandir et fini par éclore autour de 72 heures après la fertilisation.

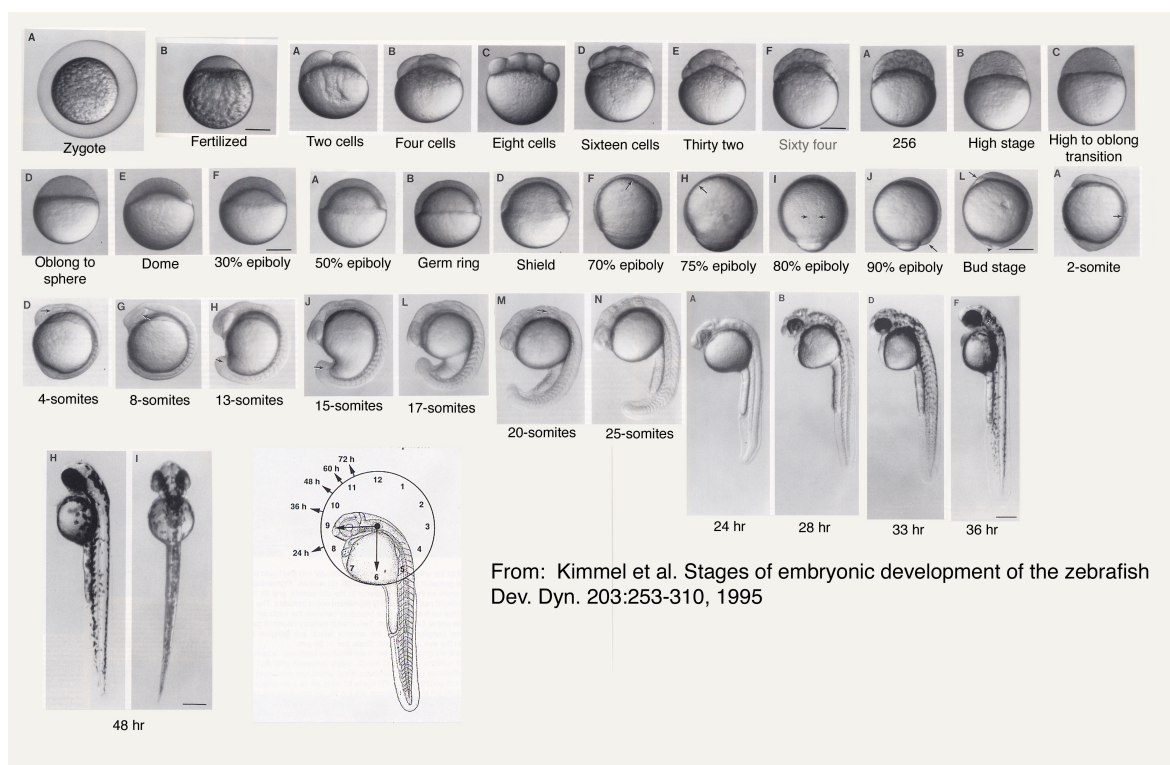


FIGURE 1.13 – Développement embryonnaire du poisson-zèbre. D'après Kimmel et al. (1995)

1.4.3.2 Développement larvaire du poisson-zèbre

120 heures après la fertilisation, les poissons-zèbres sont sous forme de larves précoces (**Figure 1.14**) (Dahm, 2002; Schilling, 2002). Ils continuent ensuite de grandir avec notamment la croissance de la vessie natatoire et de la bouche. La pigmentation se développe avec l'apparition de différents types de cellules pigmentaires. La larve devient plus active puisque l'on observe des mouvements des mâchoires, des yeux ainsi que des nageoires permettant la nage. Ceci rend possible la réaction de fuite des larves face à des stimuli et permettra ensuite la mise en place complète de la respiration, la recherche de nourriture et l'alimentation.

1.4.3.3 Le poisson-zèbre à l'état adulte

Après l'état larvaire, le poisson-zèbre atteint l'état adulte (**Figure 1.15**) (Schilling, 2002). Seul l'appareil reproducteur final reste à établir. En effet, les poissons-zèbre se développent d'abord comme femelles. Les gonades femelles peuvent ensuite devenir gonades mâles à partir d'environ 7 semaines après la fécondation. Au-delà des facteurs génétiques, tel que le gène *DMRT1*, de nombreux facteurs environnementaux influent sur le ratio mâle/femelle d'un aquarium tel que la densité de poisson, la nourriture ou la température de l'eau (Kossack & Draper, 2019).

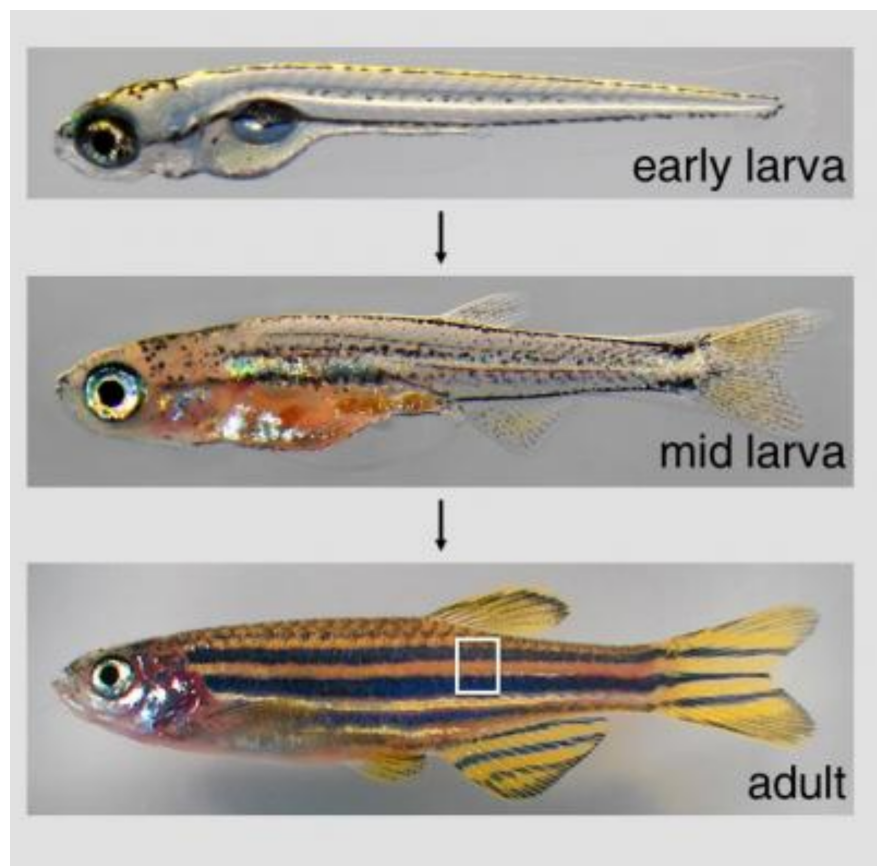


FIGURE 1.14 – Développement du poisson-zèbre des stades larvaires précoce et moyen au stade adulte.
© Dae Seok Eom, David Parichy

Zebrafish Developmental Timeline

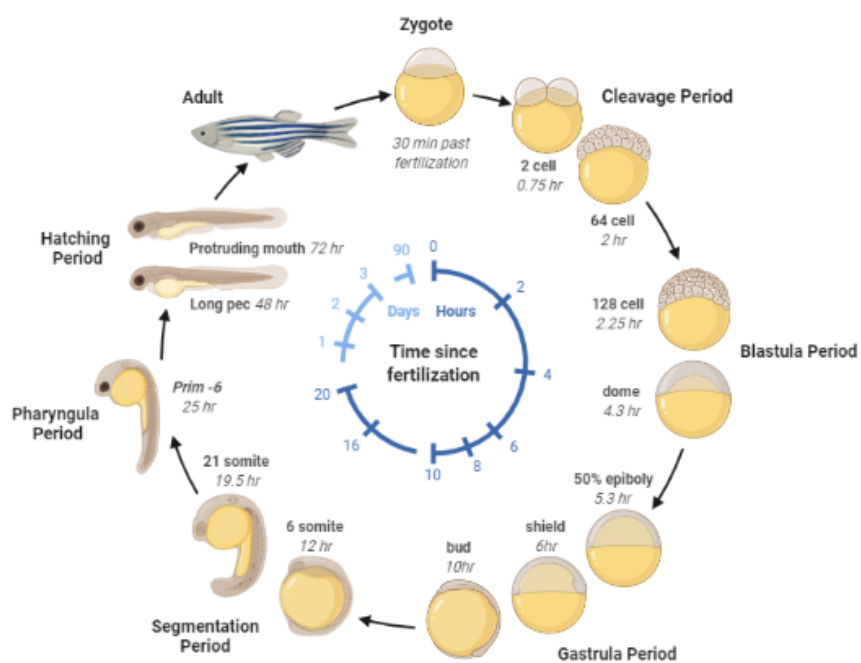


FIGURE 1.15 – Cycle de développement du poisson-zèbre. © Stephanie Lepage, Biorender

1.5 Objectifs de la thèse

Comme présenté précédemment, les vertébrés ont acquis des innovations développementales majeures au cours de leur évolution, faisant d'eux un clade au large succès (**section 1.1**). L'apparition de nouveaux gènes est un facteur contribuant à l'apparition d'innovations. Lors de l'évolution précoce des vertébrés, deux événements de duplication de génome entier ont permis une explosion du nombre de gènes, et sont donc supposés être des événements majeurs à l'origine de multiples innovations de vertébrés (Ohno, 1999). Cependant, l'origine précise de nombreuses innovations n'est pas caractérisée génétiquement. Ainsi, la formation de nouveaux gènes par d'autres mécanismes ne doit pas être sous-estimée. En particulier, les éléments transposables sont à l'origine d'innovations développementales majeures des vertébrés (**section 1.3.2**).

Nous nous sommes donc demandé si des fonctions spécifiques des vertébrés pourraient être dues à des gènes dérivés d'éléments transposables. En d'autres termes, l'objectif de ma thèse fut d'étudier l'impact de la domestication moléculaire des éléments transposables dans l'évolution précoce des vertébrés. En effet, l'identification et la caractérisation d'événements de domestications moléculaires d'ETs sont encore probablement sous évaluées, principalement dû à la difficulté d'annotation de ces éléments. Cependant, le développement ces dernières années de nouvelles technologies de séquençage, d'annotation et d'analyse des génomes, permet aujourd'hui d'approfondir ces points.

Dans ce contexte, j'ai donc cherché à identifier un nouveau cas de gène dérivé d'élément transposable, dont l'origine serait retrouvée au cours de l'évolution précoce des vertébrés. C'est ainsi que j'ai identifié le gène *MSANTD2* dérivé d'un transposon *Harbinger* et présent chez les vertébrés à mâchoires. Cependant, j'ai rapidement identifié trois autres gènes dérivés de transposons *Harbinger* et également présents chez les vertébrés à mâchoires. Mon projet de thèse a donc consisté en l'étude des transposons *Harbinger* et de leurs gènes dérivés.

Pour ce faire, j'ai étudié la dynamique évolutive des transposons *Harbinger*, en particulier dans les génomes de poissons téléostéens. En effet, les poissons téléostéens représentent le groupe de vertébrés le plus diversifié au niveau des éléments transposables, tant du point de vue quantitatif que qualitatif (**section 1.3.1**). Ceci fait l'objet du deuxième chapitre de cette thèse (**chapitre 2**).

Dans un second temps, je me suis intéressée aux gènes dérivés de transposons *Harbinger*. Plus particulièrement, j'ai étudié une famille de gènes issus de domestications moléculaires récurrentes de transposons *Harbinger* chez les vertébrés. Dans le troisième chapitre de cette thèse, je présente donc l'analyse évolutive de cette famille de gènes chez les vertébrés, ainsi que l'analyse fonctionnelle de ces gènes à l'aide de l'organisme modèle poisson-zèbre (**chapitre 3**).

2

Les transposons *Harbinger* chez les poissons téléostéens

Sommaire

2.1	Avant-propos	56
2.2	Article : Diversity of <i>Harbinger</i>-like transposons in teleost fish genomes	57
2.2.1	Abstract	57
2.2.2	Introduction	58
2.2.3	Material and methods	60
2.2.4	Results	62
2.2.4.1	Differential contribution of <i>Harbinger</i> -like transposons to fish genomes	62
2.2.4.2	Distribution of <i>Harbinger</i> -like transposons in medaka genome	62
2.2.4.3	Evolution of <i>Harbinger</i> transposons in teleost fish genomes	64
2.2.4.4	The evolution of the Myb-like proteins recapitulates the evolution of transposase proteins of <i>Harbinger</i> transposons	64
2.2.4.5	<i>Harbinger</i> -like transposons are expressed in fish	69
2.2.5	Discussion	71
2.2.5.1	<i>Harbinger</i> -like transposons are inequitably widespread in fish genomes	71
2.2.5.2	Evolutionary relationships between the two ORFs of <i>Harbinger</i> transposons	71
2.2.5.3	<i>Harbinger</i> -like transposons are transcriptionally active in teleost fish	72
2.2.6	Conclusion	73
2.2.7	Acknowledgements	73
2.2.8	Supplementary figure	74

2.1 Avant-propos

Dans ce chapitre, je me suis intéressée à la dynamique évolutive des transposons *Harbinger* chez les poissons téléostéens. En effet, ce clade représente un modèle particulièrement intéressant pour l'étude des éléments transposables puisqu'ils y sont remarquablement abondants et diversifiés (**section 1.3.1**). De plus, les éléments *Harbinger* n'ont pas été étudiés en détail dans ces génomes à ce jour. Suite à l'invitation du journal *Animals*, cette étude fait l'objet d'un article qui y sera soumis.

2.2 Article : Diversity of *Harbinger*-like transposons in teleost fish genomes

Ema Etchegaray, Corentin Dechaud, Jeremy Barbier, Magali Naville, Jean-Nicolas Volff*

Institut de Génomique Fonctionnelle de Lyon, Ecole Normale Supérieure de Lyon, CNRS UMR 5242, Université Claude Bernard Lyon 1, Lyon, France

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

2.2.1 Abstract

Harbinger elements are DNA transposons that are widespread from plants to vertebrates but absent from mammalian genomes. Among vertebrates, teleost fish are the clade presenting not only the largest number of species but also the highest diversity of transposable elements, both quantitatively and qualitatively, making them a very attractive group to investigate the evolution and the genomic impact of mobile sequences. Here we studied *Harbinger* DNA transposons and the distantly related *ISL2EU* elements in fish, focusing on eight selected teleost species compared to the spotted gar (non-teleost ray-finned fish), the coelacanth (sarcopterygian fish), the elephant shark (cartilaginous fish) and the amphioxus (cephalochordate). We observed a high variability in the genomic composition of *Harbinger*-like sequences in teleost fish. When present, *Harbinger* and *ISL2EU* transposons covered 0.002 %-0.14 % and 0.005 %-0.10 % of the genome, respectively. *Harbinger* transposons were particularly represented in the genome of the medaka fish *Oryzias latipes*, whereas *ISL2EU*, which were scarcer in teleost fish (they were absent from three out of the eight species studied), were in highest amount in the platyfish *Xiphophorus maculatus*. While *Harbinger* transposons might have been present in a common ancestor of all fish species studied here, with secondary loss in elephant shark, our results suggested that *ISL2EU* elements have been gained by horizontal transfer at the base of teleost fish 200-300 million years ago, with secondary loss in a common ancestor of pufferfishes and stickleback. Both *Harbinger* and *ISL2EU* transposons were composed of two independent open reading frames coding for a transposase and a Myb-like protein, respectively. We reconstructed and compared molecular phylogenies of both proteins to get insights into the evolution of *Harbinger*-like transposons in fish. Transposase and Myb-like protein phylogenies showed global congruent evolution, indicating unique origin of the association between both genes and suggesting rarity of recombination between transposon sublineages. Finally, we report transcriptional activity of both *Harbinger* and *ISL2EU* transposons in teleost fish, with a male-biased expression in the gonads of the medaka fish.

Keywords

Transposable elements, *Harbinger*, *ISL2EU*, genomes, teleost fish, evolution

2.2.2 Introduction

Transposable elements (TEs) are repeated DNA sequences that can be inserted into new locations in genomes. They are classified into two main classes, class I retrotransposons and class II DNA transposons, depending on their transposition mechanism. Class I TEs use an RNA intermediate that is reverse-transcribed into a new cDNA copy of the element (copy-and-paste mechanism), whereas class II transposons are generally excised from the original locus by a transposase and integrated into another site (cut-and-paste mechanism). Within each class, TEs are subdivided into superfamilies and families according to their phylogenetic relationships (Wicker et al., 2007). Since the discovery of TEs in the 1950s, TEs have been shown to be major components of genomes, and there is growing evidence of their important roles in genome evolution and organism adaptation (Chuong et al., 2016, 2017; Jangam et al., 2017; Kidwell & Lisch, 2000; Volff, 2005).

With the flourishing development of sequencing technologies and genome annotation tools, the number of sequenced genomes has exploded. Studying new species beside model organisms allows a broader and wider understanding of the molecular basis and evolutionary dynamics of biodiversity. In particular, fish, which represent more than 48 % of vertebrate species (36,000 known teleost fish species), are still understudied (Fricke et al., 2022; Van Der Laan et al., 2014). Notably, the group of species called fish does not represent a monophyletic group, since it is composed of both chondrichthyes (cartilaginous fish) and osteichthyes, the latter comprising both bony fish and non-fish sarcopterygian species. Bony fish are subdivided into actinopterygians (ray-finned fish), including the non-teleost spotted gar and teleost fishes such as zebrafish, cod, stickleback, tetraodon, fugu, platyfish, medaka and tilapia, and into sarcopterygians (lobe-finned fish) such as coelacanth and lungfish, which are the closest fish relatives of terrestrial vertebrates (tetrapods).

Teleost fish constitute the great majority of fish species. They present a high level of biodiversity and are considered to have plastic genomes (Volff, 2005). They particularly constitute an attractive group to study mobile sequences, since their genomes present a high level of TE diversity, both qualitatively and quantitatively (Carducci et al., 2020; Chalopin et al., 2015; Volff, 2005). Indeed, teleost fish genomes contain a larger number of TE superfamilies compared to tetrapods, particularly to birds and mammals, which have lost many groups of transposable elements during their evolution (Volff, 2005). Teleost genomes are also variable in term of TE coverage, which ranges from less than 10 % for tetraodon and fugu, which are species with compact genomes, to more than 50 % for zebrafish (Chalopin et al., 2015). Finally, fish genomes present a higher proportion of class II DNA transposons compared to mammals and birds, which mainly possess class I TEs. In this study, we investigated in teleost fish genomes a superfamily of DNA transposons called *Harbinger* or *PIF/Harbinger*. *Harbinger* transposons are found in various species including fish, other animals and plants, but are absent from mammalian genomes (Casola et al., 2007; Grzebelus et al., 2007; Han et al., 2015; Jiang et al., 2003; Jurka & Kapitonov, 2001; Kapitonov & Jurka, 1999, 2004; Kikuchi et al., 2003; Markova & Mason-Gamer, 2015; Pereira et al., 2013; Yuan & Wessler, 2011; Zhang et al., 2001, 2004). They are usually flanked by terminal inverted repeats (TIRS) of 25-50 base pairs (bp) in length and generally generate 3bp-long target site duplications through their integration into a genomic site. Typical autonomous *Harbinger* elements carry two open reading frames (ORFs) (Kapitonov & Jurka, 2004). The first ORF codes for a transposase containing a DDE endonuclease motif, which is composed of three carboxylate residues coordinating metal ions necessary for both catalysis of

DNA cleavage at the site of insertion and strand transfer. The second ORF encodes a DNA-binding protein possessing a conserved Myb/SANT-like domain (we will refer to this ORF as *Myb-like*). This domain is composed of a tri-helix motif – with conserved bulky aromatic residues essential for the stability of the motif – allowing interactions with DNA and proteins. Indeed, the Myb-like protein has been shown to interact with the transposase, hereby allowing their concomitant nuclear import, and to bind DNA at the *Harbinger* TIR sequences, leading to the excision of the *Harbinger* sequence by the transposase (Sinzelle et al., 2008). Transposases present a high degree of conservation even between different families of *Harbinger* transposons, whereas the Myb-like proteins are much more divergent, with only some similarities in restricted parts of the Myb-like domain between different *Harbinger* families (Kapitonov & Jurka, 2004).

ISL2EU elements are distantly related to *Harbinger* transposons but belong to the same super-family called *Harbinger-like* (Han et al., 2015; Yuan & Wessler, 2011). They are found in animals and display two ORFs, one encoding a DDE transposase with Helix-Turn-Helix (HTH) or THAP putative DNA-binding domains, and the other one coding for an exonuclease containing an YqaJ alkaline exonuclease domain. Therefore, even if *Harbinger* and *ISL2EU* elements both present two ORFs, one of them encoding a DDE transposase, the second ORF is different in the two types of elements. Finally, other *Harbinger-like* elements called *Spy* have been found only in invertebrates (Han et al., 2014, 2015). Most *Spy* elements present only one ORF encoding a transposase with DDE and HTH motifs.

Both genes of *Harbinger* transposons are necessary for transposition *in vitro* (Hancock et al., 2010; Sinzelle et al., 2008). *Harbinger* elements are transcriptionally active in *Triticeae* plants and in the salamander *Pleurodeles waltl* (Elewa et al., 2017; Markova & Mason-Gamer, 2017). In another study the expression of a coelacanth *Harbinger* transposon was detected in a mouse PAC transgenic cell line (P1-derived artificial chromosome with a coelacanth genomic insert containing a *Harbinger* element), suggesting its expression in the coelacanth itself (Smith et al., 2012).

Here we present the analysis of the evolutionary dynamics of *Harbinger* and *ISL2EU* transposons, collectively called *Harbinger-like* elements, in teleost fish genomes, focusing on zebrafish (*Danio rerio*), cod (*Gadus morhua*), stickleback (*Gasterosteus aculeatus*), tetraodon (*Tetraodon nigroviridis*), fugu (*Takifugu rubripes*), platyfish (*Xiphophorus maculatus*), medaka (*Oryzias latipes*) and tilapia (*Oreochromis niloticus*), for which genome assemblies of good quality are available. Genomes of spotted gar (*Lepisosteus oculatus*, non-teleost ray-finned fish), coelacanth (*Latimeria chalumnae*, sarcopterygian), elephant shark (*Callorhynchus milii*, cartilaginous fish) and amphioxus (*Branchiostoma floridae*, cephalochordate) were also included in this study as external groups for comparison. We observed a differential distribution of *Harbinger-like* elements depending on the fish genomes, with a higher abundance in medaka for *Harbinger* transposons and in platyfish for *ISL2EU* transposons. Moreover, we performed a comparative evolutionary analysis of the two ORFs of *Harbinger* transposons. We observed the persistence of the two ORFs in all fish transposons studied, with a global congruent evolution between the transposase and the Myb-like proteins. Finally, we also show evidence of transcriptional activity of *Harbinger* and *ISL2EU* transposons in fish, with a testis-biased expression in the gonads of the medaka.

2.2.3 Material and methods

Genomes

We used the following genome sequences in our analysis : amphioxus (*Branchiostoma floridae*_v2.0.assembly.fasta, <http://genome.jgi-psf.org/Brafl1/Brafl1.download.ftp.html>, last accessed January 30, 2015), elephant shark (EsharkAssembly, <http://esharkgenome.imcb.a-star.edu.sg>, last accessed January 30, 2015), fugu (*Takifugu rubripes*.FUGU4.66.dna.toplevel.fa, Ensembl), tetraodon (*Tetraodon nigroviridis*.TETRAODON8.73.dna.toplevel.fa, Ensembl), stickleback (*Gasterosteus aculeatus*.BROADS1.68.dna.toplevel.fa, Ensembl), tilapia (*Oreochromis niloticus*.Orenil1.0.68.dna.toplevel.fa, Ensembl), platyfish (*Xiphophorus maculatus*.Xipmac4.4.2.69.dna.nonchromosomal.fa, Ensembl), medaka (https://www.ncbi.nlm.nih.gov/assembly/GCF_002234675.1, last accessed September 2021), Atlantic cod (*Gadus morhua*.gadMor1.73.dna.toplevel.fa, Ensembl), zebrafish (*Danio rerio*.Zv9.66.dna.toplevel.fa, Ensembl), spotted gar assembly accession update (http://www.ncbi.nlm.nih.gov/assembly/GCF_000242695.1/, last accessed January 30, 2015, Genbank Assembly), African coelacanth (*Latimeria chalumnae*.LatCha1.72.dna.toplevel.fa, Ensembl).

TE annotation

TE libraries were established by a combination of both automatic and manual annotations. Manual annotations corresponded to TBLASTN search against genomes (downloaded or on the NCBI website <https://www.ncbi.nlm.nih.gov/genome/>) using TE proteins from different superfamilies as queries. TE sequences were also retrieved from the Repbase database (<http://www.girinst.org>) (Kapitonov & Jurka, 2008). Automatic annotation was performed using the RepeatModeler software (Smit, AFA, Hubley, R., <http://www.repeatmasker.org>) with default parameters. For the coelacanth, we used and reannotated the library from Amemiya et al. (2013).

TE genome masking, copy number and genome coverage

TE genome masking, copy number and genome coverage estimations were performed according to Chalopin et al. (2015). Briefly, RepeatMasker version 3.3.0 (Smit, AFA, Hubley, R, and Green, P. RepeatMasker Open-3.0. 1996–2010; <http://www.repeatmasker.org>) with “-a” and “-lib” default parameters was locally used to mask genomes. Copy number and genome coverage were calculated on RepeatMasker outfiles (.out) with custom scripts. In order to eliminate very short and too divergent sequences, data were filtered to include only elements longer than 80 nucleotides and sharing more than 80 % of identity with the reference sequence from the species-specific library. *Harbinger* and *ISL2EU* elements were annotated with RepeatModeler (<http://www.repeatmasker.org/RepeatModeler/>) and RepeatMasker (<http://repeatmasker.org>).

Sequence alignment and phylogenetic analysis

Nucleotide and amino-acid sequences were aligned using MAFFT (Katoh et al., 2002). Phylogenetic trees were built using maximum likelihood calculation with the PhyML software (Guindon & Gascuel, 2003) using LG model, and the MrBayes package (Huelsenbeck & Ronquist, 2001) using

mixed model (estimated by protest-3 software, Darriba et al. (2011)) and 500,000 generations of Bayesian inferences.

TE expression analysis in spotted gar, zebrafish, cod and medaka

TE expression was studied by comparing TE sequences against the PhyloFish database (Pasquier et al., 2016). Results were analyzed using the RNAbrowse interface (<http://phylofish.sigenae.org/index.html>) (Mariette et al., 2014). Expression was given at the copy level in PhyloFish and summarized for all superfamilies in **Figure 2.7**.

TE expression analysis in medaka gonads

TE expression quantification was performed on RNAseq data from both male and female gonads of the medaka fish *O. latipes* (Dechaud et al., 2021). Shortly, SquIRE (Yang et al., 2019) was used to estimate TE expression at the copy level using our TE library on the *O. latipes* genome (https://www.ncbi.nlm.nih.gov/assembly/GCF_002234675.1). We ran SquIRE “clean,” “map,” “count,” and “call” steps to estimate TE expression. Then, mean expression was calculated for all copies of a same TE family.

TE distribution on medaka chromosomes

Distribution of *Harbinger* and *ISL2EU* elements along medaka chromosomes was represented using the R karyoploteR package (available at the Bioconductor site <https://bioconductor.org/packages/release/bioc/html/karyoploteR.html>).

2.2.4 Results

2.2.4.1 Differential contribution of *Harbinger*-like transposons to fish genomes

In order to investigate the genomic contribution, evolution and diversification of *Harbinger* and *ISL2EU* transposons in fish genomes, we analyzed eight teleost fish species, as well as one non-teleost ray-finned fish (spotted gar), one sarcopterygian fish (coelacanth), one cartilaginous fish (elephant shark) and one cephalochordate (amphioxus), that were used as outgroups for comparison (**Figure 2.1**).

Harbinger transposon content was variable in teleost fish genomes. Tetraodon, fugu and cod genomes, which contain the lowest global amount of TEs among the species studied (5.9 %, 6.7 % and 14.3 %, respectively; Chalopin et al. (2015)), also presented the lowest genomic contribution of *Harbinger* transposons. The zebrafish genome, which otherwise possesses the highest global TE content in teleost fish (54.9 %; Chalopin et al. (2015)), was not particularly enriched in *Harbinger* transposons, indicating that these elements did not significantly contribute to TE expansion in this fish. In contrast, *Harbinger* transposons were particularly well represented in medaka in terms of both genome coverage and copy number. Outside ray-finned fish, these transposons were present with high copy numbers in both coelacanth and amphioxus, this being however not correlated with high coverage in coelacanth probably due to large genome size (more than 2,800 Mb; Amemiya et al. (2013)). Finally, absence of *Harbinger* in elephant shark might suggest secondary loss of these elements during the evolution of cartilaginous fish (Han et al., 2015; Kapitonov & Jurka, 2004).

ISL2EU elements were detected in teleost fish but absent from the non-teleost species including spotted gar, coelacanth and elephant shark. Concerning amphioxus, Han et al. (2015) indicated the presence of *ISL2EU* in this genome, however we were not able to detect them (Han et al., 2015). Since such elements are also absent from tetrapods but present in more divergent animals (Han et al., 2015), they might have been gained through horizontal transfer at the base of the teleost lineage 200-300 million years ago. Within teleost fish, *ISL2EU* distribution was patchy compared to *Harbinger*, with absence in three out of the eight species studied (fugu, tetraodon and stickleback). This suggested secondary loss of the elements in a common ancestor of these three species about 100 million years ago. *ISL2EU* elements were particularly present in platyfish and medaka genomes. Overall, there was no clear correlation between *Harbinger* and *ISL2EU* transposon in fish genomes.

2.2.4.2 Distribution of *Harbinger*-like transposons in medaka genome

We looked at the distribution of *Harbinger* and *ISL2EU* transposons on medaka chromosomes (**Supplementary Figure 2.9**). We observed that they were homogeneously distributed all along the chromosomes. This suggested that *Harbinger*-like transposons do not have any preferential insertion/retention regions at the genomic scale in the medaka genome.

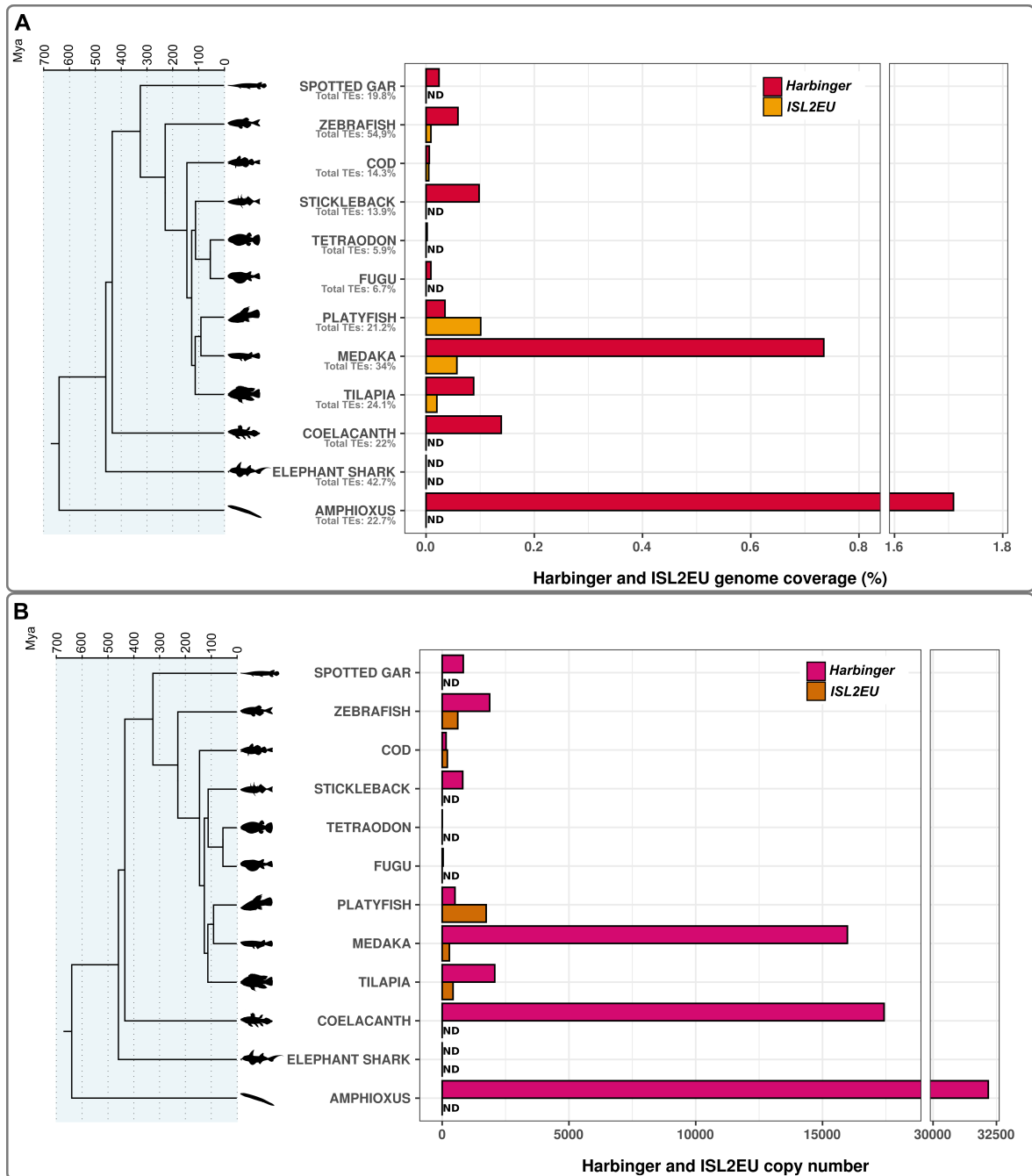


FIGURE 2.1 – Genome coverage (A) and copy number (B) of *Harbinger* and *ISL2EU* transposons in spotted gar, zebrafish, cod, stickleback, tetraodon, fugu, platyfish, medaka, tilapia, coelacanth, elephant shark and amphioxus. Data were filtered to include only copies longer than 80 nucleotides and sharing more than 80 % of identity with the reference sequence from the species-specific library. In (A) the global percentage of TEs in genomes is indicated for each species under the species name. ND (Not Detected) is indicated when no element was detected in the species. Species phylogeny is based on divergence times estimated using the TimeTree public database (Kumar et al., 2017).

2.2.4.3 Evolution of *Harbinger* transposons in teleost fish genomes

To investigate the evolution of *Harbinger* and *ISL2EU* transposons in fish genomes, representative sequences of 31 *Harbinger* and five *ISL2EU* transposons were further analyzed. These are consensus sequences of different families of *Harbinger*-like transposons annotated by RepeatModeler in genomes (Smit, AFA, Hubley, R 2008-2015). Multiple alignments of sequences of transposases (for *Harbinger* and *ISL2EU*) and Myb-like proteins (for *Harbinger*) were constructed based on the DDE motif (about 160 amino-acids, aa) and the Myb-like domain (ca. 100 aa), respectively (**Figure 2.3** and **Figure 2.4**). The results indicated conservation of the protein sequences among teleost fish, particularly for the transposase. As expected from the literature, Myb-like proteins appeared to be less constrained, but their secondary structure was well conserved (Kapitonov & Jurka, 2004).

In order to study *Harbinger*-like transposon evolution, phylogenetic trees were constructed based on transposase DDE domain multiple alignment using the Bayesian (**Figure 2.5A**) and Maximum Likelihood methods (**Supplementary Figure 2.10A**) (Guindon & Gascuel, 2003; Huelsenbeck & Ronquist, 2001). The results confirmed that *ISL2EU* transposons form a phylogenetic group distinct from *Harbinger* elements. We observed that many *Harbinger* sequences were more related to elements from other species than to sequences from the same species, hereby defining different *Harbinger* families. Some teleost *Harbinger* elements grouped with amphioxus sequences, indicating families possibly present in a common chordate ancestor, or alternatively horizontal transfer. Some other teleost *Harbinger* sequences were related neither to coelacanth nor to amphioxus families, suggesting teleost-specific family expansion and divergence, or horizontal transfer from more divergent organisms. Finally, coelacanth *Harbinger* transposon sequences preferentially grouped together and formed a distinct group from other fish transposons, indicating specific family expansion in the coelacanth genome or in one of its sarcopterygian ancestor.

2.2.4.4 The evolution of the Myb-like proteins recapitulates the evolution of transposase proteins of *Harbinger* transposons

The presence of two independent ORFs in *Harbinger* elements is an unusual feature in DNA transposons. To date, the evolutionary history of *Harbinger* transposons was only studied through large-scale analyses of their transposases (Han et al., 2015; Kapitonov & Jurka, 2004; Markova & Mason-Gamer, 2015). Moreover, while transposases present a high degree of conservation even between different families of *Harbinger* transposons, the Myb-like proteins are much more diverse and display only some similarity in restricted parts of the Myb-like domain (**Figure 2.3** and **Figure 2.4**) (Kapitonov & Jurka, 2004). Thus, the evolution of this second ORF has been poorly studied so far (Kapitonov & Jurka, 2004).

The Myb-like proteins associated to the transposases of the *Harbinger* transposons presented in **Figure 2.5A** (and **Supplementary Figure 2.10A**) were aligned on their Myb-like domain (ca. 85 aa)

FIGURE 2.2 – Multiple alignments of the *Harbinger*-like transposase proteins on the DDE domain. Black stars indicate the putative catalytic DDE residues. (DR : *Danio rerio* – zebrafish, GA : *Gasterosteus aculeatus* – stickleback, GM : *Gadus morhua* – cod, LC : *Latimeria chalumnae* – coelacanth, NF : *Nothobranchius furzeri* – killifish, OL : *Oryzias latipes* – medaka, ON : *Oreochromis niloticus* – tilapia, SS : *Salmo salar* – salmon, TF : *Takifugu flavidus* – fugu, TR : *Takifugu rubripes* – fugu, XM : *Xiphophorus maculatus* – platyfish).

2. Les transposons *Harbinger* chez les poissons téléostéens

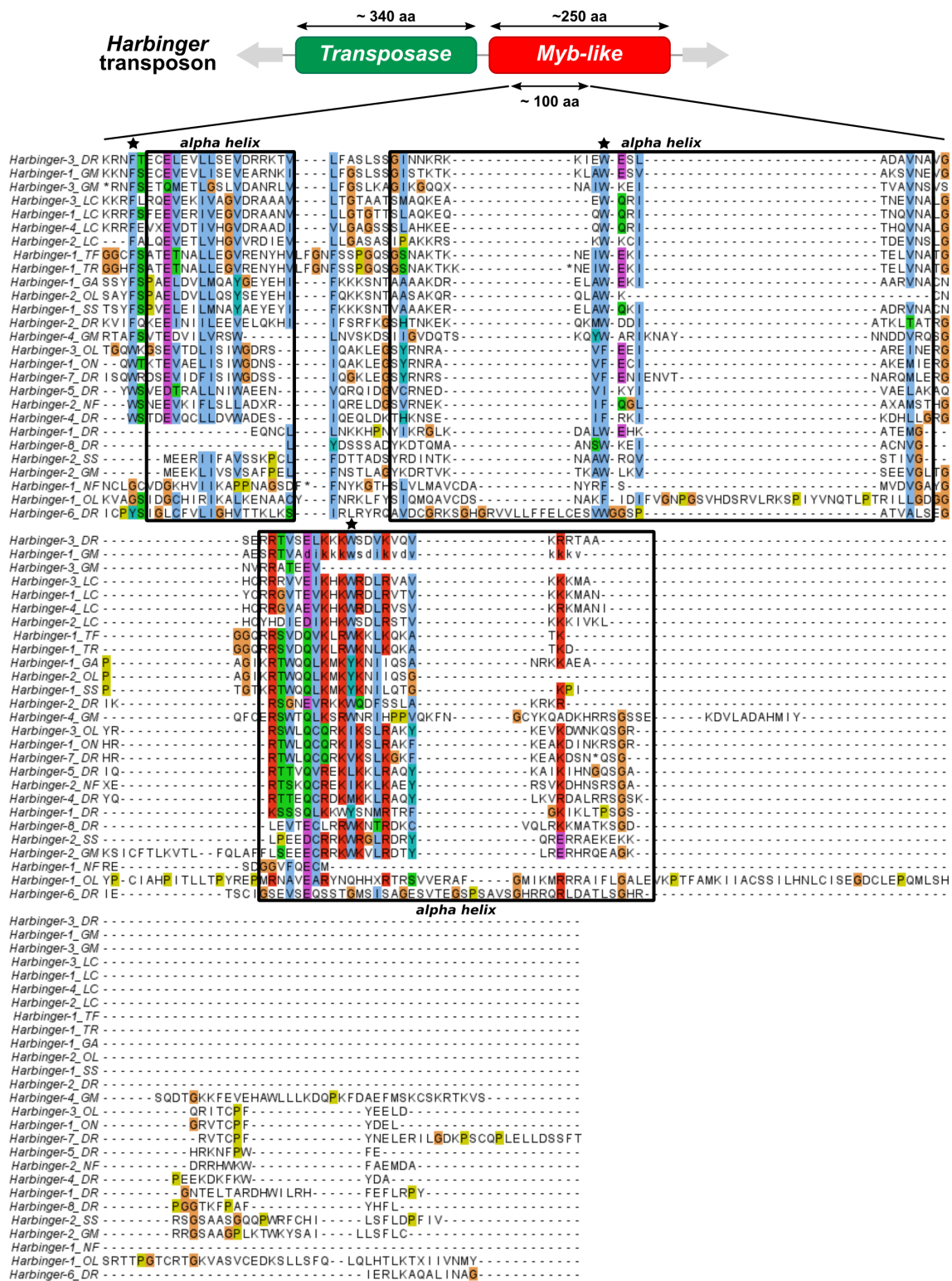


FIGURE 2.4 – Multiple alignments of the *Harbinger*-like Myb-like proteins on the Myb-like domain. Predicted alpha-helix motifs are represented with black dashed squares; bulky aromatic residues, which are essential for alpha helix structure stabilization, are indicated by black stars. (DR : *Danio rerio* – zebrafish, GA : *Gasterosteus aculeatus* – stickleback, GM : *Gadus morhua* – cod, LC : *Latimeria chalumnae* – coelacanth, NF : *Nothobranchius furzeri* – killifish, OL : *Oryzias latipes* – medaka, ON : *Oreochromis niloticus* – tilapia, SS : *Salmo salar* – salmon, TF : *Takifugu flavidus* – fugu, TR : *Takifugu rubripes* – fugu, XM : *Xiphophorus maculatus* – platyfish).

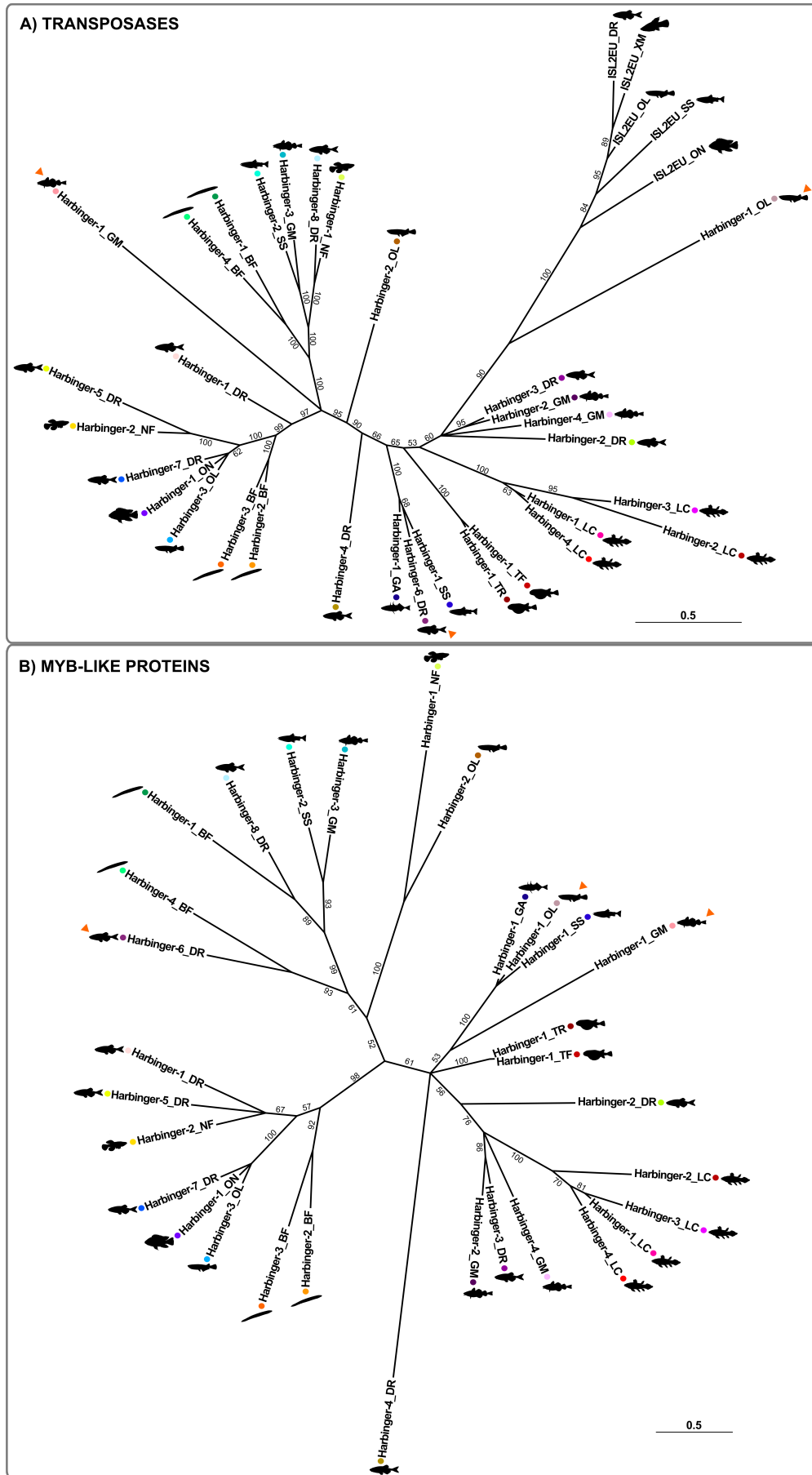


FIGURE 2.5 – Phylogenetic relationships between *Harbinger* and *ISL2EU* transposases (A) and Myb-like (B) proteins from different fish species. Legend is on next page.

FIGURE 2.5 – The tree was constructed using the Bayesian method (Huelsenbeck & Ronquist, 2001). Colored dots indicate the correspondence of the transposases and Myb-like proteins from a same *Harbinger* element. (BF : *Branchiostome floridae* – amphioxus, DR : *Danio rerio* – zebrafish, GA : *Gasterosteus aculeatus* – stickleback, GM : *Gadus morhua* – cod, LC : *Latimeria chalumnae* – coelacanth, NF : *Nothobranchius furzeri* – killifish, OL : *Oryzias latipes* – medaka, ON : *Oreochromis niloticus* – tilapia, SS : *Salmo salar* – salmon, TF : *Takifugu flavidus* – fugu, TR : *Takifugu rubripes* – fugu, XM : *Xiphophorus maculatus* – platyfish).

and a phylogenetic tree was constructed using both Bayesian and Maximum Likelihood methods (Figure 2.5B and Supplementary Figure 2.10B) (Guindon & Gascuel, 2003; Huelsenbeck & Ronquist, 2001). We calculated the congruence index Icong (de Vienne et al., 2007) between the phylogenies of *Harbinger* transposases and Myb-like proteins, which revealed that the trees are more congruent than expected by chance (p-value = 0.005) (Figure 2.6). The results therefore suggested a unique origin of the transposase and *Myb-like* gene association in *Harbinger* transposons. However, we noticed that the phylogenetic positions of three sequences (indicated with orange arrowheads in Figure 2.5) were different in the transposase and Myb-like Bayesian phylogenies. The position of one of them (Harbinger-1_GM) was not statistically supported and different in Bayesian and Maximum Likelihood trees, indicating lack of resolution. However, the phylogenetic positions of the proteins of the two other elements (Harbinger-1_OL and Harbinger-6_DR) were better supported, suggesting potential recombination events in these *Harbinger* sequences.

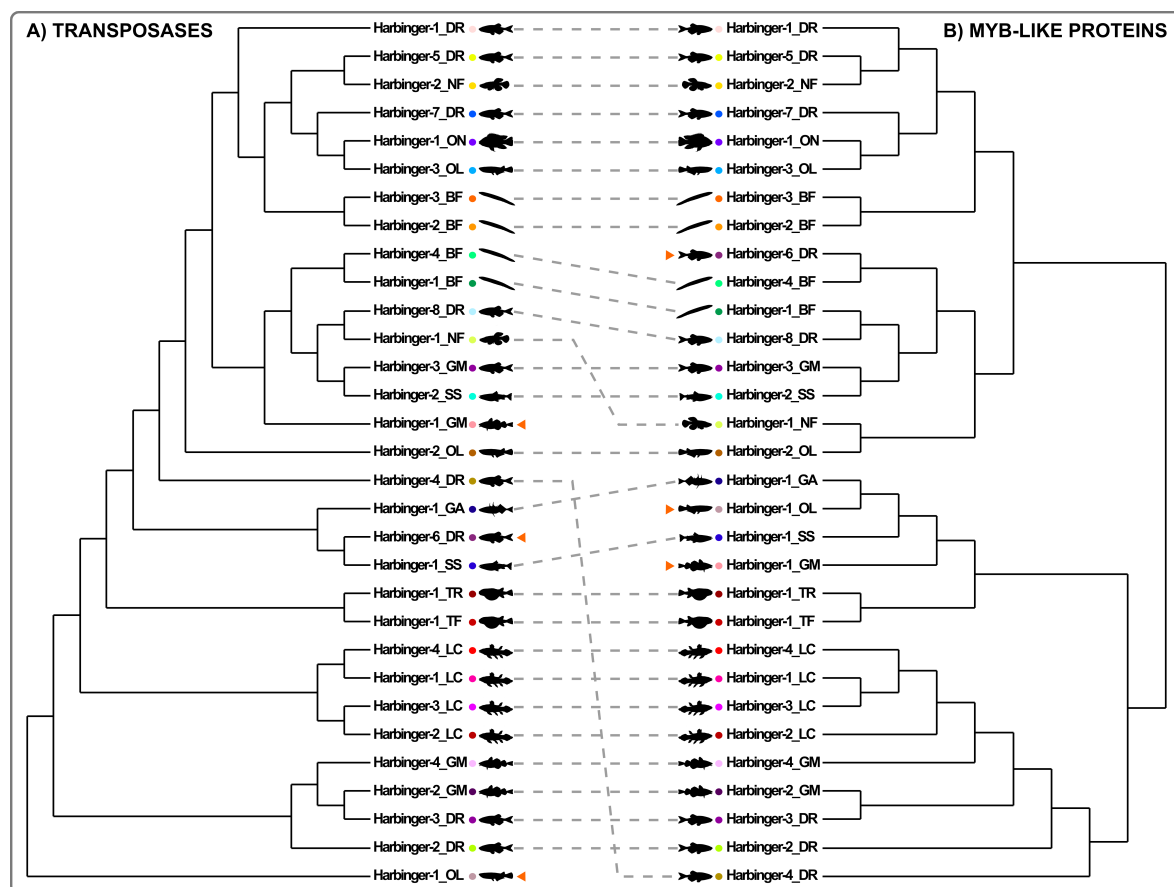


FIGURE 2.6 – Phylogenetic congruence between transposase and Myb-like proteins of *Harbinger* transposons from fish genomes. The trees, of the transposases (A, cf. Figure 2.5A) and Myb-like (B, cf. Figure 2.5B) proteins were constructed using bayesian method (Huelsenbeck & Ronquist, 2001). The correspondence of the transposases and Myb-like proteins from a same *Harbinger* element are indicated with colored dots and connected by dashed lines. Non congruent position of three elements are indicated with orange arrowheads.

2.2.4.5 Harbinger-like transposons are expressed in fish

Despite the broad distribution of *Harbinger* transposons, studies on their activity remain scarce (Elewa et al., 2017; Markova & Mason-Gamer, 2017; Smith et al., 2012). To the best of our knowledge, no proof of activity has been reported so far for *ISL2EU* transposons.

Using the PhyloFish database (Pasquier et al., 2016), which allows investigating sequence expression thanks to multiple fish transcriptome data from multiple organs, *Harbinger*-like transposon expression was detected in numerous fish species. Particularly, expression datasets were accessible for four of the species studied in **Figure 2.1** : spotted gar, zebrafish, cod and medaka. Expression of *Harbinger* and *ISL2EU* transposons was found in all four species (except for *ISL2EU* transposons, which are absent from spotted gar) (**Figure 2.7**). *Harbinger* and *ISL2EU* were particularly highly expressed in different organs such as brain and gills, as well as in testis and embryos in some species.

We further investigated expression in the medaka, the species with the highest genome coverage of *Harbinger* transposons in our analysis (**Figure 2.1**). We focused on medaka gonads, since transposition activity in germ cells allows transmission of new insertions and expansion of TE families. Analysis of RNAseq data showed expression of both *Harbinger* and *ISL2EU* transposons in both male and female gonads, each family showing similar expression levels (**Figure 2.8**). Transposon expression was testis-biased, i.e. higher in male than in female gonads.

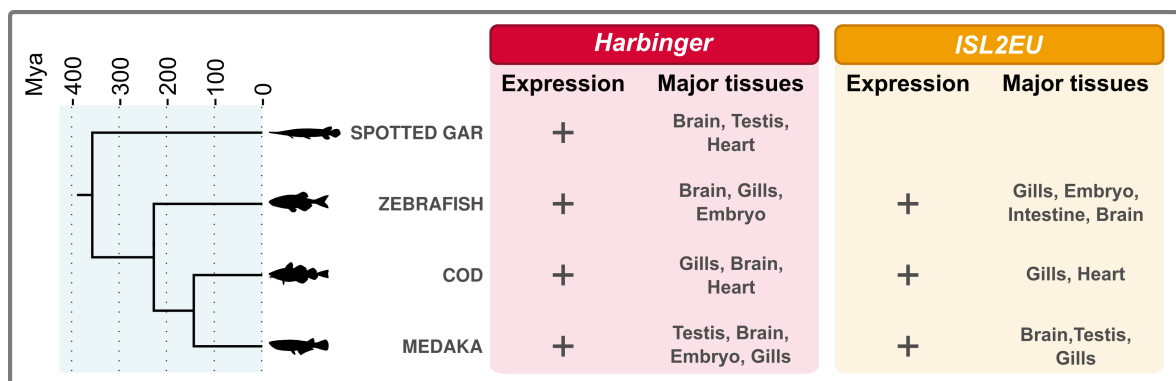


FIGURE 2.7 – Expression analysis of *Harbinger* and *ISL2EU* transposons in spotted gar, zebrafish, cod and medaka using the PhyloFish database (Pasquier et al., 2016). Expression identified in PhyloFish database is indicated with +. Absence of + indicates the absence of the element in the genome. For each species the organs where the TEs are mainly expressed are indicated.

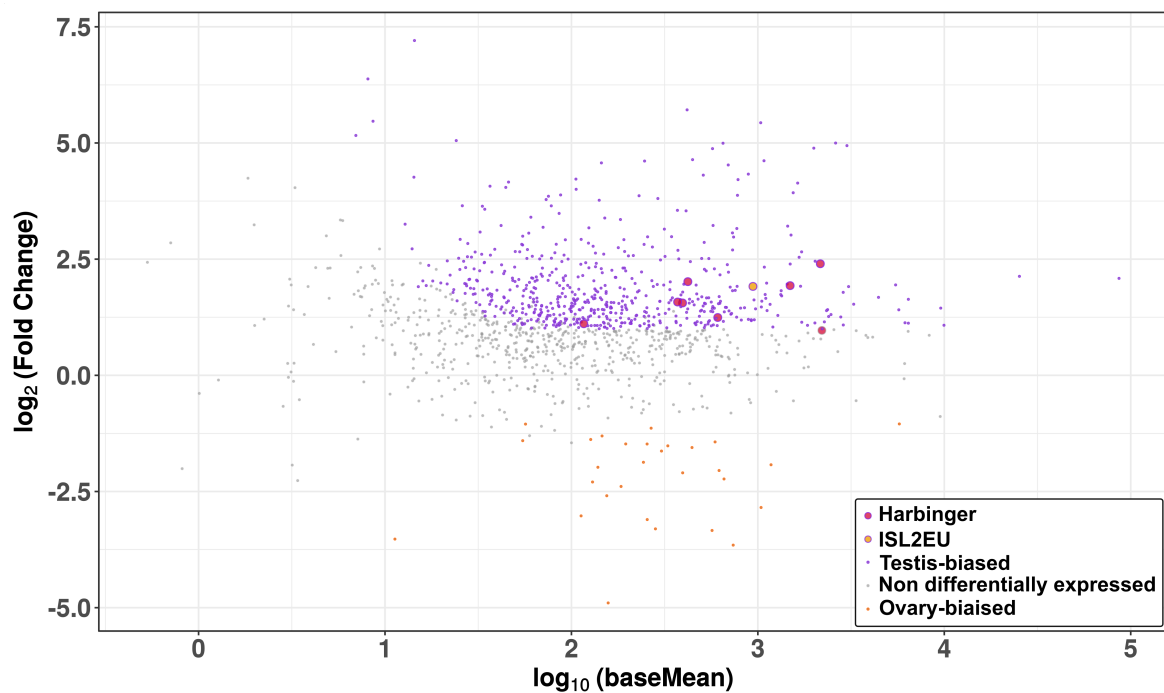


FIGURE 2.8 – MAplot representing the relative expression in male and female gonads of all TE families of the medaka genome. Each dot corresponds to the relative expression of one TE family in RNAseq data. The x-axis corresponds to the signal intensity averaged across all replicates, and the y-axis to the \log_2 Fold Change of expression between testis and ovary ($\log_2\text{FC}$). The higher the $\log_2\text{FC}$ of a TE family is, the more it is over-expressed in testes (significant over-expression in testes is indicated in purple), and the lower it is, the more it is over-expressed in ovaries (significant over-expression in ovaries is indicated in orange). The more the gene is on the right, the more it is overall expressed across all replicates. The *Harbinger* and *ISL2EU* families are highlighted with bigger red- or yellow-colored dots, respectively.

2.2.5 Discussion

2.2.5.1 *Harbinger*-like transposons are inequitably widespread in fish genomes

Harbinger-like transposons are found in diverse eukaryotic clades such as vertebrates, arthropods, fungi and plants (Casola et al., 2007; Han et al., 2015; Jiang et al., 2003; Kapitonov & Jurka, 2004; Markova & Mason-Gamer, 2015; Pereira et al., 2013; Yuan & Wessler, 2011; Zhang et al., 2001, 2004). Within vertebrates, they are absent from mammalian genomes but have been identified in other vertebrate species including fish (Han et al., 2015; Kapitonov & Jurka, 2004). Teleost fish, which are the most species-rich clade within vertebrates, have genomes with highly diversified TE composition, both quantitatively and qualitatively, particularly compared to mammals and birds. Therefore, focusing on this group of animals is of particular interest to understand the evolutionary history of *Harbinger*-like transposons. Here, we have focused our analysis on eight teleost fish species, one non-teleost ray-finned fish (spotted gar), one sarcopterygian fish (coelacanth), one cartilaginous fish (elephant shark) and one cephalochordate (amphioxus) (**Figure 2.1**). This led us to the observation that *Harbinger* transposons are widespread in ray-finned fish including teleosts but in variable amounts, with the highest genomic success in the medaka. They are also present at high copy numbers in coelacanth and amphioxus but absent from elephant shark genome, suggesting secondary loss.

ISL2EU elements are scarcer. They were not detected in the non-teleost species included in this study, i.e. amphioxus, elephant shark, spotted gar and coelacanth. Within teleost fish, *ISL2EU* transposons were found in zebrafish, cod, platyfish, medaka and tilapia but neither in stickleback nor in the two pufferfishes fugu and tetraodon. We propose that *ISL2EU* elements might have been introduced ca. 200-300 million years ago at the base of the teleost fish lineage through horizontal transfer from a more divergent species, with subsequent secondary loss in a common ancestor of stickleback and pufferfishes. Overall, our results suggested that, even if *Harbinger* and *ISL2EU* transposons are related, they present different evolutionary dynamics that might be due to different abilities to invade genomes or different mechanisms of repression in their hosts.

2.2.5.2 Evolutionary relationships between the two ORFs of *Harbinger* transposons

In order to investigate the evolutionary dynamics of *Harbinger*-like transposons, we have studied the phylogenetic relationships of the transposases of elements from different fish genomes, focusing on the DDE domain (**Figure 2.5A**). This allowed the identification of several families of fish *Harbinger* transposons. Some of them might be more ancient, dating back to a chordate common ancestor of amphioxus and fish. In contrast, others are apparently more recent and teleost-specific.

Furthermore, we studied the phylogenetic relationships between the Myb-like proteins of *Harbinger* transposons from different species. Myb-like proteins were less constrained than transposases, although their secondary structure was well conserved (**Figure 2.3** and **Figure 2.4**). Transposases were highly similar even between different families of *Harbinger* transposons. In contrast, Myb-like proteins were diverse and presented only sporadic similarities in restricted parts of the Myb-like domain, making their comparison more difficult (Kapitonov & Jurka, 2004). This questioned whether the two *Harbinger* proteins followed the same evolutionary trajectories. In teleost

fish genomes (**Figure 2.5** and **Supplementary Figure 2.10**), transposase and Myb-like protein phylogenies were consistent with respect to the element they belonged together. This indicated that, even if the two types of proteins presented different degrees of conservation, they shared common evolutionary history within *Harbinger* transposons. Hence, our results were consistent with a single origin of the association between transposase and Myb-like protein in *Harbinger*, with possible rare events of recombination during evolution between elements belonging to different families. Such a recombination might be restricted by the fact that co-evolution between the two ORFs within a same element is necessary to maintain interactions between the two proteins for a successful transposition.

Some peculiar *Harbinger*-like transposons called *Spy*, characterized by a single ORF, have been identified in invertebrates (Han et al., 2014, 2015). This ORF encodes a transposase with a DDE domain but also a helix-turn-helix (HTH) motif, which is believed to act as a sequence-specific DNA-binding domain. Hence, this HTH motif might fulfil the same function as the DNA-binding domain of the Myb-like protein in *Harbinger* elements. We did not identify any *Harbinger* transposon with a single ORF in teleost fish, confirming that both ORFs are probably essential for autonomous *Harbinger* transposition in this clade. To date, single ORF *Harbinger* transposons have only been found outside of vertebrates (Han et al., 2014, 2015). This could suggest either that the structure with two ORFs of the *Harbinger* transposons is more efficient for its spreading and maintenance in fish, or that the single ORF type has been more easily repressed and eliminated in the lineage that led to vertebrates.

2.2.5.3 *Harbinger*-like transposons are transcriptionally active in teleost fish

We report here the expression of *Harbinger*-like elements in teleost fish. Using the PhyloFish database (Pasquier et al., 2016), we have detected the expression of these transposons in spotted gar, zebrafish, cod and medaka (**Figure 2.7**). *Harbinger* and *ISL2EU* transposons were mainly expressed in the same tissues in these species, suggesting common mechanisms of activation and repression. Moreover, in the medaka both *Harbinger* and *ISL2EU* transposons showed a testis-biased expression in gonads (**Figure 2.8**).

TEs are repressed in genomes, this limiting their potential deleterious effects through insertional mutagenesis. This repression can occur through the prevention of transcription (generally with epigenetic marks, such as DNA methylation or histone modifications) or post-transcriptionally (with piRNAs for example) (Iwasaki et al., 2015; Rebollo et al., 2012; Sarkar et al., 2017; Zemleni et al., 2009). In the medaka, piRNAs, which can mediate the cleavage of transposable element mRNAs, are more expressed in testis compared to ovaries (Kneitz et al., 2016). However, TE expression is the result of both transcription and repression. Therefore, the higher expression we observed in testis for *Harbinger*-like transposons is probably due to a stronger transcription in this organ, which would not be completely compensated by piRNA inhibition, if any.

2.2.6 Conclusion

This work characterized *Harbinger*-like transposons in teleost fish genomes. Even if these elements represent small parts of these genomes, they are widespread in this clade. *Harbinger* and *ISL2EU* transposons are also transcriptionally active in fish. Since their discovery, beyond their neutral or negative effects, multiple works have demonstrated the propensity of TEs to be positively recruited by host genomes as new regulatory and coding sequences (Chuong et al., 2017; Cosby et al., 2019; Etchegaray et al., 2021). *Harbinger* transposons do not make exception to this, as multiple cases of *Harbinger*-derived genes have been reported in various organisms (Casola et al., 2007; Cosby et al., 2021; Duan et al., 2017; Kapitonov & Jurka, 2004; Liang et al., 2015; Sinzelle et al., 2008; Velanis et al., 2020; Zhou et al., 2021). The ability of *Harbinger* transposons to form new genes may be linked to the presence of two ORFs encoding proteins with useful and different molecular properties that can interact together (Sinzelle et al., 2008).

A recent study has shown that *Harbinger* transposons have invaded the genome of the sea kraits *Laticauda* about 15-25 million years ago and compose as much as 8-12 % of their DNA (Galbraith et al., 2021). In these organisms, several insertions occurred into introns, regulatory regions and exons, and have even added coding sequences into exons, conferring potential adaptation. In the tomato *Solanum lycopersicum*, light stress conditions induce expression of genes having *Harbinger* transposons (among other TEs) located in their genomic proximity (mostly in their introns) (Deweth et al., 2022). The authors suggest that these elements serve in a stress regulatory network to adapt rapidly to new environments. Thus, *Harbinger* transposons represent an interesting and beneficial reservoir of useful sequences for species adaptation in teleost fish and other organisms.

2.2.7 Acknowledgements

This work was supported by grants from the Agence Nationale de la Recherche (ANR) and from the Ecole Normale Supérieure de Lyon. EE thanks the “Fondation pour la Recherche Médicale” (FRM) for financial support through an end of PhD program fellowship.

2.2.8 Supplementary figure



FIGURE 2.9 – Distribution of *Harbinger* and *ISL2EU* transposons on medaka chromosomes. Each either red or orange colored bar corresponds to *Harbinger* or *ISL2EU* elements, respectively, at one genomic location. The height of each bar is proportional to the number of elements at a given position (smallest bars correspond to a single copy of an element).

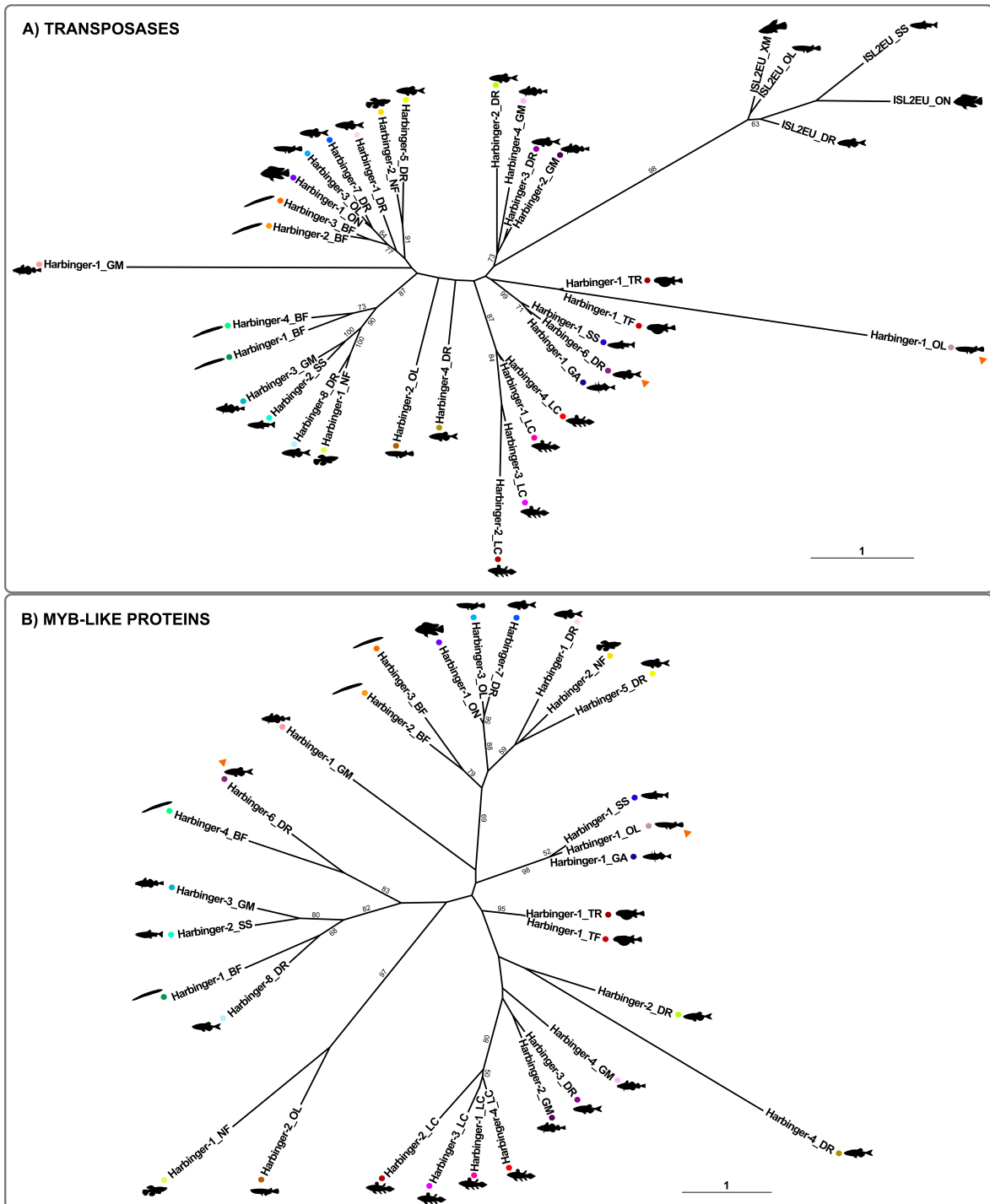


FIGURE 2.10 – Phylogenetic relationships between *Harbinger* transposase (A) and Myb-like (B) proteins of different fish species. The tree was constructed using the Maximum Likelihood method (Guindon & Gascuel, 2003). Only bootstrap values higher than 50 % are shown. Colored dots indicate correspondence between the transposases and Myb-like proteins from a same *Harbinger* element. (BF : *Branchiostome floridae* – amphioxus, DR : *Danio rerio* – zebrafish, GA : *Gasterosteus aculeatus* – stickleback, GM : *Gadus morhua* – cod, LC : *Latimeria chalumnae* – coelacanth, NF : *Nothobranchius furzeri* – killifish, OL : *Oryzias latipes* – medaka, ON : *Oreochromis niloticus* – tilapia, SS : *Salmo salar* – salmon, TF : *Takifugu flavidus* – fugu, TR : *Takifugu rubripes* – fugu, XM : *Xiphophorus maculatus* – platyfish).

3

Étude d'une famille de gènes dérivés de transposons *Harbinger* chez les vertébrés

Sommaire

3.1	Introduction	78
3.2	Domestication moléculaire récurrente des transposons <i>Harbinger</i> chez les vertébrés	78
3.3	Mise au point d'un protocole expérimental de KO-direct par CRISPR/Cas9 pour le criblage de phénotype à la première génération	79
3.3.1	Le gène <i>SAIYAN</i> comme gène-cible pour la mise au point du protocole de KO-direct par CRISPR/Cas9	81
3.3.2	Reproduction du phénotype de l'inactivation de <i>SAIYAN</i> avec la méthode de CRISPR/Cas9 utilisant de multiples sgRNAs	82
3.3.3	Optimisation du protocole de CRISPR/Cas9	83
3.4	Criblage des fonctions des gènes dérivés de transposons <i>Harbinger</i> chez le poisson-zèbre	86
3.5	Article : The neurodevelopmental gene <i>MSANTD2</i> belongs to a gene family formed by recurrent molecular domestication of <i>Harbinger</i> transposons at the base of vertebrates	89
3.6	Conclusion	139
3.7	Annexe	140

3.1 Introduction

La génération de nouveaux gènes est une source importante d'innovation et d'adaptation pour les organismes, puisqu'ils représentent un substrat majeur pour l'émergence de nouvelles fonctions. Ohno suggérait que les événements de duplications de gènes sont une des sources principales d'extension du répertoire de gènes permettant l'apparition de nombreuses innovations (Ohno, 1999). Cependant, l'impact des autres mécanismes de formation de gènes ne doit pas être sous-estimé. Les éléments transposables peuvent en particulier être exaptés pour former de nouveaux gènes via leur domestication moléculaire. En effet, les gènes *SYNCYTIN*, dérivés de gènes d'enveloppes de rétrovirus endogènes (ERV), sont impliqués dans le développement du placenta, essentiel dans l'évolution des mammifères placentaires. Plus ancestralement, les gènes *RAG*, dérivés de transposon à ADN *Transib*, ont permis l'apparition du système immunitaire adaptatif caractéristique des vertébrés à mâchoire. Ces deux exemples témoignent de l'importance de la contribution des ETs à la formation de nouveaux gènes à l'origine d'innovations majeures. L'étude et la caractérisation des gènes dérivés d'ETs sont donc essentielles pour mieux comprendre l'évolution des traits génétiques et phénotypiques des organismes. Cependant, l'identification de tels gènes est encore probablement sous-estimée du fait des difficultés d'annotation des ETs dans les génomes. Le développement récent des technologies de séquençage et des méthodes d'analyse des génomes a permis d'accroître considérablement la quantité de génomes séquencés, favorisant l'annotation et la caractérisation de séquences d'ETs. Cependant, en plus d'une difficulté d'identification, le manque général de caractérisation de ces gènes au niveau fonctionnel induit une sous-estimation de leur implication dans l'évolution des espèces.

Dans ce chapitre, je me suis intéressée à l'impact de la domestication moléculaire des ETs dans l'évolution précoce des vertébrés. Plus particulièrement, j'ai pu identifier une nouvelle famille de gènes dérivés de transposons *Harbinger* chez les vertébrés. Cette étude fait l'objet d'un article qui a été soumis au journal *Molecular Biology and Evolution* (MBE) présenté en **section 3.5**, et dont je résumerai les éléments principaux en **section 3.2**.

3.2 Domestication moléculaire récurrente des transposons *Harbinger* chez les vertébrés

L'équipe Génomique Évolutive des Poissons a pu développer au cours de ces dernières années une banque de séquences d'ETs de vertébrés comprenant des séquences provenant de bases de données publiques ainsi que de nouvelles séquences d'ETs en particulier de poissons, annotées notamment lors de participations à des projets de séquençage de génomes (Amemiya et al., 2013; Braasch et al., 2016; McGaugh et al., 2014). C'est la comparaison de cette banque d'ETs aux séquences de gènes humains, qui nous a permis d'identifier tout d'abord *MSANTD2* comme un nouveau cas de gènes dérivés d'ETs. Plus particulièrement, *MSANTD2* est issu de la domestication moléculaire du gène *Myb-like* de transposon ADN de type *Harbinger*. Des analyses supplémentaires m'ont permis d'identifier trois autres gènes, *MSANTD1*, *MSANTD3* et *MSANTD4*, comme étant également dérivés de gènes *Myb-like* de transposons *Harbinger*. Ceux-ci s'ajoutent aux gènes *NAIF1* et *HARB11*, précédemment identifiés comme gènes dérivés du même type de transposons

(Kapitonov & Jurka, 2004; Sinzelle et al., 2008). *NAIFI* et *HARBII* dérivent respectivement du gène *Myb-like* et du gène de la transposase. Tous ces gènes sont présents chez tous les vertébrés à mâchoire inclus dans cette étude, sauf *MSANTD3* qui est présent uniquement chez les sarcoptérygiens. Nous avons pu établir que ces gènes sont issus d'au moins trois événements de domestications moléculaires indépendants ayant eu lieu chez un ancêtre commun des vertébrés à mâchoire, soit il y a environ 500 millions d'années. L'origine exacte des gènes *MSANTD3*, *MSANTD4* et *NAIFI*, soit par domestications indépendantes soit par duplications, reste difficile à établir de façon certaine.

L'étude de ces gènes chez le poisson-zèbre m'a permis d'observer qu'ils sont exprimés durant le développement embryonnaire précoce ainsi que dans les tissus adultes. J'ai également observé une co-expression de ces gènes dans le cerveau mâle adulte. Ils sont aussi exprimés dans de nombreux organes humains et notamment dans le cerveau humain, en particulier durant les deux premiers trimestres du développement fœtal.

Enfin, dans le but d'étudier la fonction de ces gènes, des techniques d'inactivation de gènes ont été utilisées : « knockdown » (KD) par injection d'oligonucléotides antisense de type morpholino et « knockout » (KO)-direct par la technique de Clustered Regularly Interspaced Palindromic Repeat / Cas9 (CRISPR/Cas9) (voir **section 3.3** et **3.4**) (Wu et al., 2018). Ces techniques m'ont permis de mettre en évidence que l'inactivation du gène *MSANTD2* produit des embryons avec des défauts développementaux sévères correspondant à des retards de développement, ainsi qu'à des malformations de la tête et du système nerveux. Plus particulièrement, les tubes neuraux et les ventricules du cerveau sont mal formés, et nous observons également des patterns neuronaux anormaux. Tout ceci, indique un rôle de *MSANTD2* dans le développement du système nerveux, potentiellement dans la migration neuronale. Ces résultats font écho à d'autres travaux qui ont montré que *MSANTD2* est associé à des maladies neuro-développementales humaines, telles que les troubles du spectre de l'autisme et la schizophrénie (Lim et al., 2017; O'Brien et al., 2018; Schizophrenia Working Group of the Psychiatric Genomics Consortium, 2014; Zhang et al., 2020).

Ces résultats sont développés et analysés en détail dans la **section 3.5**.

3.3 Mise au point d'un protocole expérimental de KO-direct par CRISPR/Cas9 pour le criblage de phénotype à la première génération

Les gènes dérivés de transposons *Harbinger* identifiés correspondent à des gènes annotés dans les génomes. Certains de ces gènes ont été mis en lien avec plusieurs maladies humaines. *MSANTD1* est associé à la coronaropathie (maladie cardiaque lié au rétrécissement des artères coronaires) (van der Harst & Verweij, 2018) et est un facteur de susceptibilité à la tuberculose (Qi et al., 2017). *MSANTD2* a été relié à des maladies neuro-développementales. Il a en effet été reporté comme étant enrichi en mutations post-zygotiques dans une étude sur l'un des plus larges groupes de patients atteints de troubles du spectre de l'autisme (Lim et al., 2017). De plus, *MSANTD2* a été associé à la schizophrénie par des études d'associations pangénomiques mais également du fait de l'association de l'augmentation de son expression et des facteurs de risque génétiques (O'Brien et al., 2018; Schizophrenia Working Group of the Psychiatric Genomics Consortium, 2014; Zhang et al., 2020). *MSANTD3* a été mis en lien dans plusieurs études avec le carcinome à cellules acineuses

des glandes salivaires (Andreasen et al., 2019; Barasch et al., 2017; Lee et al., 2020) et sa protéine peut interagir avec la nucléoprotéine des virus influenza (Generous et al., 2014). Enfin, *NAIFI* code un facteur anti-apoptotique qui a été relié à de nombreux cancers comme l'ostéosarcome (Kong & Zhang, 2018), le cancer gastrique (Luo et al., 2011; Yang et al., 2014), de la prostate (Fu & Cao, 2015) et du poumon (Zhao et al., 2015). *MSANTD4* a quant à lui été associé au maintien de la température corporelle chez des populations bovines de Sibérie dans des conditions de stress lié au froid (Igoshin et al., 2019). Seuls ces quelques travaux ayant étudié ces gènes, leurs fonctions restent encore non identifiées. Afin d'investiguer cet aspect, j'ai utilisé le modèle du poisson-zèbre chez lequel cinq des six gènes dérivés de transposons *Harbinger* identifiés chez les vertébrés sont présents (*MSANTD1*, *MSANTD2*, *MSANTD4*, *NAIFI* et *HARBII*).

Le poisson-zèbre est un modèle présentant de nombreux avantages comme une forte production d'œufs, une fécondation externe, sa transparence à l'état embryonnaire, ainsi que le séquençage et l'annotation de son génome, faisant de lui un modèle de choix pour l'étude de la fonction des gènes. Les techniques d'édition du génome, notamment la technique CRISPR/Cas9, ont révolutionné la manipulation génétique de cet organisme (Hwang et al., 2013b; Li et al., 2016). La technique de CRISPR/Cas9 repose sur l'utilisation d'un sgRNA, qui sert de guide pour cibler un gène d'intérêt, et d'une protéine Cas9, qui va couper l'ADN à l'endroit ciblé. Cette technique permet ainsi une édition précise du génome. L'obtention d'une lignée stable d'individus mutés sur leurs deux allèles – individus KO – par édition du génome est essentielle afin de conclure quant à l'effet de l'inactivation d'un gène sur l'organisme ou son développement. Cependant, avec la méthode classique d'édition par CRISPR/Cas9, l'obtention de tels individus n'est généralement possible qu'au bout d'au minimum deux générations (F2), ce qui induit un délai expérimental assez long (**Figure 3.1A**). Il s'agit d'autant plus d'un facteur contraignant lorsque plusieurs gènes veulent être étudiés simultanément, puisque cela multiple la quantité de poissons à élever, génotyper et phénotyper. Ainsi, l'utilisation d'une méthode novatrice de CRISPR/Cas9 permettant l'obtention de mutants directement à la génération F0 – c'est-à-dire chez les embryons directement injectés – récapitulant les phénotypes mutants F2, représente un avantage considérable. La technique publiée par Wu et al. (2018) représente une telle méthode. En effet, leur système consiste à cibler un gène d'intérêt, non pas avec un seul guide sgRNA comme ce qui est fait classiquement, mais avec trois ou quatre sgRNAs simultanément, permettant de multiplier les sites et les probabilités de coupures du gène d'intérêt (**Figure 3.1B**). Cette technique permet de produire des mutants au phénotype nul (phénotype KO) dès la F0. En effet, les auteurs estiment que plus de 90 % des embryons injectés avec cette technique produisent un phénotype récapitulant celui obtenu par la méthode classique de transmission verticale de mutations. Il est à noter que les individus obtenus en F0 sont mutés sur leurs deux allèles mais de façon mosaïque, c'est-à-dire que les mutations sur les deux allèles et dans les différentes cellules de l'organisme ne seront pas forcément identiques. Cette méthode permet ainsi de réduire le temps expérimental de plusieurs mois à quelques jours.

Ainsi, je me suis inspirée de cette méthode utilisant de multiples sgRNAs pour cribler les phénotypes obtenus lors du développement embryonnaire et larvaire du poisson-zèbre, suite à l'inactivation des gènes dérivés de transposons *Harbinger*.

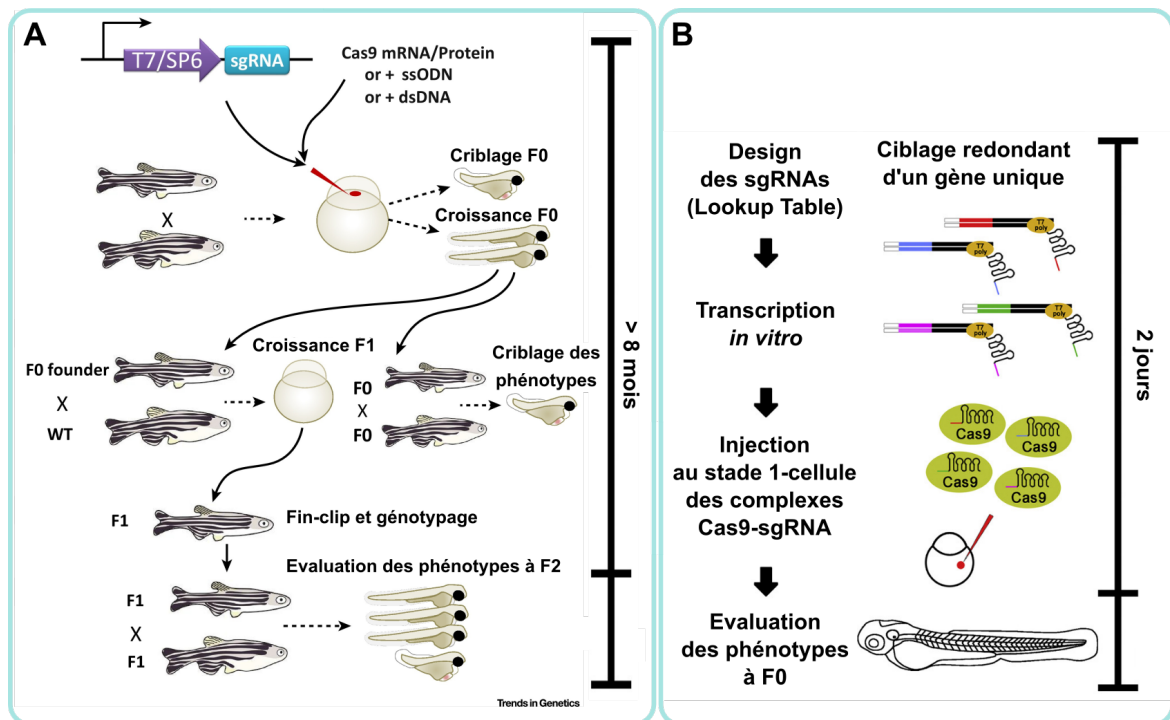


FIGURE 3.1 – **Comparaison des méthodes de génération de mutants par CRISPR/Cas9.** (A) Méthode classique : après la synthèse *in vitro* d'un sgRNA ciblant le gène d'intérêt, il est injecté avec la protéine Cas9 dans l'embryon au stade 1-cellule. L'efficacité de la mutagenèse du sgRNA est évaluée par génotypage des individus. Les animaux mutés sont élevés puis croisés avec des poissons sauvages (WT) ou bien entre eux. Les animaux obtenus en F1 sont génotypés puis croisés afin d'obtenir une génération F2 avec des animaux dont les deux allèles du gène d'intérêt sont mutés. A chaque étape d'obtention d'embryons, les phénotypes peuvent être criblés. L'évaluation finale des phénotypes se réalise en F2. Adapté de Li et al. (2016). (B) Méthode permettant l'évaluation du phénotype directement en F0 : 4 sgRNAs différents sont conçus pour cibler un gène unique d'intérêt. (Un tableau fournissant 4 sgRNAs pour chaque gène annoté dans le génome du poisson-zèbre dans la base de données Ensembl est fourni avec l'article de Wu et al. (2018)). Après la transcription *in vitro* des sgRNAs, les complexes sgRNA-Cas9 sont formés puis rassemblés en mélange équimolaire pour être injectés dans l'embryon au stade 1-cellule. L'évaluation des phénotypes se réalise directement en F0. Adapté de Wu et al. (2018).

3.3.1 Le gène *SAIYAN* comme gène-cible pour la mise au point du protocole de KO-direct par CRISPR/Cas9

Afin de mettre au point le protocole de Wu et al. 2018 dans nos conditions de laboratoire, mon gène cible a été *SAIYAN*, dont l'inactivation produit un phénotype facilement observable et quantifiable. *SAIYAN* a été identifié comme un gène surexprimé dans les bandes de peau blanches du poisson-clown *Amphiprion ocellaris* comparé aux bandes de peau orange (Salis et al., 2019). L'inactivation de ce gène chez le poisson-zèbre a montré qu'il est essentiel pour le développement des iridophores – un type de cellules pigmentaires iridescentes – puisque les mutants présentent une diminution significative du nombre de ces cellules au stade larvaire (Figure 3.2). L'inactivation de ce gène a été réalisée par la méthode de CRISPR/Cas9 en utilisant 2 sgRNAs co-injectés ciblant le gène *SAIYAN* par Salis et al. (2019).

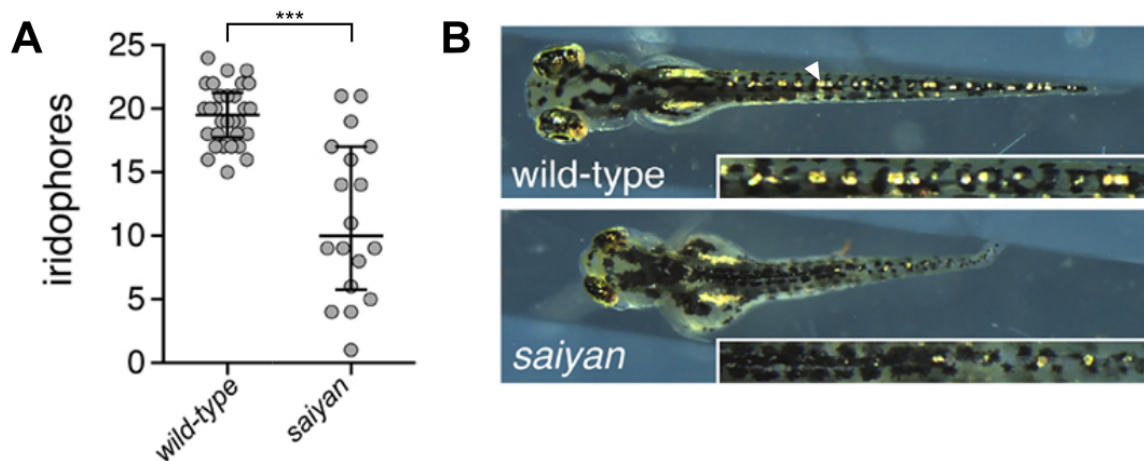


FIGURE 3.2 – L'inactivation du gène *SAIYAN* par Crispr/Cas9 chez le poisson-zèbre produit un phénotype observable et quantifiable de diminution du nombre d'iridophores au stade larvaire. Adapté de Salis et al. (2019). (A) Quantification des iridophores dorsaux chez le poisson-zèbre à 4 jours post-fertilisation (jpf). Chaque point correspond à un individu. Les astérisques indiquent une différence significative du nombre d'iridophores (comparaison non-paramétrique avec contrôle selon la méthode de Steel, *** : p value<0.001). L'effet de la différence de taille des larves observée entre les WT et les individus mutés, a été contrôlé et n'influe pas la variation du nombre d'iridophores. (B) Phénotypes observés pour les individus contrôles (wild-type) et les individus mutés de façon mosaïque par CRISPR/Cas9 pour *SAIYAN*. Les individus mutés pour *SAIYAN* présentent un nombre plus faible d'iridophores même après la prise en compte de la différence de taille des larves. Les iridophores correspondent aux points iridescents, ici visibles sur la partie dorsale de la queue des larves de poisson-zèbres, un iridophore a été indiqué à l'aide d'une tête de flèche.

3.3.2 Reproduction du phénotype de l'inactivation de *SAIYAN* avec la méthode de CRISPR/Cas9 utilisant de multiples sgRNAs

Dans un premier temps, mon objectif a été de reproduire le phénotype produit par l'inactivation de *SAIYAN* observé par Salis et al. (2019) avec un protocole adapté de celui de Wu et al. (2018).

Pour cela, j'ai utilisé 3 sgRNAs ciblant ce gène du poisson-zèbre : 2 sgRNAs utilisés par Salis et al. (2019) et un autre que j'ai conçu (**Annexe 3.1**). De plus, contrairement au protocole de Wu et al. où les sgRNA sont transcrits *in vitro*, les sgRNAs que j'ai utilisés sont entièrement synthétiques. En effet, ces derniers sont connus pour avoir une meilleure efficacité d'édition de l'ADN (sgRNA synthétisé par SYNTHOGO).

Afin de tenter de reproduire le phénotype des mutants *SAIYAN* j'ai donc co-injecté les trois complexes sgRNA-Cas9 en mélange équimolaire dont la concentration finale était de 5 μ M de Cas9 et 30 μ M de sgRNA, comme proposé par Wu et al. (2018). Les embryons ont été injectés au stade 1-cellule dans le vitellus. Les résultats ont été analysés à 4 jpf (**Figure 3.3**). J'ai pu observer que, dès la génération F0, les individus présentaient une diminution significative du nombre d'iridophores dorsaux comparé aux individus contrôles (WT). Certains individus ne présentaient même aucun iridophore. De plus, j'ai observé que, dans nos conditions de laboratoire, les individus présentaient une diminution du nombre d'iridophores sans que la taille des individus ne soit affectée, au contraire de ce qui avait été observé par Salis et al. (2019). Comme Salis et al. (2019), j'ai pu enfin constater une forte variabilité du nombre d'iridophores en fonction des individus injectés. Ceci est à mettre en lien avec la méthode CRISPR/Cas9, qui produit des individus mosaïques en F0. En effet, dès les premières minutes après l'injection et lors des premières divisions cellulaires, l'efficacité

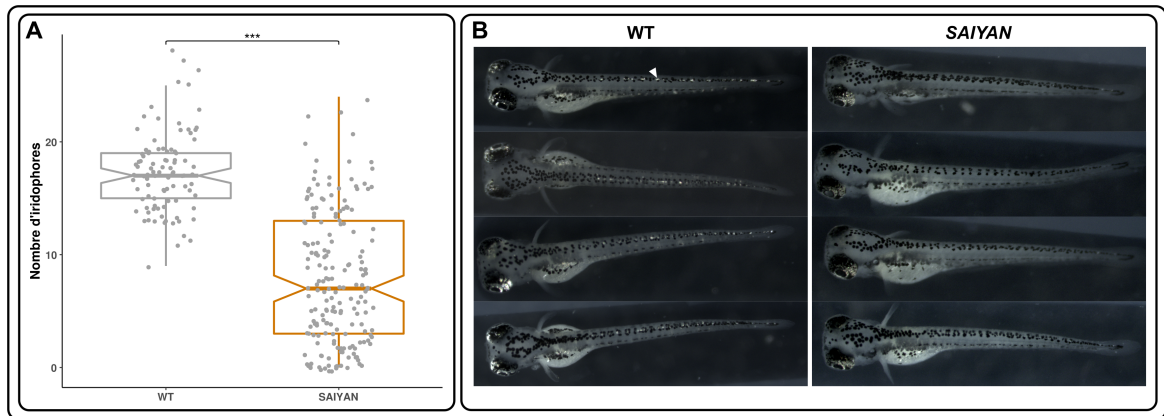


FIGURE 3.3 – **Reproduction du phénotype de réduction du nombre d'iridophores à 4 jpf lors de l'inactivation du gène *SAIYAN* par CRISPR/Cas9 selon la méthode de Wu et al. (2018) chez le poisson-zèbre.** (A) Quantification des iridophores dorsaux chez le poisson-zèbre à 4 jours post-fertilisation (jpf). Chaque point correspond à un individu. Les astérisques indiquent une différence significative du nombre d'iridophores (test de Student, *** : p value<0.0001). (B) Phénotypes observés pour les individus contrôles (WT) et les individus injectés avec 3 sgRNAs ciblant *SAIYAN*. Les iridophores correspondent aux points iridescents visibles sur la partie dorsale de la queue des larves de poissons-zèbres, un iridophore a été indiqué à l'aide d'une tête de flèche.

et la variabilité de coupure de la protéine Cas9 détermineront la transmission des mutations à l'ensemble ou une partie des cellules de l'organisme.

En conséquence, j'ai pu reproduire le phénotype des mutants *SAIYAN* chez le poisson-zèbre et vérifier l'applicabilité du protocole de Wu et al. (2018). J'ai donc pu valider mon modèle d'étude, que j'ai utilisé dans un second temps pour optimiser ce protocole de CRISPR/Cas9.

3.3.3 Optimisation du protocole de CRISPR/Cas9

Dans l'optique de s'assurer de la reproductibilité de la méthode et de l'optimiser, plusieurs paramètres ont été testés.

Tout d'abord, je me suis intéressée à une possible variabilité d'efficacité des sgRNAs. Pour ce faire, j'ai comparé le nombre d'iridophores dorsaux des individus à 4 jpf suite à l'injection individuelle de chaque sgRNA ou de plusieurs combinaisons de sgRNAs ciblant le gène *SAIYAN* (Figure 3.4). Les résultats indiquent que le nombre d'iridophores est plus faible lorsque seul le sgRNA 1 est injecté comparé à toutes les autres conditions. Le sgRNA 2 ne semble quant à lui que faiblement efficace. Ainsi, ces résultats illustrent la variabilité d'efficacité des sgRNAs. Dans le cas présent l'utilisation d'un seul sgRNA semble plus efficace que la combinaison de plusieurs. Cependant, dans les travaux de Wu et al. (2018), c'est bien lors de l'utilisation de plusieurs sgRNAs que le plus d'individus avec un phénotype nul ont été produits. L'utilisation de quatre sgRNAs ciblant un unique gène est un bon outil pour cribler plus facilement les phénotypes liés à l'inactivation d'un gène aux fonctions inconnues (car l'inefficacité d'un sgRNA peut être compensée par l'efficacité des autres). Cependant, il paraît ensuite important de tester les sgRNAs individuellement lorsqu'un phénotype a pu être caractérisé afin de vérifier l'efficacité des différents sgRNAs et éventuellement mieux choisir les sgRNAs à injecter. L'injection des sgRNAs individuellement peut également permettre d'étudier la reproductibilité des phénotypes et ainsi éliminer l'hypothèse que les phénotypes

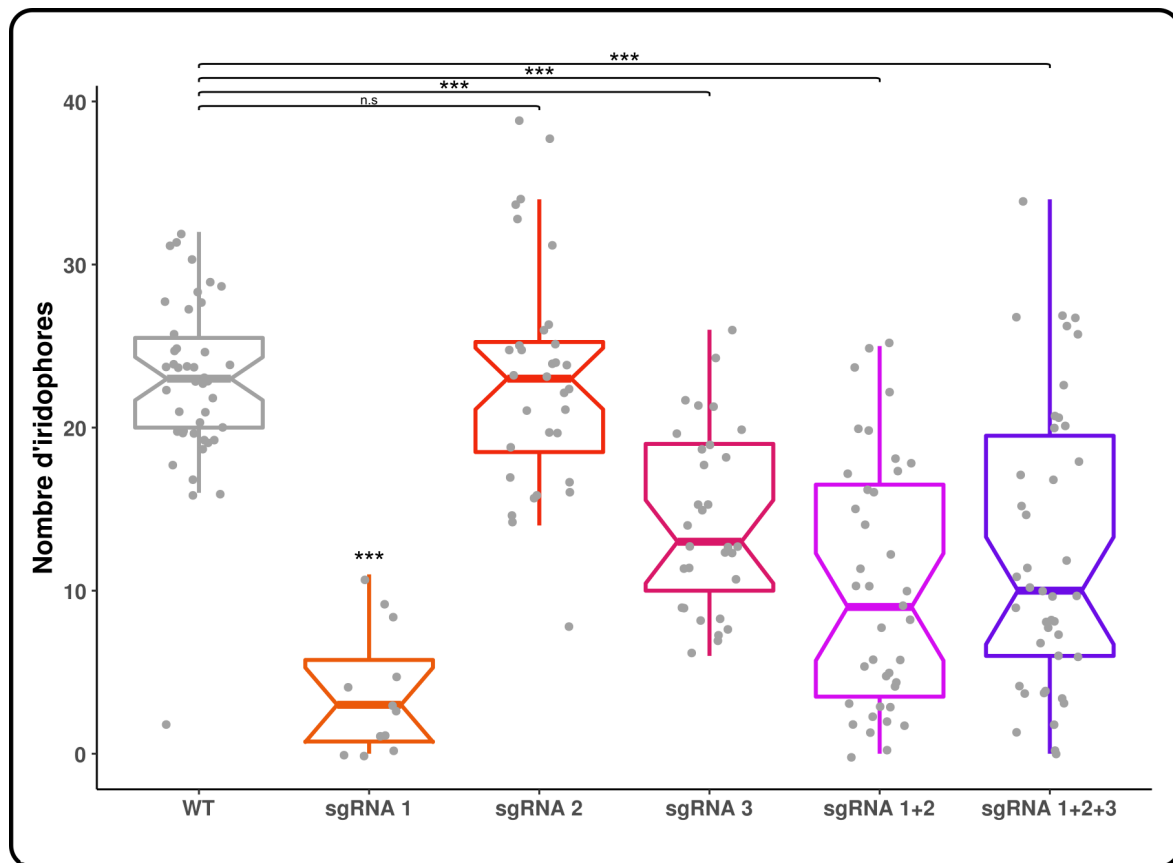


FIGURE 3.4 – Comparaison de l'efficacité des différents sgRNA ciblant le gène *SAIYAN* par CRISPR/Cas9 chez le poisson-zèbre. Quantification des iridophores dorsaux chez le poisson-zèbre à 4 jpf selon l'injection d'un seul sgRNA (sgRNA 1, sgRNA 2, sgRNA 3), de deux sgRNA (sgRNA 1+2) ou de trois sgRNA (sgRNA 1+2+3) ciblant le gène *SAIYAN* (injection dans le vitellus, 15 μ M Cas9 – 20 μ M sgRNAs au total). Chaque point correspond à un individu. Les astérisques indiquent une différence significative du nombre d'iridophores (test de Student, n.s : non significatif, *** : pvalue<0.0005. Les *** placées au-dessus de la condition sgRNA 1 indiquent une différence significative avec toutes les autres conditions).

observés pourraient provenir de la toxicité ou de « off-targets » (c'est-à-dire la coupure de sites génomiques non-ciblés initialement) d'un sgRNA.

Dans un deuxième temps, j'ai voulu tester l'influence de la zone d'injection. En effet, au stade 1-cellule les embryons peuvent être injectés dans le vitellus ou bien directement dans la cellule. La cellule de l'œuf nouvellement fertilisé se positionne au-dessus du sac vitellin (ou vitellus). Au cours des premiers stades de développement, il n'y a pas de séparation membranaire entre les cellules et le sac vitellin, ce qui permet les échanges de flux. On peut tout de même se questionner quant à la rapidité et l'efficacité des échanges entre la/les cellules et le sac vitellin. En effet, il est important lors de l'édition du génome que la/les mutations puissent se faire avant ou très rapidement après la première division cellulaire, pour qu'elles puissent être transmises au maximum de cellules de l'organisme. Ceci tendrait donc à privilégier une injection directement dans la cellule. Par ailleurs, l'injection dans le sac vitellin présente de nombreux avantages, elle est plus simple et surtout plus rapide, ce qui permet d'injecter beaucoup plus d'embryons au stade 1-cellule, avant que la première division cellulaire n'ait lieu. Ainsi, pour déterminer quelle était la méthode d'injection la plus efficace, j'ai injecté des sgRNAs ciblant le gène *SAIYAN* dans des embryons au stade 1-cellule soit dans la cellule soit dans le vitellus (**Figure 3.5**). L'efficacité des deux méthodes d'injection a été

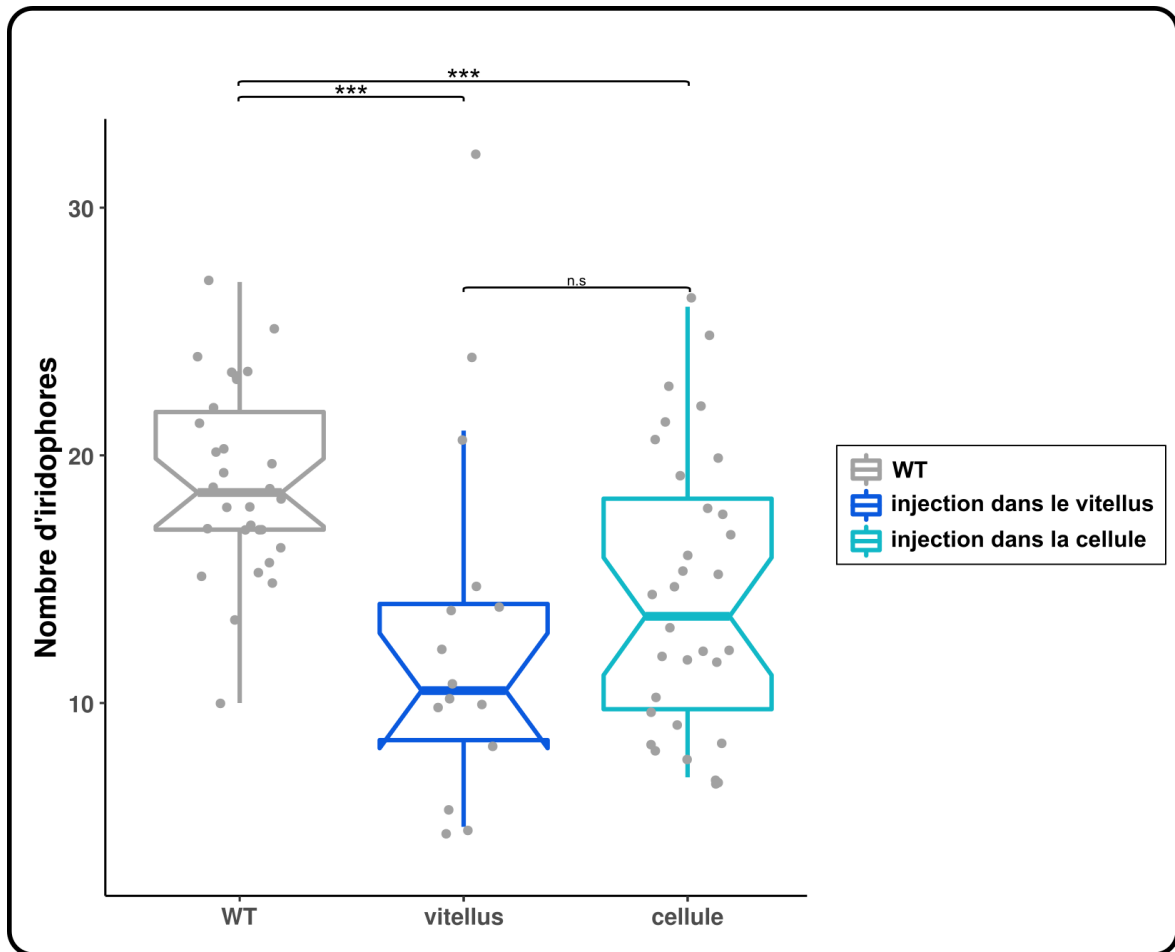


FIGURE 3.5 – **Comparaison de la méthode d'injection des complexes sgRNA/Cas9 lors de l'inactivation du gène *SAIYAN* par CRISPR/Cas9 chez le poisson-zèbre.** Quantification des iridophores dorsaux chez le poisson-zèbre à 4 jpf selon l'injection dans le vitellus ou dans la cellule des complexes sgRNA/Cas9 ciblant le gène *SAIYAN* (15 μ M Cas9 – 30 μ M sgRNAs total). Chaque point correspond à un individu. Les astérisques indiquent une différence significative du nombre d'iridophores (test de Student, n.s : non significatif, *** : pvalue<0.0005).

évaluée en comparant le nombre d'iridophores dorsaux des poissons-zèbres à 4 jpf. Les résultats obtenus indiquent que le nombre d'iridophores des individus n'est pas significativement différent entre les modes d'injection cellule/vitellus.

Suite à ces résultats j'ai sélectionné la méthode d'injection dans le vitellus, car les avantages pratiques de cette technique n'ont pas d'impact négatif important sur la diminution du nombre d'iridophores et donc sur l'efficacité d'édition du système CRISPR/Cas9. Ainsi, pour toutes les analyses réalisées par la suite, les injections ont été réalisées dans le sac vitellin des embryons.

Enfin, j'ai évalué plusieurs concentrations, caractérisées par différents ratios sgRNA :Cas9, pour les injections afin de sélectionner la plus performante (**Figure 3.6**). En effet, dans la littérature, il existe de nombreuses divergences quant au ratio sgRNA :Cas9 optimal à utiliser pour maximiser la mutagenèse (Hoshijima et al., 2019; Kroll et al., 2021; Wu et al., 2018). Trois concentrations ont été testées : la concentration C1 (identique à celle du protocole proposé par Wu et al. (2018)) avec un ratio sgRNA :Cas9 de 6 :1 ; ce ratio est de 2 :1 pour la concentration C2 ; et 1.3 :1 pour la concentration C3. J'ai pu observer que le nombre d'iridophores dorsaux des individus est généralement plus faible lors de l'utilisation de la concentration C3. Ainsi, pour toutes les analyses réalisées par la suite, les

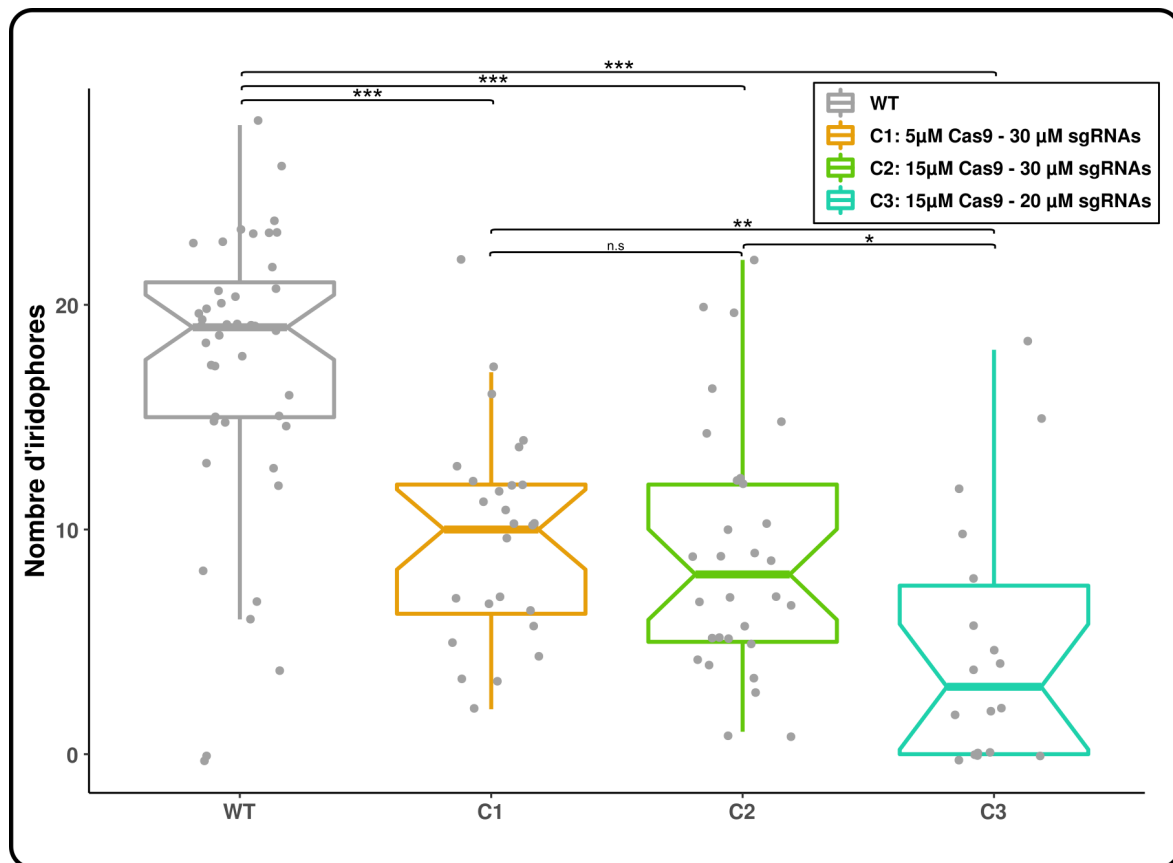


FIGURE 3.6 – Comparaison de l'efficacité de différentes concentrations d'injection de sgRNA/Cas9 lors de l'inactivation du gène *SAIYAN* par CRISPR/Cas9 chez le poisson-zèbre. Quantification des iridophores dorsaux chez le poisson-zèbre à 4 jpf selon les concentrations C1 (5µM Cas9 – 30µM sgRNAs total), C2 (15µM Cas9 – 30µM sgRNAs total) ou C3 (15µM Cas9 – 20µM sgRNAs total). Chaque point correspond à un individu. Les astérisques indiquent une différence significative du nombre d'iridophores (test de Student, * : pvalue<0.05, ** : pvalue<0.005, *** : pvalue<0.0005).

injections ont été réalisées selon la concentration C3.

3.4 Criblage des fonctions des gènes dérivés de transposons *Harbinger* chez le poisson-zèbre

Lors de ma thèse, j'ai pu identifier une nouvelle famille de gènes dérivés de transposons *Harbinger* chez les vertébrés, (voir **section 3.2** et **3.5**). La fonction des gènes *MSANTD1*, *MSANTD2*, *MSANTD3*, *MSANTD4*, *NAIF1* et *HARB11* conservés chez les vertébrés à mâchoire (chez les sarcoptérygiens pour *MSANTD3*) est indéterminée jusqu'à présent. Afin d'étudier cet aspect, j'ai utilisé le protocole présenté en **section 3.3** pour cribler les phénotypes potentiellement produits par l'inactivation de ces gènes chez le poisson-zèbre. Afin de s'affranchir d'éventuels biais dus à cette méthode novatrice de KO-direct par CRISPR/Cas9, plusieurs contrôles ont été réalisés. Les crispants (les individus F0 ayant subis l'injection du système CRISPR/Cas9) des différents gènes dérivés de transposons *Harbinger* ont été comparés à : (1) des WT, c'est-à-dire des individus non-injectés ; (2) des individus injectés avec des sgRNAs « scrambled » (SCRBL), c'est-à-dire des sgrNA aux séquences aléatoires n'ayant pas de cible prédite dans le génome du poisson-zèbre ; (3) des individus injectés

avec des sgrNAs ciblant le gène *SAIYAN*; (4) des individus injectés avec des sgrNAs ciblant le gène de la tyrosinase (*TYR*), enzyme impliquée dans la production de mélanine par les cellules pigmentaires de type mélanophores, dont l'inactivation produit des individus sans aucune pigmentation mélanique.

Les résultats ont pu montrer que l'inactivation des gènes *MSANTD1*, *MSANTD4*, *NAIF1* et *HARBII* produit des individus présentant un léger retard de développement visible à l'aspect général des embryons à 24 heures post fertilisation (hpf) (**Figure 3.7**). Ce retard de développement est un peu plus marqué pour *NAIF1* (observé pour environ 35 % des embryons injectés), car il semble plus proche phénotypiquement d'un stade 25 somites (environ 22 hpf) que d'un stade 24 hpf (Kimmel et al., 1995). Les retards de développement des crispants *MSANTD1*, *MSANTD4* et *HARBII* sont compensés au bout de 4 à 6 heures, mais pas forcément pour *NAIF1*. Pour tous les crispants des individus dysmorphiques ont été observés (malformations générales sévères) à des taux inférieurs à 14 % pour *MSANTD1*, *MSANTD4* et *HARBII* et d'environ 25 % pour *NAIF1*. Selon Wu et al. (2018) le protocole de KO-direct par CRISPR/Cas9 produit des phénotypes dysmorphiques dus à la toxicité de l'expérimentation pour moins de 17 % des individus injectés. Pour les contrôles que j'ai réalisés ce taux était inférieur à 5 %. Pour *NAIF1*, ce taux est plus élevé qu'attendu, pouvant suggérer un phénotype lié à l'inactivation du gène. Cependant les phénotypes de malformations observés ne sont pas toujours similaires entre les individus, au sein d'une même expérimentation avec un même traitement, ne nous permettant pas de conclure à ce stade quant à la spécificité des phénotypes observés. Ainsi, l'inactivation des gènes *MSANTD1*, *MSANTD4*, *NAIF1* et *HARBII* par cette méthode ne nous a pas permis, dans une première approche de criblage, de conclure sur les rôles biologiques potentiels de ces gènes chez le poisson-zèbre.

Les résultats concernant l'inactivation du gène *MSANTD2* sont détaillés en **section 3.5**.

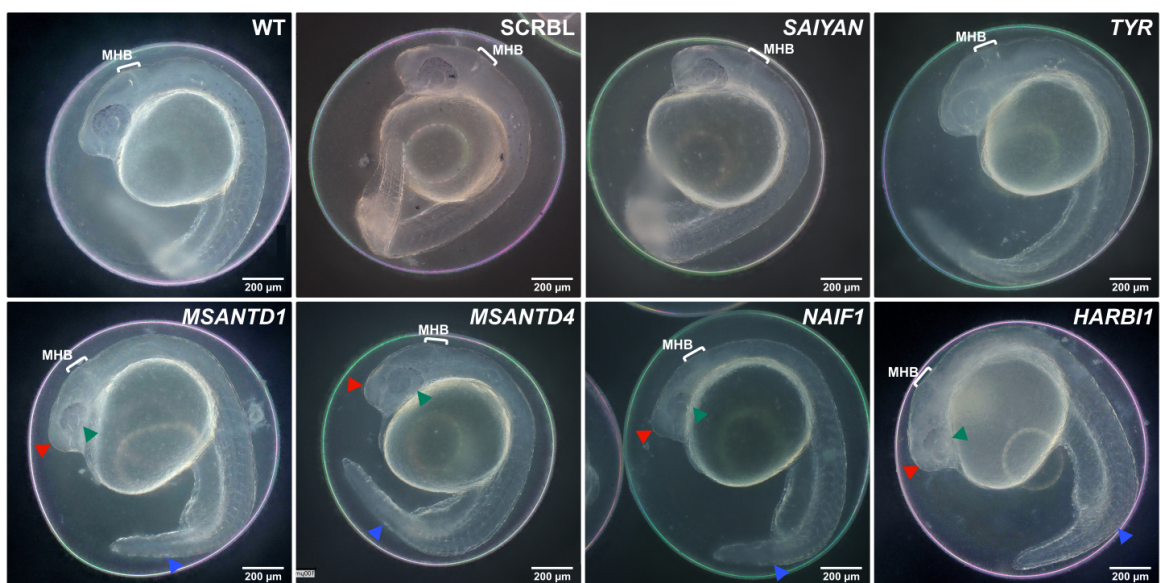


FIGURE 3.7 – **Phénotypes observés lors de l'inactivation des gènes dérivés de *Harbinger* (*MSANTD1*, *MSANTD4*, *NAIF1*, *HARBI1*) par CRISPR/Cas9 chez le poisson-zèbre en comparaison des individus contrôles (*WT*, *SCRBL*, *SAIYAN*, *TYR*) au stade 24 hpf.** Les individus *MSANTD1*, *MSANTD4*, *NAIF1* et *HARBI1* présentent un léger retard de développement visible à l'aspect général des embryons Ceci s'observe notamment par l'aspect général de la tête qui correspond à des stades antérieurs (tête de flèche rouge); le manque de pigmentation au niveau des yeux (tête de flèche vert); la queue plus courte/moins développée (tête de flèche bleu); la jonction midbrain-hindbrain (MHB) caractéristique du stade 24hpf peu ou pas visible (la zone où devrait se trouver cette jonction est indiquée par une accolade blanche).

3.5 Article : The neurodevelopmental gene *MSANTD2* belongs to a gene family formed by recurrent molecular domestication of *Harbinger* transposons at the base of vertebrates

1 **Submission type:** Article

2 **Section:** Discoveries

3

4

5 **Title:**

6 The neurodevelopmental gene *MSANTD2* belongs to a gene family
7 formed by recurrent molecular domestication of *Harbinger* transposons
8 at the base of vertebrates

9

10

11 **Authors:**

12 Ema Etchegaray¹, Dominique Baas², Magali Naville¹, Zofia Haftek-Terreau¹, Jean-
13 Nicolas Volff¹

14

15

16 **Affiliations:**

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

19

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

21

22

23 **Corresponding author:**

24 Jean-Nicolas Volff , E-mail: jean-nicolas.volff@ens-lyon.fr

25

26

27

28

29

30

31

32

33 **ABSTRACT**

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

57

58 **INTRODUCTION**

59 The formation of new genes is an important source of evolutionary innovation and
60 adaptation for species. Indeed, they represent a major substrate for the emergence
61 of new functions and contribute to the birth of novel phenotypic traits that are source
62 of adaptation and speciation. For example, new genes can be generated *de novo*
63 from scratch from initially non-functional sequences, a rare phenomenon, or through
64 the duplication of preexisting genes, which can lead to new functions thanks to
65 mutations and relaxed selective constraints. Another source of new genes is the

66 recruitment, also called molecular domestication, of transposable element (TE)-
67 coding sequences.

68 TEs are repeated DNA sequences that can insert into novel genomic locations and
69 thus can cause genomic instability through insertion and recombination (Kato et al.
70 2002). TEs have been found in every species that have been investigated. However,
71 the quantitative and qualitative composition of TEs in genomes is variable depending
72 on the species (Huelsenbeck and Ronquist 2001). While TEs are mutagenic agents
73 that can have neutral or deleterious effects on genomes (Ohno 1972; Doolittle and
74 Sapienza 1980; Orgel and Crick 1980), they can also serve as material for the
75 formation of new regulatory sequences, new exons or even new genes (Kidwell and
76 Lisch 2000; Warren et al. 2015; Chuong et al. 2017). TEs have been source of major
77 innovations during evolution, as exemplified by vertebrate development (Etchegaray
78 et al. 2021). By the process of molecular domestication, TEs can give rise to new
79 functional genes positively selected in host genomes. Major examples of TE
80 domestication have been documented in vertebrates, such as the *RAG* genes
81 involved in the adaptive immune system and the *SYNCYTIN* genes necessary for
82 placenta development in mammals (Mallet et al. 2004; Dupressoir et al. 2011;
83 Kapitonov and Koonin 2015; Etchegaray et al. 2021). Thus, TE molecular
84 domestication can lead to important adaptive innovations. In the human genome,
85 which is composed at least of 45% of TEs (Lander et al. 2001), a hundred cases of
86 protein-coding genes derived from TEs have been identified so far (Volf 2006).
87 However, most of these genes have been poorly characterized, particularly at the
88 functional level. Considering the quantity and diversity of TEs in genomes, their role
89 in the diversification and adaptation of organisms is probably still underestimated
90 (Brandt et al. 2005; Britten 2006; Volf 2006; Alzohairy et al. 2013). Therefore, the
91 identification and functional characterization of new cases of TE-derived genes is
92 important to better understand the formation of novel genes and the factors driving
93 genetic innovation.

94 In the course of a study aiming to assess the impact of TE molecular domestication
95 on the early evolution of the vertebrate lineage, we have identified several genes
96 domesticated from *Harbinger* transposable elements through the comparison of
97 human protein sequences to a vertebrate-wide TE sequence database. *Harbinger*
98 transposons are DNA transposons present in the genome of protists, plants, insects,
99 worms and vertebrates but absent from mammals (Kapitonov and Jurka 2004). They

100 are generally flanked by Terminal Inverted Repeat sequences (TIRs) and encode two
101 proteins, a transposase with a DDE endonuclease motif and a SANT-Myb-trihelix
102 motif-containing protein, which we will now refer to as the Myb-like protein. Both
103 genes have been shown to be necessary for *Harbinger* transposition (Sinzelle et al.
104 2008; Hancock et al. 2010). The Myb-like protein contains a tri-helix motif with
105 conserved bulky aromatic residues that allows DNA and protein binding. Myb-like
106 proteins are responsible for the nuclear import of the transposase through interaction
107 with its N-terminal end. Thanks to the tri-helix motif they also bind the TIRs of the
108 transposon, allowing the recruitment of the transposase and thus the
109 excision/insertion of the sequence (Sinzelle et al. 2008).

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

120 This study describes a family of genes derived from *Myb-like* genes of *Harbinger*
121 DNA transposons in jawed vertebrates. We have identified four new genes that have
122 been formed through three to four independent molecular domestication events
123 during vertebrate evolution, three at the base of jawed vertebrates about 500 million
124 years ago and a fourth one possibly in a common ancestor of sarcopterygians ca.
125 430 million years ago. The *Harbinger*-derived genes are expressed during zebrafish
126 embryonic development and in zebrafish adult tissues, predominantly in male brain,
127 as well as in human brain during fetal development. Inactivation of one of these
128 genes, *MSANTD2*, by CRISPR/Cas9 direct knock-out in F0 and morpholino
129 antisense oligonucleotide knock-down techniques in zebrafish led to embryos with
130 severe brain developmental defects and modification of the expression of
131 characteristic genes involved in vertebrate nervous system development.
132 Interestingly, *MSANTD2* has been associated with neuro-developmental diseases
133 such as autism spectrum disorders and schizophrenia in human (Schizophrenia

134 Working Group of the Psychiatric Genomics Consortium 2014; Lim et al. 2017;
135 O'Brien et al. 2018; Zhang et al. 2020).

136

137 **RESULTS**

138 **Multiple molecular domestication events of *Harbinger* transposons have** 139 **formed a new gene family in vertebrates**

140 Identification of new *Harbinger Myb-like*-derived genes in jawed vertebrates

141 Comparing human protein sequences to TE sequence databases we identified
142 *MSANTD2* (Myb/SANT DNA Binding Domain Containing 2) as a new potential case
143 of molecular domestication from a *Harbinger* DNA transposon. Indeed, the
144 *MSANTD2* predicted protein sequence presented homologies with the Myb-like
145 protein of a *Harbinger* transposon from the genome of the medaka fish *Oryzias*
146 *latipes*, with a conservation score (considering both residue identities and
147 conservation of physico-chemical classes) of 54% in the Myb-like domain region (150
148 aa). The *MSANTD2* gene is located on chromosome 11 in the human genome and is
149 2,384 base pairs (bp) in length, with four exons encoding a protein of 559 amino-
150 acids (aa). Prediction of conserved domains on the whole sequence of the
151 *MSANTD2* protein revealed a single domain, a Myb-like DNA binding domain
152 containing a tri-helix motif. *MSANTD2* was different from the two other genes derived
153 from *Harbinger* transposons previously described in human, *HARBI1* that has been
154 formed from a transposase gene and *NAIF1* from a *Myb-like* gene (Kapitonov and
155 Jurka 2004; Sinzelle et al. 2008).

156 We further identified three additional cases of *Harbinger*-derived genes in the human
157 genome, called *MSANTD1*, *MSANTD3* and *MSANTD4*. These three genes encode
158 predicted proteins with similarities to Myb-like proteins of *Harbinger* transposons. In
159 human, *MSANTD1* is on chromosome 4 and 3,164 bp in length, with three exons
160 coding for a 278 aa protein. *MSANTD3*, on chromosome 9, is a 1,880 bp gene with
161 three exons encoding a 275 aa protein. Finally, *MSANTD4*, on chromosome 11 and
162 4103 bp in length, contains three exons coding for a 345 aa protein. Prediction of
163 conserved domains on each *MSANTD1*, *MSANTD3* and *MSANTD4* protein also
164 revealed a Myb-like domain containing a tri-helix motif (**Figure 1**). This suggested
165 functional homology of the *MSANTD* proteins with the Myb-like proteins of *Harbinger*
166 transposons, with possible DNA- and/or protein-binding properties.

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

179 Moreover, *MSANTD2P1*, an intronless processed pseudogene (according to the loss
180 of its protein-coding capacity), probably originating from the retrotransposition of
181 *MSANTD2* mRNA, was detected in simians, i.e. from human (on chromosome 21) to
182 marmoset but neither in macaques nor in baboons (**Figure S1**). This suggested that
183 *MSANTD2P1* appeared at the base of simians about 36-50 Mya (Perelman et al.
184 2011). By analyzing the 21q23 human genome region where *MSANTD2P1* was
185 located, we observed depletion in protein-coding genes. Indeed, within a region
186 slightly larger than four megabases (Mb) of genomic DNA, no protein-coding gene
187 was detected. In contrast, 51 non-protein-coding RNA genes were annotated in this
188 region: 12 pseudogenes (including 7 retropseudogenes), all of them being
189 transcribed as lncRNAs (long non-coding RNAs), as well as 36 other lncRNA and
190 three miRNA (microRNA) genes. These numbers were consistent with global
191 estimates for pseudogenes as well as lncRNA and miRNA genes in the whole human
192 genome (Lander et al. 2001; Milligan and Lipovich 2014; Hon et al. 2017; Plotnikova
193 et al. 2019). In order to test if the observed depletion in protein-coding genes was an
194 outstanding feature of the 21q23 region containing *MSANTD2P1*, we studied the
195 length distribution of intergenic regions between two consecutive protein-coding
196 genes in the human genome (**Figure S1E**). Among the 16,000 regions studied, most
197 protein-coding genes were relatively close (89.5% of the intergenic regions are
198 smaller than 200,000 bp). Only 0.17% of the analyzed regions, including that
199 containing *MSANTD2P1* (4.04 Mb), were larger than four Mb (27 regions). Thus, the
200 21q23 region was one of the largest (the 25th largest) “protein-coding gene desert”

201 regions in the human genome. Among the 27 largest regions depleted in protein-
202 coding genes, most of them (64%) corresponding to centromeres, the 21q23 region
203 was the non-centromeric region with the highest density of non-protein-coding RNA
204 genes.

205

206 Vertebrate *Myb-like*-derived genes originated from three to five independent
207 molecular domestication events of *Harbinger* transposons

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

221 In order to investigate the evolutionary origin(s) of *MSANTD* genes, phylogenetic
222 trees were built based on protein alignments using the Bayesian method (MrBayes;
223 Huelsenbeck and Ronquist 2001) (**Figure 3**). While the sequence of the transposase
224 of *Harbinger* transposons is highly conserved between different families, this is not
225 the case for the *Myb-like* transposon proteins, which are much more divergent
226 (Kapitonov and Jurka 2004). Such an important sequence divergence was also
227 observed between most MSANTD proteins. Therefore, we were not able to
228 reconstruct reliable general sequence alignment and phylogeny for all MSANTD and
229 transposon *Myb-like* proteins together. However, MSANTD1 and MSANTD2 were
230 most similar to the same group of *Harbinger* sequences, and MSANTD3, MSANTD4
231 and NAIF1 to the same other group of transposon sequences. This allowed
232 generating two different sets of multiple sequence alignments and phylogenies: one
233 for MSANTD1 and MSANTD2 with related *Harbinger* transposon proteins, and
234 another one for MSANTD3, MSANTD4, NAIF1 and related *Harbinger* sequences

235 **(Figure 3)**. Phylogenies indicated that each *MSANTD* sequence from different
236 species formed an independent monophyletic group, and that the closest related
237 *Harbinger* transposons were different for each *MSANTD* sequence. Hence, this
238 supported five independent events of molecular domestication, four at the base of
239 jawed vertebrates and a fifth one later in a common ancestor of sarcopterygians for
240 *MSANTD3*. Phylogenies were also constructed with the Maximum Likelihood method
241 and showed similar results **(Figure S2)**. However, in this analysis, *MSANTD3* and
242 *MSANTD4* did not clearly group with a specific *Harbinger* transposon, and the
243 clustering with transposon was not highly statistically supported for *NAIF1*. Hence,
244 Maximum Likelihood analysis suggested at least three events of molecular
245 domestication, two for *MSANTD1* and *MSANTD2* and at least a third one for
246 *MSANTD3*, *MSANTD4* and *NAIF1*.

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

252 Taken together, the results indicated that *MSANTD1*, *MSANTD2*, *MSANTD3* and
253 *MSANTD4* are four new cases of vertebrate genes derived from *Harbinger Myb-like*
254 transposon sequences, in addition to *NAIF1*. These genes arose from three to four
255 independent molecular domestication events at the base of jawed vertebrates around
256 500 Mya, with another potential one at the base of sarcopterygians around 430 Mya
257 that generated *MSANTD3*.

258

259 Vertebrate *Myb-like*-derived genes evolved under negative selection

260 To further investigate the evolutionary constraints having acted on vertebrate
261 *MSANTD* genes, we performed a positive/negative selection test using CODEML
262 (Yang et al. 2007). We calculated the dN/dS ratio (ratio between non-synonymous vs.
263 synonymous substitution rates) as a proxy for selection pressure **(Table S1)**. All
264 ratios were smaller than 1, reflecting a higher rate of synonymous than non-
265 synonymous substitutions. These ratios were comparable to those of other genes in
266 genomes (0.066 on average in human-zebrafish comparisons) (Wolf et al. 2009).
267 Hence, *MSANTD* sequences evolved under negative/purifying selection in
268 vertebrates, i.e. these genes were functionally constrained. The ratio for the

269 *MSANTD2P1* pseudogene was closer to 1 compared to other *MSANTD* genes, in
270 accordance with relaxed constraints and loss of protein-coding capacity.

271

272 ***Harbinger*-derived genes are expressed in zebrafish during embryonic** 273 **development and predominantly in adult male brain**

274 The expression of the *MSANTD1*, *MSANTD2*, *MSANTD4*, *NAIF1* and *HARBI1*
275 *Harbinger*-derived genes was studied by qPCR in zebrafish embryos (**Figure 4A**). All
276 these genes were expressed during zebrafish embryonic development, with *HARBI1*
277 being the most expressed gene. Except for *MSANTD1*, which is more expressed at
278 later stages, most genes were more strongly expressed at the first stages of
279 development before the midblastula transition (MBT), suggesting maternal effect.
280 Using *in situ* RNA hybridization, *MSANTD2*, which was chosen for further functional
281 analyses (see below), was found to be expressed during zebrafish embryonic
282 development in the whole embryo from 1.25 hours post fertilization (hpf) to 17hpf
283 (**Figure 4C**). From 6hpf, *MSANTD2* was more strongly expressed in the anterior side
284 of the embryo, the region leading to the head and the central nervous system. At
285 24hpf, the expression of *MSANTD2* was specifically restricted to the head region,
286 and more particularly to the forebrain, midbrain and hindbrain regions of the brain.

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

293

294 ***Harbinger*-derived genes are expressed in human, particularly during brain** 295 **development**

296 According to the NIH Genotype-Tissue Expression (GTEx) project (dbGaP Accession
297 phs000424.v8.p2; (GTEx Consortium 2013)), all *Harbinger*-derived genes appeared
298 to be expressed in human brain as well as in some other tissues depending on the
299 gene. Using the BrainSpan Atlas of the Developing Human Brain
300 (www.brainspan.org) (Miller et al. 2014), expression was detected in human brain
301 before and after birth, except for the *MSANTD2P1* pseudogene (**Figure 5**). *MSANTD*
302 genes were expressed in the whole brain particularly during early fetal development

303 at the first/second trimesters of pregnancy, with decreasing expression in the third
304 trimester (around 10 weeks before birth) (**Figure 5A**). *MSANTD3* and *MSANTD4*
305 were the most expressed genes and *MSANTD2* and *NAIF1* presented the same
306 expression pattern but with lower expression. *HARBI1* had a more ubiquitous
307 expression, with higher expression in early fetal development as observed for the
308 *MSANTD* genes, but also later after 13 years. *MSANTD1* presented a more localized
309 expression in a specific brain structure, the striatum, during the second trimester of
310 pregnancy (from 13 to 24 postconceptional weeks (pcw)) (**Figure 5B**). These results
311 showed that *Harbinger*-derived genes are particularly expressed during fetal brain
312 development, in whole human brain for most genes or in a more specific brain region
313 (striatum) for *MSANTD1*.

314

315 ***MSANTD2* inactivation leads to zebrafish embryos with severe neuro-** 316 **developmental defects**

317 The biological function of vertebrate *Harbinger*-derived genes was further
318 investigated by gene inactivation. Gene knock-out was achieved with CRISPR/Cas9
319 technique and knock-down using morpholino antisense oligonucleotides (Nasevicius
320 and Ekker 2000). The CRISPR/Cas9 protocol was adapted from Wu et al. (2018) in
321 order to produce null phenotypes in the F0 generation of zebrafish, i.e. directly in
322 injected embryos (Wu et al. 2018). This protocol implies the co-injection of four
323 different sgRNAs targeting the same gene at four different loci. In a first screening,
324 single gene inactivation of *HARBI1*, *MSANTD1*, *MSANTD4* and *NAIF1* did not
325 produce any clear and visible phenotype in zebrafish embryos, possibly explained by
326 gene redundancy. Therefore, we focused our analyses on the *MSANTD2* gene,
327 which has been associated to human neuro-developmental diseases such as autism
328 spectrum disorders and schizophrenia (Schizophrenia Working Group of the
329 Psychiatric Genomics Consortium 2014; Lim et al. 2017; O'Brien et al. 2018; Zhang
330 et al. 2020).

331 In order to inactivate *MSANTD2* by CRISPR/Cas9, zebrafish embryos were injected
332 with one sgRNA or different combinations of two or four sgRNAs. The observed
333 phenotypes were similar with almost all the different combinations of sgRNAs,
334 although the penetrance was variable (**Figure 6G, S3**). Therefore, only three
335 treatments (combination of four sgRNAs, combination of sgRNAs 1 and 4, sgRNA 4
336 only) will be further detailed (**Figure 6G**). Sequencing of injected embryos revealed

337 mutations at all sgRNA loci, with multiple frameshift nucleotide deletions leading to
338 premature stop codons (**Figure S4-5**). When looking at embryos injected with four
339 sgRNAs, read coverage analysis showed that almost all reads (ca. 90%) showed
340 mutations in the first exon at the sgRNA 1 locus, in addition to mutations at the three
341 other sgRNA loci.

342 Embryos injected with four sgRNAs showed developmental delays as well as tail and
343 nervous system malformations compared to control embryos at 24hpf (**Figure 6A-F**).
344 Heads were smaller and tails curved with not well defined somites (**Figure 6D**, black
345 arrowheads). Moreover, *MSANTD2* CRISPR/Cas9 embryos presented defaults in
346 neural tube folding (**Figure 6F**, white arrowheads), with cell aggregates (i.e. cells that
347 are loosely grouped together) visible around the nervous system (**Figure 6E-F**, black
348 arrows). Similar phenotypes were observed when embryos were injected with
349 sgRNAs 1 and 4 together, and intermediate phenotypes with sgRNA 4 alone (**Figure**
350 **S6**). In a typical experiment (**Figure 6**), about 25% of embryos injected with all four
351 sgRNAs presented strong phenotypes (developmental delays, tail malformations,
352 nervous system malformations, cell aggregates) and around 50% intermediate
353 phenotypes (developmental delays, nervous system malformations, no or few cell
354 aggregates, no tail malformation) (**Figure 6G**). Moreover, *MSANTD2* inactivation
355 appeared to severely compromise development, since 40%-85% of injected embryos
356 with phenotypes died few days or weeks post injection (compared to 6-16% for
357 control embryos). Similar phenotypes were also observed after injection of
358 morpholino antisense oligonucleotide directed against *MSANTD2* (**Figure S7**).

359 Concerning the nervous system, the Midbrain-Hindbrain Boundary (MHB), which is a
360 well-defined structure of the 24hpf stage of zebrafish development, was not well
361 formed (**Figure 6A-F**, white stars). In order to characterize nervous system
362 malformations, dextran Texas Red was injected into zebrafish brain ventricles at
363 24hpf and 30hpf (**Figure 6H-O**) (Gutzman and Sive 2009). At each stage, three
364 pictures of *MSANTD2* CRISPR/Cas9 embryos were compared to a control. All
365 *MSANTD2* CRISPR/Cas9 embryos presented brain abnormalities with neural tubes
366 misfolding particularly in the MHB region (**Figure 6H-O**, white arrowheads), as well
367 as defects in forebrain, midbrain and hindbrain inflation. These phenotypes were
368 observed from 24hpf and were still present at 30hpf. Dead cells were marked in
369 24hpf embryos with acridine orange staining (**Figure 7A-D**). We observed numerous
370 dead cells in *MSANTD2* CRISPR/Cas9 embryos compared to control embryos.

371 Hence, the cell aggregates we observed might correspond to dead cell areas (**Figure**
372 **7C-D**, regions of cell aggregates were indicated with white stars).

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

386 We studied in 24hpf *MSANTD2* CRISPR/Cas9 embryos the expression of
387 characteristic genes involved in vertebrate nervous system development (**Figure 7J-**
388 **N**). Because the MHB is a well-defined structure at the 24hpf stage of zebrafish
389 development, we characterized it with the expression of the *FGF8*, *PAX2A* and *HER5*
390 genes. *FGF8* encodes a fibroblast growth factor involved in several processes,
391 including nervous system development particularly for two brain structures, the
392 tectum and cerebellum. *FGF8* is also responsible for the maintenance of the MHB
393 together with *PAX2A* (Nakamura 2001; Chi et al. 2003). In 24hpf control embryos,
394 *FGF8* was expressed in telencephalon, dorsal diencephalon, optic stalks, otic vesicle
395 and MHB regions. In addition to MHB, *PAX2A* was expressed in optic stalks, otic
396 vesicle and hindbrain neurons at 24hpf (**Figure 7J,K,M**). The MHB is also
397 characterized by the expression of *HER5*, which is involved in multiple developmental
398 processes and particularly brain development (**Figure 7L**). Finally, as we observed
399 dead cell aggregates around the nervous system in the *MSANTD2* CRISPR/Cas9
400 embryos (**Figure 6-7**), we questioned whether they might correspond to dead neural
401 crest cells by studying the expression of *DLX2*, which marks the cranial migratory
402 neural crest cells that form the pharyngeal arches and migrate into the forebrain
403 (Akimenko et al. 1994; Yan et al. 2005; Sperber et al. 2008; Dai et al. 2013).

404 In *MSANTD2* CRISPR/Cas9 embryos, we observed alterations of *FGF8*, *PAX2A* and
405 *HER5* expression particularly in the MHB region (**Figure 7J-M**). These genes were
406 still expressed, but the marked areas were different between control and *MSANTD2*
407 CRISPR/Cas9 embryos. The expression bands were narrower for *FGF8* and *PAX2A*
408 (**Figure 7K,M**) and also less deep for *FGF8* (**Figure 7J**). For *HER5*, the staining into
409 two distinct areas was lost in *MSANTD2* CRISPR/Cas9 embryos (**Figure 7L**).
410 Moreover, the expression of *FGF8* in telencephalon and optic stalks was not
411 separated into two different zones but formed a unique and larger area, suggesting
412 defects in the definition and individualization of these structures (**Figure 7J,K**).
413 *MSANTD2* CRISPR/Cas9 embryos lacked the expression of *PAX2A* in hindbrain
414 neurons (**Figure 7M**). Finally, the expression of *DLX2* was markedly reduced in the
415 telencephalon and pharyngeal arch regions (**Figure 7N**). In conclusion, these results
416 indicated anatomical defects of the *MSANTD2* CRISPR/Cas9 embryos in the MHB
417 and telencephalon regions. Finally, the accumulation of dead cells and the
418 expression patterns of *PAX2A* and *DLX2* also suggested potential implication of
419 *MSANTD2* in neural crest cell migration, homing or differentiation into neurons
420 (visible for hindbrain neurons).

421 422 **DISCUSSION**

423 ***Harbinger* transposons have given rise to a new gene family through recurrent** 424 **and concomitant molecular domestication events in vertebrates**

425 In this work we report recurrent and concomitant molecular domestication of
426 *Harbinger* transposons in early vertebrate evolution. We have identified in jawed
427 vertebrates a new family of genes derived from *Harbinger* elements, and more
428 particularly from their *Myb-like* gene (**Figure 1-3**). Indeed, *MSANTD1*, *MSANTD2*,
429 *MSANTD3* and *MSANTD4* presented sequence similarities with the *Myb-like* domain
430 of proteins from *Harbinger* transposons (**Figure 1**). Each *MSANTD* gene is present
431 as a single copy gene at a conserved position in vertebrate genomes (**Figure 2**).
432 Hence, *MSANTD* genes are not transposons anymore but *bona fide* vertebrate
433 genes. In order to investigate whether *MSANTD* genes arose from independent
434 molecular domestication or/and sequential duplication events, sequence alignments
435 and phylogenies were constructed (**Figure 1, 3**). The sequences of all *MSANTD*
436 genes and proteins could not be aligned unambiguously together due to high
437 divergence. Indeed, the *Myb-like* proteins of different families of *Harbinger*

438 transposons present significant similarities only restricted to a short part of their Myb-
439 like domain. In contrast, *Harbinger* transposases are much more conserved. After
440 constructing separate phylogenies through Bayesian analyses for
441 MSANTD1/MSANTD2 and MSANTD3/MSANTD4/NAIF1, respectively, we observed
442 that orthologous MSANTD sequences formed monophyletic groups and that for each
443 of them the closest *Harbinger* transposon was different. However, Maximum
444 Likelihood analysis failed to support some of these preferential phylogenetic
445 relationships between *MSANTD* genes and *Harbinger* transposons. Taken together,
446 we propose that the vertebrate family of *Myb-like* genes derived from *Harbinger*
447 transposons originated from three to five independent molecular domestication
448 events. Three to four domestications probably occurred at the base of jawed
449 vertebrates about 500 Mya, and a more recent might have led to the formation of
450 *MSANTD3* at the base of sarcopterygians approx. 430 Mya.

451 A processed pseudogene resulted from a duplication by *MSANTD2* mRNA
452 retrotransposition into a “protein-coding gene desert” in simians (**Figure S1**).
453 Duplicated transcribed pseudogenes can directly regulate related functional genes by
454 transcriptional interference through the production of small interfering RNAs, or by
455 recruiting factors initially silencing the protein-coding gene transcript (Sen and Ghosh
456 2013). However, *MSANTD2P1* is not expressed in human brain, and no more
457 information is available on its expression. Its function, if any, remains to be further
458 investigated. *MSANTD2P1* insertion occurred into a “protein-coding gene desert”
459 region containing a cluster of non-coding RNA genes. Centromere and
460 pericentromeric regions are generally depleted in protein-coding genes and enriched
461 in TEs and other repetitive sequences (Schueler et al. 2001; Li et al. 2013). However,
462 the 21q23 region in which the insertion occurred is neither a centromere nor a
463 pericentromeric region. Hence, it might correspond to a cluster of lncRNAs and
464 miRNAs with some regulatory capacities.

465 *Harbinger* DNA transposons have given rise to multiple novel genes in divergent
466 organisms: at least six in vertebrates, nine in *Arabidopsis* and seven in *Drosophila*
467 (Kapitonov and Jurka 2004; Casola et al. 2007; Sinzelle et al. 2008; Liang et al.
468 2015; Duan et al. 2017; Velanis et al. 2020; Zhou et al. 2021). Hence, it seems that
469 these elements have high propensity to be recruited as new genes. The
470 characteristics of *Harbinger* transposons, with their two protein-coding open reading
471 frames (ORFs), may be an advantage. These two ORFs encode proteins with

472 domains with widespread functions. Particularly, the Myb-like domain, a DNA- and
473 protein-binding domain, could be repurposed in diverse ways for gene regulation as a
474 transcription factor and/or as a member of a protein interactome (Sinzelle et al.
475 2008). In addition, the separation of the different molecular activities, i.e. DNA
476 breaking/recombination and DNA/protein binding, in two independent ORFs is
477 uncommon for TEs. This might allow a more specific co-option by the host of a single
478 molecular activity without interference of the other.

479

480 ***Harbinger*-derived *MSANTD* genes encode potential DNA- and protein-binding** 481 **proteins**

482 We observed that the secondary structure (tri-helix motif) of the *Harbinger* Myb-like
483 protein has been conserved in the different *MSANTD* proteins, suggesting
484 conservation of the original molecular properties (**Figure 1**). Generally, the
485 SANT/myb/trihelix motives have been shown to have DNA- and protein-binding
486 capacities in multiple transcription factors (Boyer et al. 2004). The *Harbinger* Myb-like
487 protein is able to bind both the transposon DNA and the transposase protein
488 (Sinzelle et al. 2008). Thus, the *MSANTD* proteins could act as DNA- and protein-
489 interactors. Accordingly, NAIF1, like the *Harbinger* Myb-like protein, is able to bind
490 DNA and interact with the *Harbinger* transposase as well as with the transposase-
491 derived HARBI1 protein (Sinzelle et al. 2008). *MSANTD3* has been suggested to
492 work as a transcription factor that binds to DNA, where it can recruit the Polycomb
493 Repressive Complex 2 to regulate neuronal differentiation in P19 mouse cells (Gou
494 2014). Outside of vertebrates, other genes derived from *Harbinger* transposons have
495 been identified in *Arabidopsis* (Liang et al. 2015; Duan et al. 2017; Velanis et al.
496 2020; Zhou et al. 2021). *ALP1* (Antagonist of Like Heterochromatin Protein 1), its
497 paralog *HHP1* (HDA6-associated *Harbinger* transposon-derived Protein 1) and *HDP1*
498 (*Harbinger* transposon-Derived Protein 1) are derived from transposases, while *ALP2*
499 (Antagonist of Like Heterochromatin Protein 2), *HDP2* (*Harbinger* transposon-Derived
500 Protein 2), *SANT1*, *SANT2*, *SANT3*, and *SANT4* have been formed from *Harbinger*
501 *Myb-like* genes. *ALP2* and *HDP2* interact with *ALP1* and *HDP1*, respectively, and are
502 involved in chromatin modifying complexes (Liang et al. 2015; Duan et al. 2017;
503 Velanis et al. 2020; Zhou et al. 2021). *ALP1* and *ALP2* mediate Polycomb
504 Repressive Complex 2 formation (Velanis et al. 2020). *HDP1* and *HDP2* are part of a
505 histone acetyltransferase complex acting in DNA methylation through the DNA-

506 binding capacity of HDP2 (Duan et al. 2017). Similarly, HHP1, SANT1, SANT2,
507 SANT3, and SANT4 belong to a HDA6 histone deacetylase complex controlling
508 flowering time (Zhou et al. 2021).

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

514

515 ***Harbinger*-deriving genes are expressed in developing and adult vertebrate** 516 **brain**

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

525 *Harbinger*-derived genes are expressed in multiple tissues in zebrafish (**Figure 4**)
526 and human (GTEx Consortium 2013). However, as observed in zebrafish adult male
527 brain, we detected a common expression of *Harbinger*-derived genes in human brain
528 particularly during early fetal development (**Figure 4B, 5**), which might favor
529 functional interactions of their proteins in this organ in vertebrates. *Harbinger*-derived
530 genes are predominantly expressed from 8-9 pcw through the two first trimesters of
531 fetal development. Around 8-9 pcw, a process called neuronal migration starts in fetal
532 brain (Métin et al. 2008; Rahimi-Balaei et al. 2018). Neurons are formed in the
533 neuroepithelium, a neural tube layer, during embryonic development. Neuronal
534 migration corresponds to the processes by which neurons will migrate from their
535 germinal layer to all over the central nervous system, where they will establish
536 connections with other cells. As more and more neurons migrate to their final
537 localization, the different brain structures start to be formed throughout the first and
538 second trimesters of fetal development. Disturbance of neuronal migration can lead
539 to neurological disorders such as schizophrenia, autism spectrum disorders and

540 epilepsy (Fatemi 2005; Guerrini and Parrini 2010; Muraki and Tanigaki 2015; Pan et
541 al. 2019).

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

547 Together, the redundant expression in zebrafish and human brain suggests the
548 potential implication of *Harbinger*-derived genes in vertebrate nervous system
549 development, potentially in neuronal migration (**Figure 5**). This is also compatible
550 with works associating *MSANTD2* to schizophrenia and autism spectrum disorders in
551 human (Schizophrenia Working Group of the Psychiatric Genomics Consortium
552 2014; Lim et al. 2017; O'Brien et al. 2018; Zhang et al. 2020).

553

554 ***MSANTD2*, a gene involved in vertebrate nervous system development**

555 In order to better understand the biological roles of *Harbinger*-derived genes in
556 vertebrates, we have further analyzed the effects of the inactivation of the *MSANTD2*
557 gene in zebrafish. This gene has been associated to neuro-developmental diseases
558 such as autism spectrum disorders and schizophrenia in human (Schizophrenia
559 Working Group of the Psychiatric Genomics Consortium 2014; Lim et al. 2017;
560 O'Brien et al. 2018; Zhang et al. 2020). Expression analyses revealed *MSANTD2*
561 expression in brain during development in human but also in zebrafish at 24hpf
562 (**Figure 4-5**). This suggested a possible function of *MSANTD2* in vertebrate nervous
563 system development.

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

571 In *MSANTD2* CRISPR/Cas9 embryos we observed modified expression patterns for
572 the *FGF8*, *DLX2*, *PAX2A* and *HER5* genes, which are involved in vertebrate nervous
573 system development. These results revealed brain, and particularly MHB

574 organization defects. Moreover, we found accumulation of dead cells in *MSANTD2*
575 CRISPR/Cas9 embryos (**Figure 7A-D**). The altered expression of *DLX2*, a gene
576 involved in cranial migratory neural crest cell development, suggested that dead cell
577 accumulation could correspond to neural crest cells. Neural crest cells contribute to
578 multiple cell lineages, including sensory and automatic neurons, glia cells, pigment
579 cells and chondrocytes (Iulianella and Trainor 2003). In zebrafish, cranial neural crest
580 cell migration starts around 13hpf (Rocha et al. 2020). Furthermore, the expression of
581 *MSANTD2* in human brain during fetal development in a time lapse where neuronal
582 migration arises as well as the aberrant pattern of early neurons in *MSANTD2*
583 CRISPR/Cas9 zebrafish embryos might support a role of *MSANTD2* in neural crest
584 cell or neuron migration.

585 The phenotypes observed in our analysis correspond to *MSANTD2* F0 generation
586 mutants. Reproducible phenotypes were obtained with different combinations of
587 sgRNAs as well as with morpholino oligonucleotides in a gene knock-down approach.
588 Moreover, inactivation of other *Harbinger*-derived genes in zebrafish did not produce
589 similar phenotypes, indicating specificity of the phenotypes observed for *MSANTD2*.
590 Hence, in addition to the strong mortality of *MSANTD2* mutated embryos, these
591 results strongly support a role for this gene in the development of the vertebrate
592 nervous system.

593

594 **Conclusion**

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

604 Ohno proposed that whole genome duplications, generating an extensive expansion
605 of gene repertoires, are major events giving rise to massive innovations and
606 important evolutionary transitions (Ohno 1999). Accordingly, two events of genome
607 duplications have taken place at the base of vertebrates (Dehal and Boore 2005).

608 However, new gene formation by duplication is not the unique mechanism allowing
609 the apparition of major novelties. *SYNCYTIN* genes, involved in placenta formation in
610 mammals, as well as *RAG* genes, implicated in the adaptive immune system in
611 vertebrates, testify of the role of TE-derived novel genes in organismal innovation.

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

620

621 **MATERIALS AND METHODS**

622 **Zebrafish maintenance**

623 Zebrafish of the strain AB/TU were raised according to standard procedures [PRECI,
624 SFR Biosciences (UAR3444/CNRS, US8/INSERM, ENS de Lyon, UCBL)]. Embryos
625 were raised at 28°C. Developmental stages were expressed in hours post-fertilization
626 (hpf) or days post-fertilization (dpf) based on morphological criteria (Kimmel et al.
627 1995). The Tg(elavl3:GCaMP6s) transgenic line, containing a modified GCaMP
628 (GCaMP is a genetically-encoded calcium indicator) known as GCaMP3 under elavl3
629 regulatory region (ZFIN ID: ZDB-TGCONSTRUCT-180326-1), were also raised
630 according to the same procedures (Park et al. 2000; Panier et al. 2013).

631

632 ***In situ* hybridization (ISH)**

633 ISH probes for *MSANTD2* were cloned from wild-type (WT) zebrafish cDNA by PCR
634 using the GoTaq polymerase (Promega). *PAX2A* and *HER5* probes were given by
635 Dr. Sebastian Dworkin lab, La Trobe University, Melbourne, Australia; *FGF8* and
636 *DLX2* probes by Dr. Dominique Baas, MeLiS, UCBL, France.

637 Zebrafish embryos were collected, removed from their chorion, sorted and fixed in
638 paraformaldehyde (PFA) 4%, dehydrated in methanol and stored at - 20°C.

639 ISH was performed following the Thisse Lab protocol (Thisse and Thisse 2008;
640 Thisse and Thisse 2014). Embryos were rehydrated and washed in phosphate

641 buffered saline (PBS) – Tween (PBT) solution. They were permeabilized with
642 proteinase K and fixed in PFA. Each embryo was incubated with probes overnight at
643 65°C in hybridization mix supplemented with 5% Dextran Sulfate (Millipore). Non-
644 hybridized probes were removed with several washes in formamide and Saline-
645 Sodium-Citrate (SSC) solutions. Embryos were incubated overnight at 4°C with α -
646 DIG antibodies (Roche). Non-fixed antibodies were removed with PBT washes.
647 Probes were revealed with NBT-BCIP (Roche). Embryos were fixed in PFA. After
648 removing of the background with ethanol bath, embryos were stored in glycerol 80%
649 at 4°C. Pictures were taken under Leica stereomicroscope and Keyence VHX-7000
650 microscope.

651

652 **qPCR**

653 Pools of 3-5 zebrafish adults and 15-20 embryos were used for RNA extraction.
654 RNAs were extracted with Trizol according to the Bio-Rad company protocol and
655 treated with DNaseI. Reverse transcription was performed using the RevertAid First
656 Strand cDNA Synthesis Kit (Thermo Scientific). The following specific primers were
657 designed: for *NAIF1* TGAATCACTTTAACGCGGGC, CCGTCTTCAGATCCGACCAT;
658 for *HARBI1* CGCTGCGTTTCTAACGTAC, AGAGTCATCCGCATTGGGAG; for
659 *MSANTD1* CAAACCTCTCATCGTCTGGC, AGGCCGTCATCCTCATCATT; for
660 *MSANTD2* AGACCCGAGTTCTTCAGATACGAC,
661 GAGAGAAGTCCGTCCACGTTTG; for *MSANTD4* TCAAGATGGAGGACGACGAG,
662 GGGAGGATGGAGGGAAAACA. qPCR was performed using SYBR Green following
663 the Bio-Rad protocol. 18S housekeeping ribosomal RNA gene
664 (TCGCTAGTTGGCATCGTTTATG, CGGAGGTTCGAAGACGATCA) was used to
665 normalize gene expression. Results were analyzed with the Δ Ct method (Schmittgen
666 and Livak 2008).

667

668 **Morpholino knockdown**

669 Two non-overlapping morpholino antisense oligonucleotides targeting the 5'-UTR of
670 *MSANTD2* (GCCATCTTGCTTCTGTTGCTAAGGG,
671 CAGACACGACTGACGGCTTCTTATG) and a control mismatched morpholino
672 (CACACACCAGTGACGCCTTGTATG) were purchased from Gene Tools and
673 injected from 0.2M to 3M into one-cell embryos (Nasevicius and Ekker 2000).

674 Morphological and phenotypic observations were performed at 1dpf under Zeiss Axio
675 Zoom microscope.

676

677 **CRISPR/Cas9**

678 For each gene four non-overlapping single guide RNAs (sgRNAs) were purchased
679 from Synthego (**Table S2**). The sequences of the sgRNAs were selected from Wu et
680 al. 2018 (Wu et al. 2018). The Cas9-GFP protein was purchased from TacGene. A
681 mix of four sgRNAs (20 to 30 μ M in total) and Cas9-GFP protein (5 to 15 μ M) was
682 injected into WT embryos at the one-cell stage. For *MSANTD2*, sgRNA 1 and sgRNA
683 2 were located in the first exon and sgRNA 3 and sgRNA 4 in the third and fourth
684 exons, respectively. For each gene knock-out, the experiment was performed at least
685 in duplicate. Scrambled (random sequence) sgRNAs were used as a negative control
686 and sgRNAs targeting the tyrosinase (*TYR*) gene as a positive control (its inactivation
687 led to individuals without melanic pigmentation). Embryo survival and phenotypic
688 observations were monitored from 6hpf under Leica stereomicroscope and Zeiss
689 Axio Zoom microscope.

690

691 **Brain ventricle imaging**

692 Zebrafish brain ventricle injection was performed according to the protocol developed
693 by Gutzman and Size 2009 (Gutzman and Sive 2009). Briefly, embryos were
694 anesthetized with Tricaine (Sigma). Micro-injection was performed in hindbrain
695 ventricle with 1-10nl of dextran Texas Red (5% in 0,2mol/l KCl, Invitrogen). 15 to
696 30min after injection, images were taken with transmitted and fluorescent lights under
697 a Zeiss AxioZoom Microscope.

698

699 **Acridine orange staining**

700 Embryos were dechorionated and stained with 10 μ g/mL acridine orange solution for
701 30 min. Then, embryos were washed three times in E3 medium. Images were taken
702 with transmitted and fluorescent lights under a Zeiss AxioZoom Microscope.

703

704 **DNA extraction, PCR amplification, NGS sequencing and sequencing data 705 analyses**

706 In order to search for mutations after application of the CRISPR/Cas9 protocol,
707 injected embryos were collected at 24hpf and five DNA extraction replicates were

708 conducted starting from a single embryo for the four conditions (4sgRNAs, sgRNA1-
709 4, sgRNA4 and scrambled). Lysis of embryos was performed in lysis buffer (10mM
710 Tris-HCl pH8 - 2 mM EDTA pH8 – 0.2% Triton x-100) with 250 µg/µl proteinase K
711 (Invitrogen) 12 hours at 55°C, followed by proteinase K inactivation of 10 minutes at
712 95°C. Three fragments of the *MSANTD2* gene (exon 1, exon 3 and exon 4) were
713 amplified by PCR. PCR reactions were performed in 25 µl using the GoTaq G2 DNA
714 polymerase kit (Promega), 2 µl of DNA extract and 0.5 µM of each primer set with the
715 following PCR program: 2 min at 94°C, 35 cycles at 94°C 30 s, 60°C 30 s and 72°C
716 30 s, with a final extension step at 72°C for 5 min. For each condition and for each
717 exon, five PCR tubes (each PCR corresponding to one embryo DNA amplification)
718 were pooled. PCR product purification was carried out according to manufacturer's
719 recommendations (Nucleospin Gel and PCR Clean-up, Macherey Nagel) and eluted
720 in 30 µl of elution buffer (NE buffer). For each condition, equimolar amounts of the
721 three purified amplicons were used to create a barcoded library with an input of 50ng
722 using the NEBNext Ultra II DNA Library Prep Kit protocol for Illumina. Quantitation
723 and quality assessment of each library was performed on a 4150 TapeStation
724 analyzer using the High Sensitivity D5000 ScreenTape kit (Agilent Technologies).
725 Libraries were mixed at the same equimolar proportions, spiked with approximately
726 5% PhiX control and sequenced with the Illumina MiSeq sequencer using the Nano
727 Kit v2 reagent (pair-end reads, R1 and R2 read lengths, 260bp and 259bp
728 respectively). More than 800K reads were obtained and analyzed using the Galaxy
729 platform (Afgan et al. 2018) using the FastQC, Cutadapt, Bowtie2 and Sort tools to
730 assess the quality of reads, remove adapter sequences, map reads against reference
731 and store aligned sequences, respectively.

732

733 **Transposable element and gene sequence *in silico* analyses**

734 TE-derived genes were identified through sequence similarity with TE sequences
735 from the Repbase database (www.girinst.org) and from annotation of various
736 sequenced vertebrate genomes (Chalopin et al. 2015) using blastp, blastn and
737 tblastn (Altschul et al. 1990). Additional *Harbinger* Myb-like protein sequences were
738 recovered through blast analysis of the NCBI Genomes (RefSeq Genomes) database
739 (www.ncbi.nlm.nih.gov) using *MSANTD* and *Harbinger* transposon sequences as
740 queries. NCBI, Ensembl, Censor (www.girinst.org) and Genomicus (Muffato et al.
741 2010) were used to determine the copy number, sequence alignments, phylogeny

742 and synteny of TE-derived genes. Conserved proteins motifs were detected with the
743 NCBI Conserved Domain Search (Marchler-Bauer et al. 2011), InterPro (Apweiler et
744 al. 2000) and PROSITE (Sigrist et al. 2002). Genes and open reading frames were
745 predicted with Augustus (Stanke and Morgenstern 2005) and ORFfinder
746 (www.ncbi.nlm.nih.gov/orffinder). NPS-PRABI (Combet et al. 2000) and Jpred4
747 (Drozdetskiy et al. 2015) were used to predict the secondary structure of proteins.
748 For positive selection tests, the protein-coding sequence of genes from different
749 species were collected from the Ensembl database, aligned as proteins using
750 MUSCLE (Edgar 2004) and then converted back into a nucleic sequence alignment.
751 A phylogenetic tree was then built with the PhyML package (see below).
752 Positive/negative selection tests were performed using CODEML (Yang 2007). The
753 tests were run based on an alignment of coding sequences from spotted gar,
754 zebrafish, tetraodon, stickleback, platyfish, coelacanth, chinese softshell turtle,
755 mouse, macaca, marmoset, human, chimpanzee and chicken.
756 For phylogenetic analysis, nucleotide and amino-acid sequences were aligned with
757 MAFFT (Kato et al. 2002). Phylogenetic trees were built using maximum likelihood
758 with PhyML (Guindon and Gascuel 2003) and with MrBayes (Huelsenbeck and
759 Ronquist 2001) using a mixed model (estimated by the Prottest-3 software (Darriba et
760 al. 2011) and 500,000 generations of Bayesian inferences.
761 Gene accessions numbers are HGNC:33741, HGNC:26266, HGNC:23370,
762 HGNC:29383, HGNC:25446 and HGNC:26522 for *MSANTD1*, *MSANTD2*,
763 *MSANTD3*, *MSANTD4*, *NAIF1* and *HARBI1*, respectively. Transposon sequences
764 and alignments are available upon request.

765

766 **ACKNOWLEDGEMENTS**

767 This work was supported by grants from the Agence Nationale de la Recherche
768 (ANR) and from the Ecole Normale Supérieure de Lyon. We acknowledge the
769 contribution of the PRECI fish facility of the SFR Biosciences (UAR3444/CNRS,
770 US8/INSERM, ENS de Lyon, UCBL) and of the IGFL's PSI Sequencing platform. We
771 are grateful to Marilyne Malbouyres, Sandrine Bretaud and Florence Ruggiero (Matrix
772 biology and pathology team, Institut de Génomique Fonctionnelle de Lyon) for their
773 help in Crispr-Cas9 methodology. EE thanks the "Fondation pour la Recherche
774 Médicale" (FRM) for financial support through an end of PhD program fellowship.

775

776 **AUTHOR CONTRIBUTIONS**

777 Experiments were designed by EE and JNV and performed by EE, manuscript was
778 drafted by EE and amended by JNV, the project was supervised by JNV and co-
779 supervised by DB, MN and ZHT.

780

781 **FIGURE LEGENDS**

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

790

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

797

798 **Figure 3:** Phylogenetic relationships between MSANTD proteins and their closest Myb-like
799 proteins from *Harbinger* transposons. Trees were constructed using the Bayesian method
800 (Huelsenbeck and Ronquist 2001). Only branch support values higher than 50% are shown.

801 AG: *Anopheles gambiae*, AM: *Alligator mississippiensis*, BF: *Branchiostoma floridae*, CG:
802 *Crassostrea gigas*, CM: *Chelonia mydas*, CP: *Chrysemys picta*, DR: *Danio rerio*, EL: *Esox lucius*,
803 GA: *Gasterosteus aculeatus*, GG: *Gallus gallus*, HS: *Homo sapiens*, LC: *Latimeria chalumnae*,
804 MM: *Mus musculus*, NV: *Nematostella vectensis*, OL: *Oryzias latipes*, ON: *Oreochromis niloticus*,
805 OS: *Oryza sativa*, PG: *Puccinia graminis*, PS: *Pelodiscus sinensis*, PSt: *Puccinia striiformis*, SS:
806 *Salmon salar*, TF: *Takifugu flavidus*, VV: *Vitis vinifera*, XT: *Xenopus tropicalis*). Silhouette images
807 from phylopic.org.

808

809 **Figure 4:** Expression of *Harbinger*-derived genes in zebrafish embryos and adults. **(A)** Relative
810 gene expression determined by qPCR compared to the 18S housekeeping gene during
811 embryonic development (from 1hpf to 3dpf). The midblastula transition (MBT) is shown with an

812 arrow. **(B)** Relative gene expression determined by qPCR compared to the 18S housekeeping
813 gene in female (F) and male (M) adult tissues. **(C)** Whole-mount *in situ* hybridization of
814 *MSANTD2* in zebrafish embryos from 1.25hpf to 24hpf using a sense probe as a control. The
815 head or future head region is indicated with “h”. Mhb: midbrain-hindbrain boundary, fb: forebrain,
816 mb: midbrain, hb: hindbrain. Scale bars: 200µm.

817

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

825

826 **Figure 6:** *MSANTD2* CRISPR/Cas9 embryo phenotypes. Embryos injected with control **(A-C,**
827 **CTRL)** or four *MSANTD2*-directed sgRNAs **(D-F, MSANTD2)**. **A/D** and **B/E** present lateral views
828 of whole embryos and of the head region, respectively. **C/F** show dorsal views of the embryo
829 head region. Embryos injected with four sgRNAs showed developmental delays as well as tail
830 and nervous system malformations compared to control embryos at 24hpf **(A-F)**. Tail curvature
831 and not well-defined somites are shown with black arrowheads **(D)**. Defect in neural tube folding
832 and cell aggregates around the nervous system are indicated with white arrowheads and black
833 arrows, respectively **(E-F)**. Not well-formed MHB are shown with white stars (*) **(E-F)**.

834 **(G)** Proportions of F0 zebrafish phenotypes. Embryos were injected with sgRNA 1+2+3+4,
835 sgRNA 1+4 and sgRNA 4 targeting *MSANTD2*, four scrambled sgRNAs or non-injected (WT),
836 and were scored for phenotypes at 24hpf. Percentages with strong (red), intermediate (orange) or
837 no phenotypes (gray–WT) are shown. Strong phenotypes: developmental delays, tail
838 malformations, nervous system malformations, aggregates; intermediate phenotypes:
839 developmental delays, nervous system malformations, no or few aggregates, no tail
840 malformation.

841 **(H-O):** Dextran Texas Red injection in brain ventricles of 24hpf **(H-K)** or 30hpf **(L-O)** embryos
842 injected with control **(H, L, CTRL)** or four *MSANTD2*-directed sgRNAs **(I-K, M-O, MSANTD2)**
843 (dorsal view). At each stage, three pictures of *MSANTD2* CRISPR/Cas9 embryos were compared
844 to a control picture in order to represent phenotype variability. All cases illustrate neural tubes
845 misfolding (shown with white arrowheads) particularly in the MHB region. We also observed
846 smaller red fluorescent areas in the forebrain and midbrain regions, indicating reduction of these
847 ventricles. MHB: midbrain-hindbrain boundary; F: forebrain, M: midbrain; H: hindbrain. Scale bar
848 200µm.

849

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

869

870 **SUPPLEMENTARY FIGURE LEGENDS**

871 **Figure S1:** *MSANTD2P1* is a transcribed processed pseudogene originating from the duplication
872 of *MSANTD2* in a “protein-coding gene desert”. **(A)** Multiple alignment of *MSANTD2* proteins and
873 *MSANTD2P1* translated sequences. *MSANTD2P1* translated sequences were obtained by
874 assembling the results of *MSANTD2P1* nucleic sequences “blasted” against *MSANTD2* protein
875 sequences. Stop codons are marked with black stars. **(B)** Phylogenetic relationships between
876 *MSANTD2* proteins and *MSANTD2P1* translated sequences. Bootstrap values are shown. The
877 tree was constructed using the Maximum Likelihood method (Guindon and Gascuel 2003). **(C)**
878 Synteny analysis of the *MSANTD2P1*-containing region in simians. Species names and genomic
879 locations are shown on the right. **(D)** Histogram representing the distribution of the length of
880 human intergenic regions (in bp) between two consecutive protein-coding genes in log scale.
881 16.000 regions were studied. The green line indicates the length of the 21q23 region where
882 *MSANTD2P1* is located. The dashed grey line shows the limit of 200,000 base pair-long region.

883

884 **Figure S2:** Phylogenetic relationships between *MSANTD* proteins and their closest Myb-like
885 proteins from *Harbinger* transposons. The trees were constructed using the Maximum Likelihood
886 method (Guindon and Gascuel 2003). Bootstrap values higher than 50% are shown.

887 (AG: *Anopheles gambiae*, AM: *Alligator mississippiensis*, BF: *Branchiostoma floridae*, CG:
888 *Crassostrea gigas*, CM: *Chelonia mydas*, CP: *Chrysemys picta*, DR: *Danio rerio*, EL: *Esox lucius*,
889 GA: *Gasterosteus aculeatus*, GG: *Gallus gallus*, HS: *Homo sapiens*, LC: *Latimeria chalumnae*,
890 MM: *Mus musculus*, NV: *Nematostella vectensis*, OL: *Oryzias latipes*, ON: *Oreochromis niloticus*,
891 OS: *Oryza sativa*, PG: *Puccinia graminis*, PS: *Pelodiscus sinensis*, PST: *Puccinia striiformis*, SS:
892 *Salmon salar*, TF: *Takifugu flavidus*, VV: *Vitis vinifera*, XT: *Xenopus tropicalis*). Silhouette images
893 from phylopic.org.

894

895 **Figure S3:** Proportions of F0 zebrafish phenotypes using different sgRNAs in CRISPR-Cas9
896 experiments directed against *MSANTD2*. Embryos were injected with sgRNA 1+2, sgRNA 1+3,
897 sgRNA 2+3, sgRNA 2+4, sgRNA 3+4, sgRNA 1, sgRNA 2 and sgRNA 3 targeting *MSANTD2*,
898 four scrambled sgRNAs or non-injected (WT) and scored for phenotypes at 24hpf. The
899 percentages of embryos with strong (red), intermediate (orange) or no phenotypes (gray-WT) are
900 presented. Strong phenotypes: developmental delays, tail malformations, nervous system
901 malformations, aggregates; intermediate phenotypes: developmental delays, nervous system
902 malformations, no or few aggregates, no tail malformation.

903

904

905 **Figure S4:** Read coverage at each sgRNA locus for the four conditions of zebrafish embryo
906 injection in CRISPR-Cas9 experiments against *MSANTD2* (scrambled, *MSANTD2* 4sgRNAs,
907 *MSANTD2* sgRNA 1+4, *MSANTD2* sgRNA 4). Peak height represents the read coverage at each
908 position of the sequence, height decrease indicates nucleotide deletion. Colored peaks show
909 single nucleotide polymorphisms, whereas gray peaks indicate no nucleotide substitution. sgRNA
910 1 and sgRNA 2 are located in exon 1 and sgRNA 3 and sgRNA 4 in exon 3 and 4, respectively.

911

912 **Figure S5:** Alignment of the most frequent reads at each sgRNA locus compared to the reference
913 sequence (ref_exon) in CRISPR-Cas9 experiments against zebrafish *MSANTD2*.

914

915 **Figure S6:** *MSANTD2* CRISPR/Cas9 embryos at 24hpf, injected with *MSANTD2* sgRNA 1 and
916 sgRNA 4 together (**A, B, C**) or sgRNA 4 alone (**D, E, F**). Scale bar 200µm. **A/D** and **B/E** show
917 lateral views of whole embryos and of the head region, respectively. **C/F** present dorsal views of
918 the head region of the embryos. Embryos showed developmental delays as well as tail and
919 nervous system malformations compared to control embryos (see **Figure 6**) at 24hpf (**A-F**). Tail
920 curvature and not well-defined somites are shown with black arrowheads (**D**). Defaults in neural
921 tubes folding and cell aggregates around the nervous system are indicated with white arrowheads
922 and black arrows, respectively (**E-F**).

923

924 **Figure S7:** Phenotypes after morpholino-mediated knock-down of the *MSANTD2* gene in
 925 zebrafish embryos. 24hpf embryos were injected with control (CTRL) (**A, B, G, H**) or *MSANTD2*
 926 ATG-blocking (*MSANTD2*) morpholinos MO1 and MO2 (**C-F, I-L**) at 1mM (**A-F**) or 3mM (**G-L**).
 927 Scale bar 200 μ m. **A/C/E/G/I/K** and **B/D/F/H/J/L** present lateral views of whole embryos or dorsal
 928 views of the head region, respectively. Embryos injected with MO1 at 1mM and 3mM as well as
 929 MO2 at 3mM showed developmental delays as well as tail and nervous system malformations
 930 compared to control embryos at 24hpf. Tail curvature and not well-defined somites are shown
 931 with black arrowheads (**C, I, K**). Defaults in neural tube folding and cell aggregates around the
 932 nervous system are indicated with white arrowheads and black arrows, respectively (**C, D, J, L**).
 933 Fewer aggregates were observed compared to *MSANTD2* CRISPR/Cas9 embryos (**Figure 6**).
 934

935 TABLES

936 **Table S1:** Selection tests for vertebrate *Harbinger*-derived genes. Estimation of the non-
 937 synonymous to synonymous rate ratio (ω , dN/dS ratio) of *Harbinger*-derived protein-coding genes
 938 under neutral or nearly neutral models between the following vertebrate species : *Homo sapiens*,
 939 *Pan troglodytes*, *Macaca fascicularis*, *Callithrix jacchus*, *Mus musculus*, *Gallus gallus*, *Latimeria*
 940 *chalumnae*, *Lepisosteus oculatus*, *Gasterosteus aculeatus*, *Tetraodon nigroviridis*, *Chrysemys*
 941 *picta bellii*, *Xiphophorus maculatus* and *Danio rerio*. The neutral model calculates the non-
 942 synonymous to synonymous rate ratio (ω_0). The nearly neutral model provides the percentage of
 943 sites with non-synonymous to synonymous rate ratio of 1 (ω_0) and the percentage of sites with
 944 non-synonymous to synonymous rate ratio smaller than 1 (ω_1).

Gene	Neutral model	Nearly neutral model
<i>MSANTD1</i>	$\omega_0 = 0,047$	$\omega_0 = 1 - 5\%$ of sites $\omega_1 = 0,4734 - 95\%$ of sites
<i>MSANTD2</i>	$\omega_0 = 0,0526$	$\omega_0 = 1 - 7\%$ of sites $\omega_1 = 0,04201 - 93\%$ of sites
<i>MSANTD3</i>	$\omega_0 = 0,02888$	$\omega_0 = 1 - 5\%$ of sites $\omega_1 = 0,01597 - 95\%$ of sites
<i>MSANTD4</i>	$\omega_0 = 0,04948$	$\omega_0 = 1 - 6\%$ of sites $\omega_1 = 0,04482 - 94\%$ of sites
<i>NAIF1</i>	$\omega_0 = 0,01990$	$\omega_0 = 1 - 8\%$ of sites $\omega_1 = 0,01971 - 92\%$ of sites

<i>HARBI1</i>	$\omega_0 = 0,05859$	$\omega_0 = 1 - 4\%$ of sites $\omega_1 = 0,05596 - 96\%$ of sites
<i>MSANTD2P1</i>	$\omega_0 = 0,68617$	$\omega_0 = 1 - 80\%$ of sites $\omega_1 = 0,20994 - 20\%$ of sites

945

946 **Table S2:** sgRNA sequences used for CRISPR/Cas9 gene inactivation in zebrafish.

Name	Sequence
MSANTD2_sgRNA1	GGAGAACGCUCAGCGUUACU
MSANTD2_sgRNA2	GUGUUUUCUGGCAAGGCUC
MSANTD2_sgRNA3	GAACUCGGGUCUUCGGAAGC
MSANTD2_sgRNA4	GGAGCACUCCAAACGUGGA
scrambled_sgRNA1	GGCAGGCAAAGAAUCCCUGCC
scrambled_sgRNA2	GGUACAGUGGACCUCGGUGUC
scrambled_sgRNA3	GGCUUCAUACAAUAGACGAUG
scrambled_sgRNA4	GGUCGUUUUGCAGUAGGAUCG
TYR_sgRNA1	GGACUGGAGGACUUCUGGGG
TYR_sgRNA2	GGAUGCAUUAUUACGUGUCC
TYR_sgRNA3	GGAAAGUUACAACCUCGCG
TYR_sgRNA4	GGUAGUGUGUGCGGGGCGGC

947

948

949 REFERENCES

- 950 Afgan E, Baker D, Batut B, van den Beek M, Bouvier D, Cech M, Chilton J, Clements
951 D, Coraor N, Grüning BA, et al. 2018. The Galaxy platform for accessible,
952 reproducible and collaborative biomedical analyses: 2018 update. *Nucleic
953 Acids Res* 46:W537–W544.
- 954 Akimenko MA, Ekker M, Wegner J, Lin W, Westerfield M. 1994. Combinatorial
955 expression of three zebrafish genes related to distal-less: part of a homeobox
956 gene code for the head. *J Neurosci* 14:3475–3486.

- 957 Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment
958 search tool. *J Mol Biol* 215:403–410.
- 959 Alzohairy AM, Gyulai G, Jansen RK, Bahieldin A. 2013. Transposable elements
960 domesticated and neofunctionalized by eukaryotic genomes. *Plasmid* 69:1–
961 15.
- 962 Apweiler R, Attwood TK, Bairoch A, Bateman A, Birney E, Biswas M, Bucher P,
963 Cerutti L, Corpet F, Croning MD, et al. 2000. InterPro--an integrated
964 documentation resource for protein families, domains and functional sites.
965 *Bioinformatics* 16:1145–1150.
- 966 Báez-Mendoza R, Schultz W. 2013. The role of the striatum in social behavior. *Front*
967 *Neurosci* 7:233.
- 968 Boyer LA, Latek RR, Peterson CL. 2004. The SANT domain: a unique histone-tail-
969 binding module? *Nat Rev Mol Cell Biol* 5:158–163.
- 970 Brandt J, Schrauth S, Veith A-M, Froschauer A, Haneke T, Schultheis C, Gessler M,
971 Leimeister C, Volff J-N. 2005. Transposable elements as a source of genetic
972 innovation: expression and evolution of a family of retrotransposon-derived
973 neogenes in mammals. *Gene* 345:101–111.
- 974 Britten R. 2006. Transposable elements have contributed to thousands of human
975 proteins. *Proc Natl Acad Sci U S A* 103:1798–1803.
- 976 Casola C, Lawing AM, Betrán E, Feschotte C. 2007. PIF-like transposons are
977 common in drosophila and have been repeatedly domesticated to generate
978 new host genes. *Mol Biol Evol* 24:1872–1888.
- 979 Chalopin D, Naville M, Plard F, Galiana D, Volff J-N. 2015. Comparative analysis of
980 transposable elements highlights mobilome diversity and evolution in
981 vertebrates. *Genome Biol Evol* 7:567–580.
- 982 Chi CL, Martinez S, Wurst W, Martin GR. 2003. The isthmic organizer signal FGF8 is
983 required for cell survival in the prospective midbrain and cerebellum.
984 *Development* 130:2633–2644.
- 985 Chuong EB, Elde NC, Feschotte C. 2017. Regulatory activities of transposable
986 elements: from conflicts to benefits. *Nat Rev Genet* 18:71–86.
- 987 Combet C, Blanchet C, Geourjon C, Deléage G. 2000. NPS@: network protein
988 sequence analysis. *Trends Biochem Sci* 25:147–150.
- 989 Dai J, Kuang Y, Fang B, Gong H, Lu S, Mou Z, Sun H, Dong Y, Lu J, Zhang W, et al.
990 2013. The effect of overexpression of Dlx2 on the migration, proliferation and
991 osteogenic differentiation of cranial neural crest stem cells. *Biomaterials*
992 34:1898–1910.
- 993 Darriba D, Taboada GL, Doallo R, Posada D. 2011. ProtTest 3: fast selection of
994 best-fit models of protein evolution. *Bioinformatics* 27:1164–1165.
- 995 Dehal P, Boore JL. 2005. Two rounds of whole genome duplication in the ancestral
996 vertebrate. *PLoS Biol* 3:e314.
- 997 Doolittle WF, Sapienza C. 1980. Selfish genes, the phenotype paradigm and
998 genome evolution. *Nature* 284:601–603.

- 999 Drozdetskiy A, Cole C, Procter J, Barton GJ. 2015. JPred4: a protein secondary
1000 structure prediction server. *Nucleic Acids Res* 43:W389-394.
- 1001 Duan C-G, Wang X, Xie S, Pan L, Miki D, Tang K, Hsu C-C, Lei M, Zhong Y, Hou Y-
1002 J, et al. 2017. A pair of transposon-derived proteins function in a histone
1003 acetyltransferase complex for active DNA demethylation. *Cell Res* 27:226-
1004 240.
- 1005 Dupressoir A, Vernochet C, Harper F, Guégan J, Dessen P, Pierron G, Heidmann T.
1006 2011. A pair of co-opted retroviral envelope syncytin genes is required for
1007 formation of the two-layered murine placental syncytiotrophoblast. *Proc Natl*
1008 *Acad Sci U S A* 108:E1164-1173.
- 1009 Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and
1010 high throughput. *Nucleic Acids Res* 32:1792-1797.
- 1011 Etchegaray E, Naville M, Volff J-N, Haftek-Terreau Z. 2021. Transposable element-
1012 derived sequences in vertebrate development. *Mob DNA* 12:1.
- 1013 Fatemi SH. 2005. Reelin glycoprotein: structure, biology and roles in health and
1014 disease. *Mol Psychiatry* 10:251-257.
- 1015 Fu Y, Cao F. 2015. MicroRNA-125a-5p regulates cancer cell proliferation and
1016 migration through NAIF1 in prostate carcinoma. *Onco Targets Ther* 8:3827-
1017 3835.
- 1018 Gou Y. 2014. MSANTD3, a noval transcription factor, recruits PRC2 complex to
1019 regulate neuron differentiation in mouse P19 cells. Available from:
1020 <https://scholarship.rice.edu/handle/1911/87825>
- 1021 Grillner S, Robertson B, Stephenson-Jones M. 2013. The evolutionary origin of the
1022 vertebrate basal ganglia and its role in action selection. *J Physiol* 591:5425-
1023 5431.
- 1024 GTEx Consortium. 2013. The Genotype-Tissue Expression (GTEx) project. *Nat*
1025 *Genet* 45:580-585.
- 1026 Guerrini R, Parrini E. 2010. Neuronal migration disorders. *Neurobiol Dis* 38:154-166.
- 1027 Guindon S, Gascuel O. 2003. A simple, fast, and accurate algorithm to estimate
1028 large phylogenies by maximum likelihood. *Syst Biol* 52:696-704.
- 1029 Gutzman JH, Sive H. 2009. Zebrafish brain ventricle injection. *J Vis Exp*:1218.
- 1030 Hancock CN, Zhang F, Wessler SR. 2010. Transposition of the Tourist-MITE mPing
1031 in yeast: an assay that retains key features of catalysis by the class 2
1032 PIF/Harbinger superfamily. *Mob DNA* 1:5.
- 1033 Haniffa M, Taylor D, Linnarsson S, Aronow BJ, Bader GD, Barker RA, Camara PG,
1034 Camp JG, Chédotal A, Copp A, et al. 2021. A roadmap for the Human
1035 Developmental Cell Atlas. *Nature* 597:196-205.
- 1036 Hon C-C, Ramilowski JA, Harshbarger J, Bertin N, Rackham OJL, Gough J,
1037 Denisenko E, Schmeier S, Poulsen TM, Severin J, et al. 2017. An atlas of
1038 human long non-coding RNAs with accurate 5' ends. *Nature* 543:199-204.
- 1039 Huelsenbeck JP, Ronquist F. 2001. MRBAYES: Bayesian inference of phylogenetic

- 1040 trees. *Bioinformatics* 17:754–755.
- 1041 Iulianella A, Trainor PA. 2003. Hox gene control of neural crest cell, pharyngeal arch
1042 and craniofacial patterning">Hox gene control of neural crest cell, pharyngeal
1043 arch and craniofacial patterning. In: *Advances in Developmental Biology and*
1044 *Biochemistry*. Vol. 13. Murine Homeobox Gene Control of Embryonic
1045 Patterning and Organogenesis. Elsevier. p. 155–206. Available from:
1046 <https://www.sciencedirect.com/science/article/pii/S1569179903130067>
- 1047 Kapitonov VV, Jurka J. 2004. Harbinger transposons and an ancient HARBI1 gene
1048 derived from a transposase. *DNA Cell Biol* 23:311–324.
- 1049 Kapitonov VV, Koonin EV. 2015. Evolution of the RAG1-RAG2 locus: both proteins
1050 came from the same transposon. *Biol Direct* 10:20.
- 1051 Katoh K, Misawa K, Kuma K, Miyata T. 2002. MAFFT: a novel method for rapid
1052 multiple sequence alignment based on fast Fourier transform. *Nucleic Acids*
1053 *Res* 30:3059–3066.
- 1054 Kidwell MG, Lisch DR. 2000. Transposable elements and host genome evolution.
1055 *Trends Ecol Evol* 15:95–99.
- 1056 Kimmel CB, Ballard WW, Kimmel SR, Ullmann B, Schilling TF. 1995. Stages of
1057 embryonic development of the zebrafish. *Dev Dyn* 203:253–310.
- 1058 Kong D, Zhang Z. 2018. NAIF1 suppresses osteosarcoma progression and is
1059 regulated by miR-128. *Cell Biochem Funct* 36:443–449.
- 1060 Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar
1061 K, Doyle M, FitzHugh W, et al. 2001. Initial sequencing and analysis of the
1062 human genome. *Nature* 409.
- 1063 Li B, Choulet F, Heng Y, Hao W, Paux E, Liu Z, Yue W, Jin W, Feuillet C, Zhang X.
1064 2013. Wheat centromeric retrotransposons: the new ones take a major role in
1065 centromeric structure. *Plant J* 73:952–965.
- 1066 Liang SC, Hartwig B, Perera P, Mora-García S, de Leau E, Thornton H, de Lima
1067 Alves F, de Alves FL, Rappsilber J, Rapsilber J, et al. 2015. Kicking against
1068 the PRCs - A Domesticated Transposase Antagonises Silencing Mediated by
1069 Polycomb Group Proteins and Is an Accessory Component of Polycomb
1070 Repressive Complex 2. *PLoS Genet* 11:e1005660.
- 1071 Lim ET, Uddin M, De Rubeis S, Chan Y, Kamumbu AS, Zhang X, D’Gama AM, Kim
1072 SN, Hill RS, Goldberg AP, et al. 2017. Rates, distribution and implications of
1073 postzygotic mosaic mutations in autism spectrum disorder. *Nat Neurosci*
1074 20:1217–1224.
- 1075 Lu J, Zhu M, Ahlberg PE, Qiao T, Zhu Y, Zhao W, Jia L. 2016. A Devonian predatory
1076 fish provides insights into the early evolution of modern sarcopterygians. *Sci*
1077 *Adv* 2:e1600154.
- 1078 Luo Q, Zhao M, Zhong J, Ma Y, Deng G, Liu J, Wang J, Yuan X, Huang C. 2011.
1079 NAIF1 is down-regulated in gastric cancer and promotes apoptosis through
1080 the caspase-9 pathway in human MKN45 cells. *Oncol Rep* 25:1117–1123.
- 1081 Lv B, Shi T, Wang X, Song Q, Zhang Y, Shen Y, Ma D, Lou Y. 2006.

- 1082 Overexpression of the novel human gene, nuclear apoptosis-inducing factor
1083 1, induces apoptosis. *Int J Biochem Cell Biol* 38:671–683.
- 1084 Mallet F, Bouton O, Prudhomme S, Cheynet V, Oriol G, Bonnaud B, Lucotte G,
1085 Duret L, Mandrand B. 2004. The endogenous retroviral locus ERVWE1 is a
1086 bona fide gene involved in hominoid placental physiology. *Proc Natl Acad Sci*
1087 *U S A* 101:1731–1736.
- 1088 Marchler-Bauer A, Lu S, Anderson JB, Chitsaz F, Derbyshire MK, DeWeese-Scott C,
1089 Fong JH, Geer LY, Geer RC, Gonzales NR, et al. 2011. CDD: a Conserved
1090 Domain Database for the functional annotation of proteins. *Nucleic Acids Res*
1091 39:D225-229.
- 1092 Métin C, Vallee RB, Rakic P, Bhide PG. 2008. Modes and mishaps of neuronal
1093 migration in the mammalian brain. *J Neurosci* 28:11746–11752.
- 1094 Miller JA, Ding S-L, Sunkin SM, Smith KA, Ng L, Szafer A, Ebbert A, Riley ZL, Royall
1095 JJ, Aiona K, et al. 2014. Transcriptional landscape of the prenatal human
1096 brain. *Nature* 508:199–206.
- 1097 Milligan MJ, Lipovich L. 2014. Pseudogene-derived lncRNAs: emerging regulators of
1098 gene expression. *Front Genet* 5:476.
- 1099 Muffato M, Louis A, Poisnel C-E, Roest Crolius H. 2010. Genomicus: a database
1100 and a browser to study gene synteny in modern and ancestral genomes.
1101 *Bioinformatics* 26:1119–1121.
- 1102 Muraki K, Tanigaki K. 2015. Neuronal migration abnormalities and its possible
1103 implications for schizophrenia. *Front Neurosci* 9:74.
- 1104 Nakamura H. 2001. Regionalization of the optic tectum: combinations of gene
1105 expression that define the tectum. *Trends Neurosci* 24:32–39.
- 1106 Nasevicius A, Ekker SC. 2000. Effective targeted gene ‘knockdown’ in zebrafish. *Nat*
1107 *Genet* 26:216–220.
- 1108 O’Brien HE, Hannon E, Hill MJ, Toste CC, Robertson MJ, Morgan JE, McLaughlin G,
1109 Lewis CM, Schalkwyk LC, Hall LS, et al. 2018. Expression quantitative trait
1110 loci in the developing human brain and their enrichment in neuropsychiatric
1111 disorders. *Genome Biol* 19:194.
- 1112 Ohno S. 1972. So much “junk” DNA in our genome. *Brookhaven Symp Biol* 23:366–
1113 370.
- 1114 Ohno S. 1999. Gene duplication and the uniqueness of vertebrate genomes circa
1115 1970-1999. *Semin Cell Dev Biol* 10:517–522.
- 1116 Orgel LE, Crick FH. 1980. Selfish DNA: the ultimate parasite. *Nature* 284:604–607.
- 1117 Pan Y-H, Wu N, Yuan X-B. 2019. Toward a Better Understanding of Neuronal
1118 Migration Deficits in Autism Spectrum Disorders. *Front Cell Dev Biol* 7:205.
- 1119 Panier T, Romano SA, Olive R, Pietri T, Sumbre G, Candelier R, Debrégeas G.
1120 2013. Fast functional imaging of multiple brain regions in intact zebrafish
1121 larvae using selective plane illumination microscopy. *Front Neural Circuits*
1122 7:65.

- 1123 Park HC, Kim CH, Bae YK, Yeo SY, Kim SH, Hong SK, Shin J, Yoo KW, Hibi M,
 1124 Hirano T, et al. 2000. Analysis of upstream elements in the HuC promoter
 1125 leads to the establishment of transgenic zebrafish with fluorescent neurons.
 1126 *Dev Biol* 227:279–293.
- 1127 Perelman P, Johnson WE, Roos C, Seuánez HN, Horvath JE, Moreira MAM,
 1128 Kessing B, Pontius J, Roelke M, Rumpler Y, et al. 2011. A molecular
 1129 phylogeny of living primates. *PLoS Genet* 7:e1001342.
- 1130 Plotnikova O, Baranova A, Skoblov M. 2019. Comprehensive Analysis of Human
 1131 microRNA-mRNA Interactome. *Front Genet* 10:933.
- 1132 Rahimi-Balaei M, Bergen H, Kong J, Marzban H. 2018. Neuronal Migration During
 1133 Development of the Cerebellum. *Front Cell Neurosci* 12:484.
- 1134 Rocha M, Singh N, Ahsan K, Beiriger A, Prince VE. 2020. Neural crest development:
 1135 Insights from the zebrafish. *Dev Dyn* 249:88–111.
- 1136 Schizophrenia Working Group of the Psychiatric Genomics Consortium. 2014.
 1137 Biological insights from 108 schizophrenia-associated genetic loci. *Nature*
 1138 511:421–427.
- 1139 Schmittgen TD, Livak KJ. 2008. Analyzing real-time PCR data by the comparative
 1140 C(T) method. *Nat Protoc* 3:1101–1108.
- 1141 Schueler MG, Higgins AW, Rudd MK, Gustashaw K, Willard HF. 2001. Genomic and
 1142 genetic definition of a functional human centromere. *Science* 294:109–115.
- 1143 Sen K, Ghosh TC. 2013. Pseudogenes and their composers: delving in the “debris”
 1144 of human genome. *Brief Funct Genomics* 12:536–547.
- 1145 Sigrist CJA, Cerutti L, Hulo N, Gattiker A, Falquet L, Pagni M, Bairoch A, Bucher P.
 1146 2002. PROSITE: a documented database using patterns and profiles as motif
 1147 descriptors. *Brief Bioinform* 3:265–274.
- 1148 Sinzelle L, Kapitonov VV, Grzela DP, Jursch T, Jurka J, Izsvák Z, Ivics Z. 2008.
 1149 Transposition of a reconstructed Harbinger element in human cells and
 1150 functional homology with two transposon-derived cellular genes. *Proc Natl*
 1151 *Acad Sci U S A* 105:4715–4720.
- 1152 Sperber SM, Saxena V, Hatch G, Ekker M. 2008. Zebrafish *dlx2a* contributes to
 1153 hindbrain neural crest survival, is necessary for differentiation of sensory
 1154 ganglia and functions with *dlx1a* in maturation of the arch cartilage elements.
 1155 *Dev Biol* 314:59–70.
- 1156 Stanke M, Morgenstern B. 2005. AUGUSTUS: a web server for gene prediction in
 1157 eukaryotes that allows user-defined constraints. *Nucleic Acids Res* 33:W465-
 1158 467.
- 1159 Thisse B, Thisse C. 2014. In situ hybridization on whole-mount zebrafish embryos
 1160 and young larvae. *Methods Mol Biol* 1211:53–67.
- 1161 Thisse C, Thisse B. 2008. High-resolution in situ hybridization to whole-mount
 1162 zebrafish embryos. *Nat Protoc* 3:59–69.
- 1163 Velanis CN, Perera P, Thomson B, de Leau E, Liang SC, Hartwig B, Förderer A,
 1164 Thornton H, Arede P, Chen J, et al. 2020. The domesticated transposase

- 1165 ALP2 mediates formation of a novel Polycomb protein complex by direct
1166 interaction with MSI1, a core subunit of Polycomb Repressive Complex 2
1167 (PRC2). *PLoS Genet* 16:e1008681.
- 1168 Volff J-N. 2006. Turning junk into gold: domestication of transposable elements and
1169 the creation of new genes in eukaryotes. *Bioessays* 28:913–922.
- 1170 Warren IA, Naville M, Chalopin D, Levin P, Berger CS, Galiana D, Volff J-N. 2015.
1171 Evolutionary impact of transposable elements on genomic diversity and
1172 lineage-specific innovation in vertebrates. *Chromosome Res* 23:505–531.
- 1173 Wolf JBW, Künstner A, Nam K, Jakobsson M, Ellegren H. 2009. Nonlinear dynamics
1174 of nonsynonymous (dN) and synonymous (dS) substitution rates affects
1175 inference of selection. *Genome Biol Evol* 1:308–319.
- 1176 Wu RS, Lam II, Clay H, Duong DN, Deo RC, Coughlin SR. 2018. A Rapid Method for
1177 Directed Gene Knockout for Screening in G0 Zebrafish. *Dev Cell* 46:112-
1178 125.e4.
- 1179 Yan Y-L, Willoughby J, Liu D, Crump JG, Wilson C, Miller CT, Singer A, Kimmel C,
1180 Westerfield M, Postlethwait JH. 2005. A pair of Sox: distinct and overlapping
1181 functions of zebrafish sox9 co-orthologs in craniofacial and pectoral fin
1182 development. *Development* 132:1069–1083.
- 1183 Yang G, Zhang F, Hancock CN, Wessler SR. 2007. Transposition of the rice
1184 miniature inverted repeat transposable element mPing in *Arabidopsis*
1185 *thaliana*. *Proc Natl Acad Sci U S A* 104:10962–10967.
- 1186 Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol*
1187 24:1586–1591.
- 1188 Zhang Y, You X, Li S, Long Q, Zhu Y, Teng Z, Zeng Y. 2020. Peripheral Blood
1189 Leukocyte RNA-Seq Identifies a Set of Genes Related to Abnormal
1190 Psychomotor Behavior Characteristics in Patients with Schizophrenia. *Med*
1191 *Sci Monit* 26:e922426.
- 1192 Zhao G, Liu L, Zhao T, Jin S, Jiang S, Cao S, Han J, Xin Y, Dong Q, Liu X, et al.
1193 2015. Upregulation of miR-24 promotes cell proliferation by targeting NAIF1 in
1194 non-small cell lung cancer. *Tumour Biol* 36:3693–3701.
- 1195 Zhou X, He J, Velanis CN, Zhu Y, He Y, Tang K, Zhu M, Graser L, de Leau E, Wang
1196 X, et al. 2021. A domesticated Harbinger transposase forms a complex with
1197 HDA6 and promotes histone H3 deacetylation at genes but not TEs in
1198 *Arabidopsis*. *J Integr Plant Biol* 63:1462–1474.

1199

1200

1201

1202 **FIGURES**

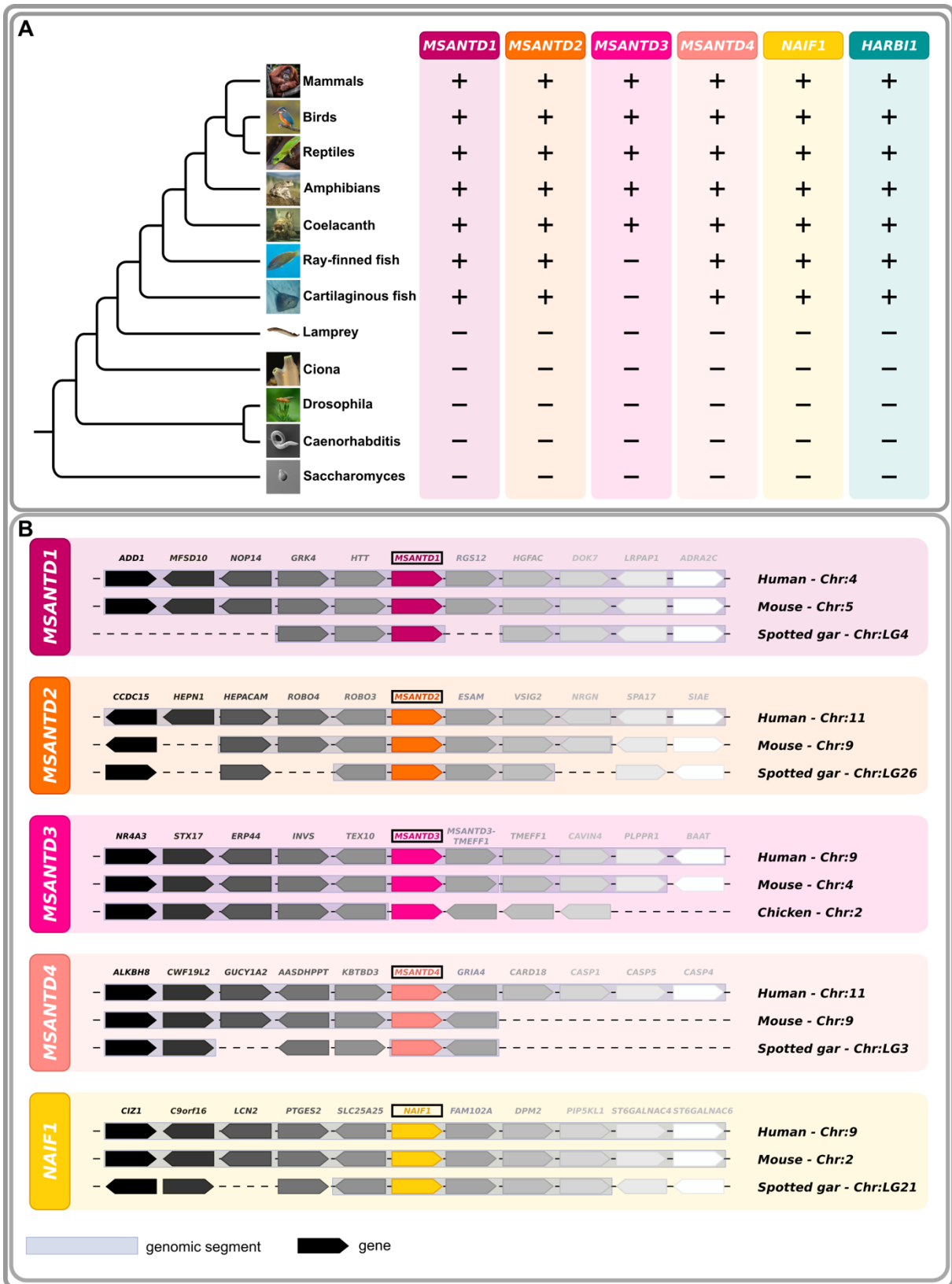


Figure 2

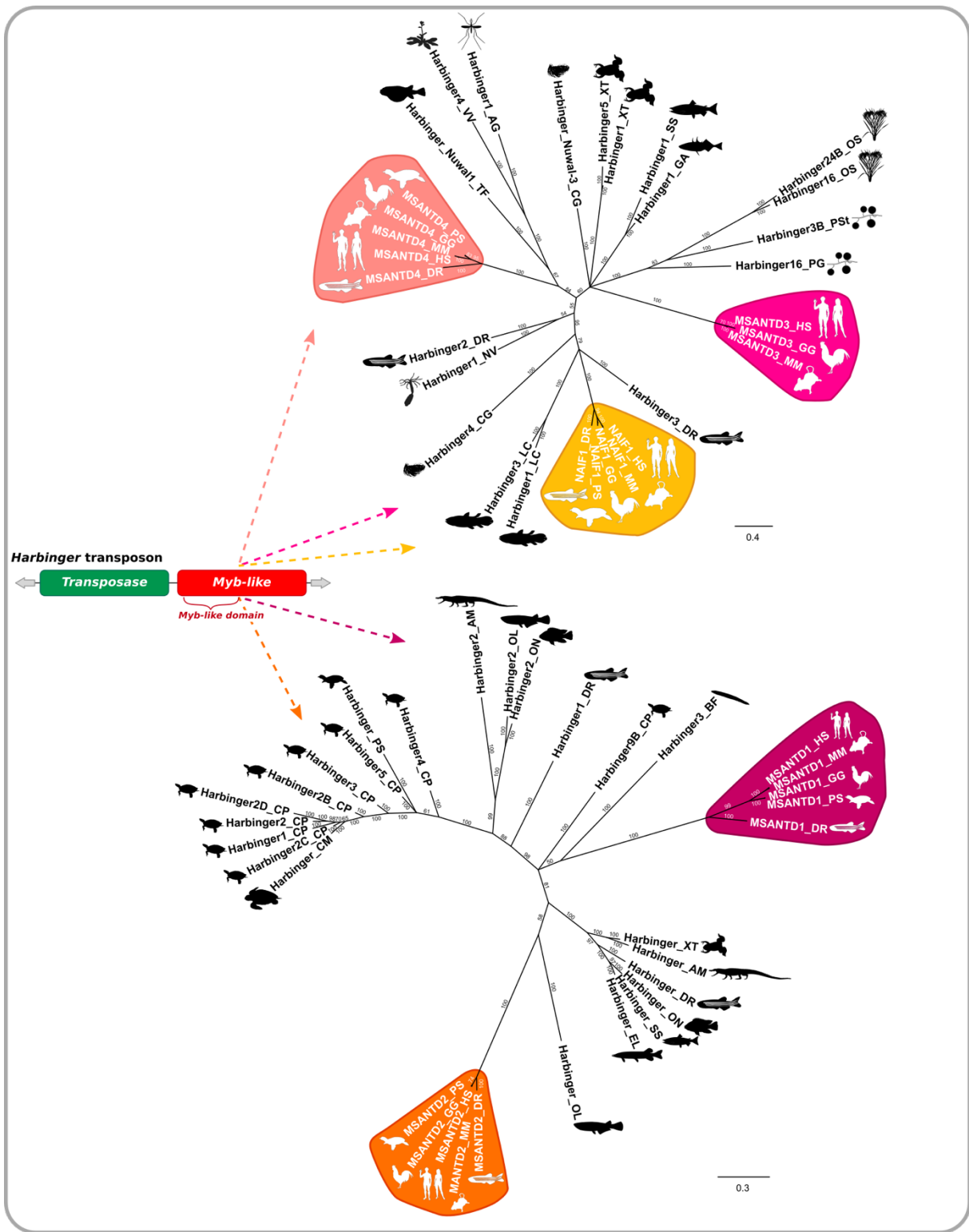


Figure 3

1205

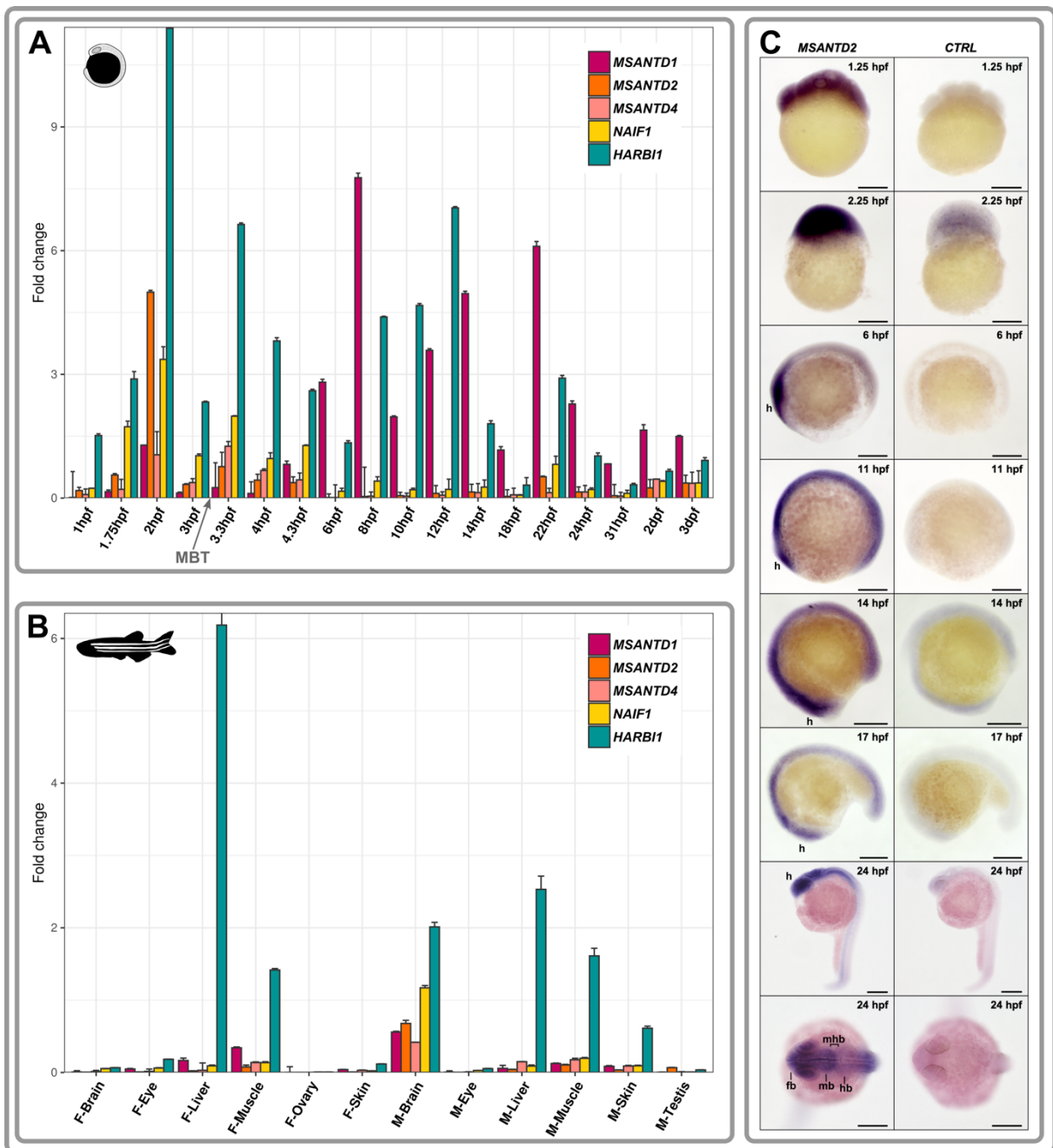


Figure 4

1206

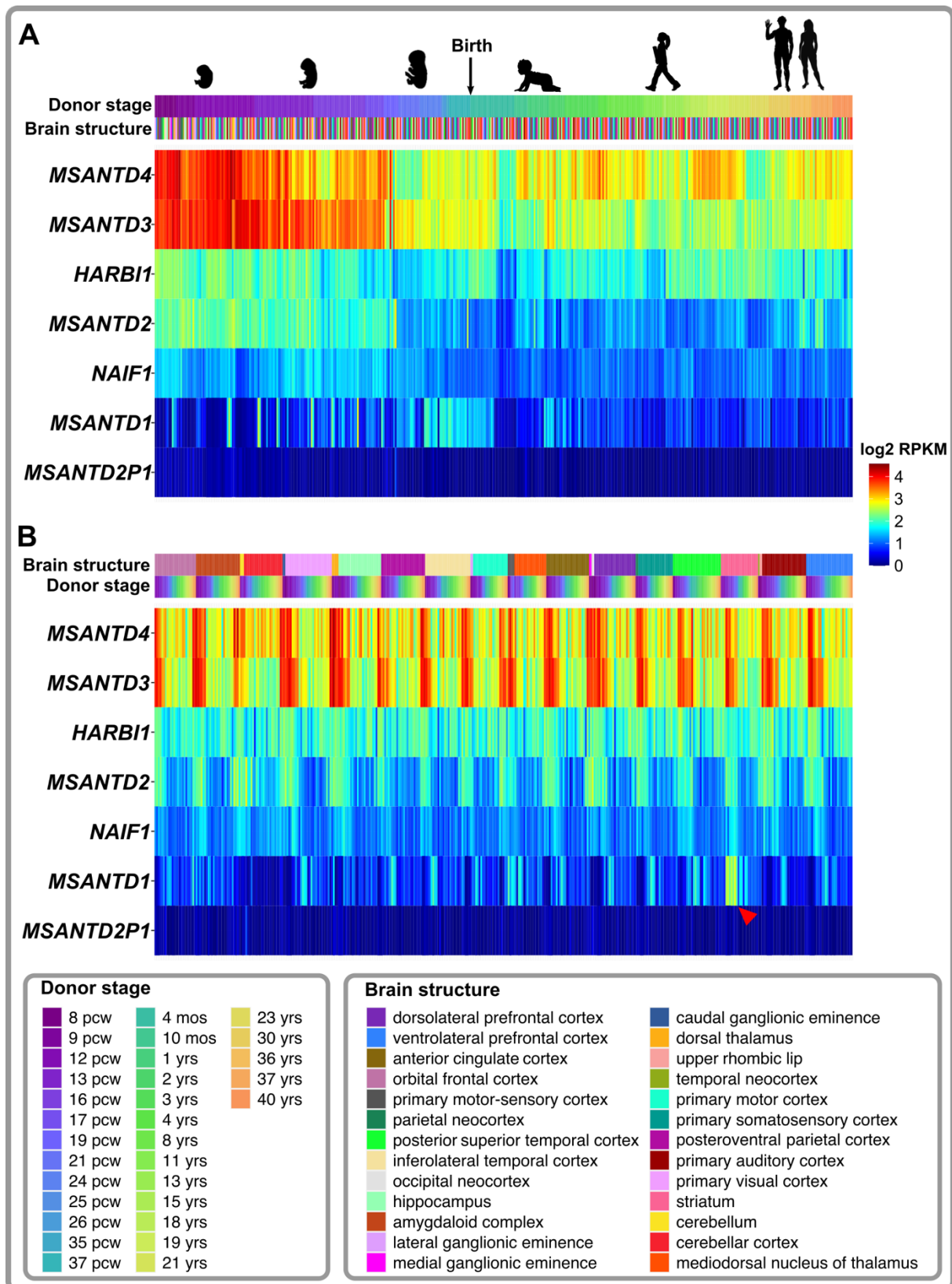


Figure 5

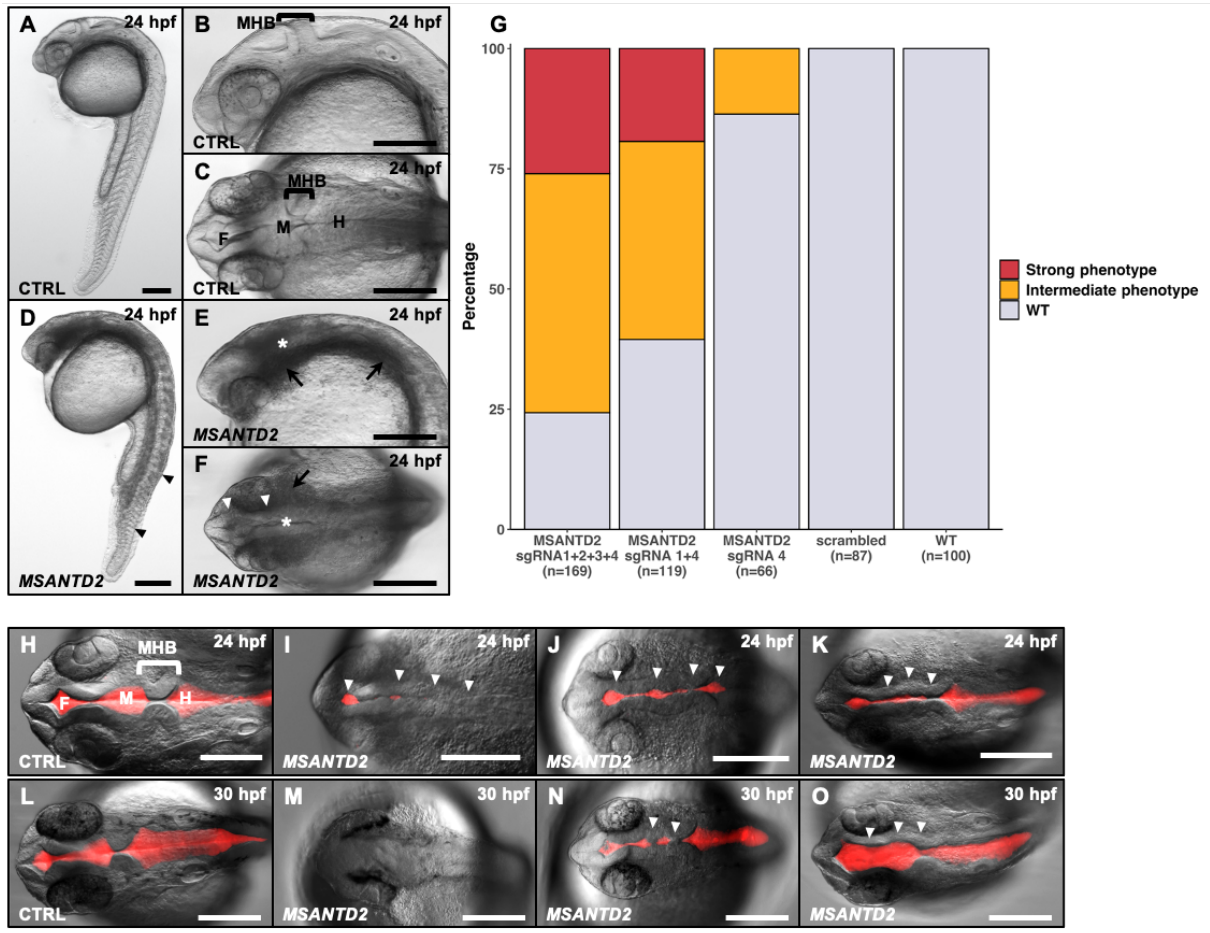


Figure 6

1208

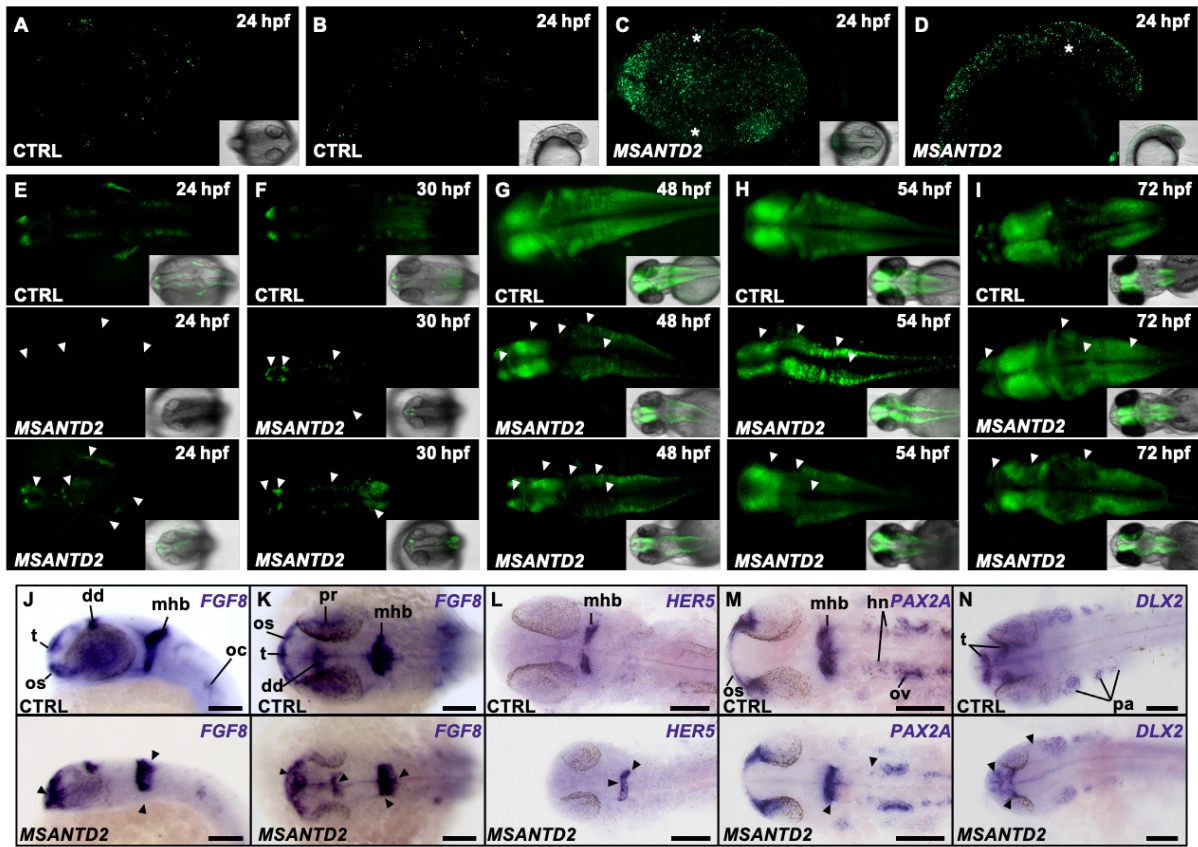


Figure 7

1209

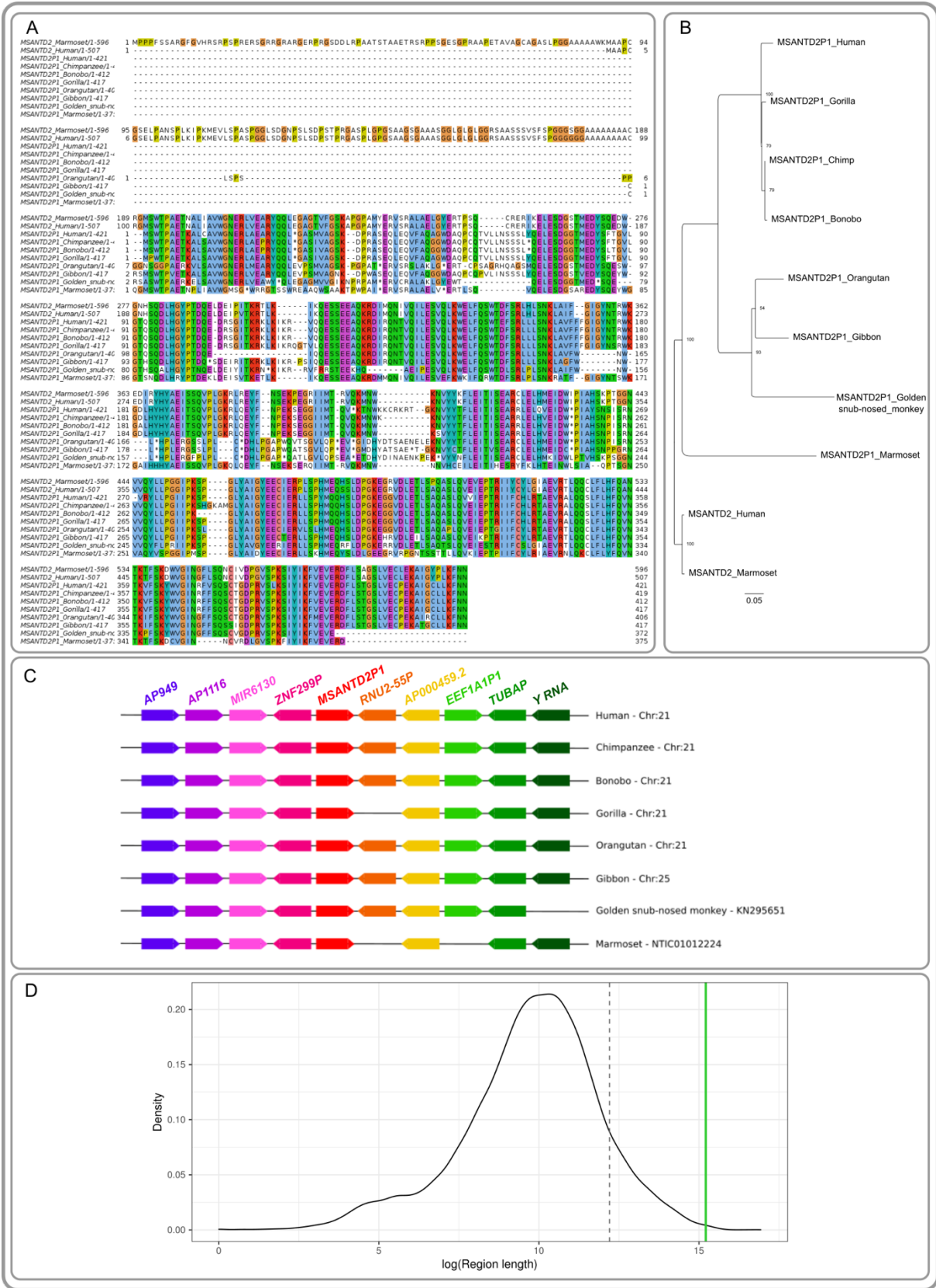


Figure S1

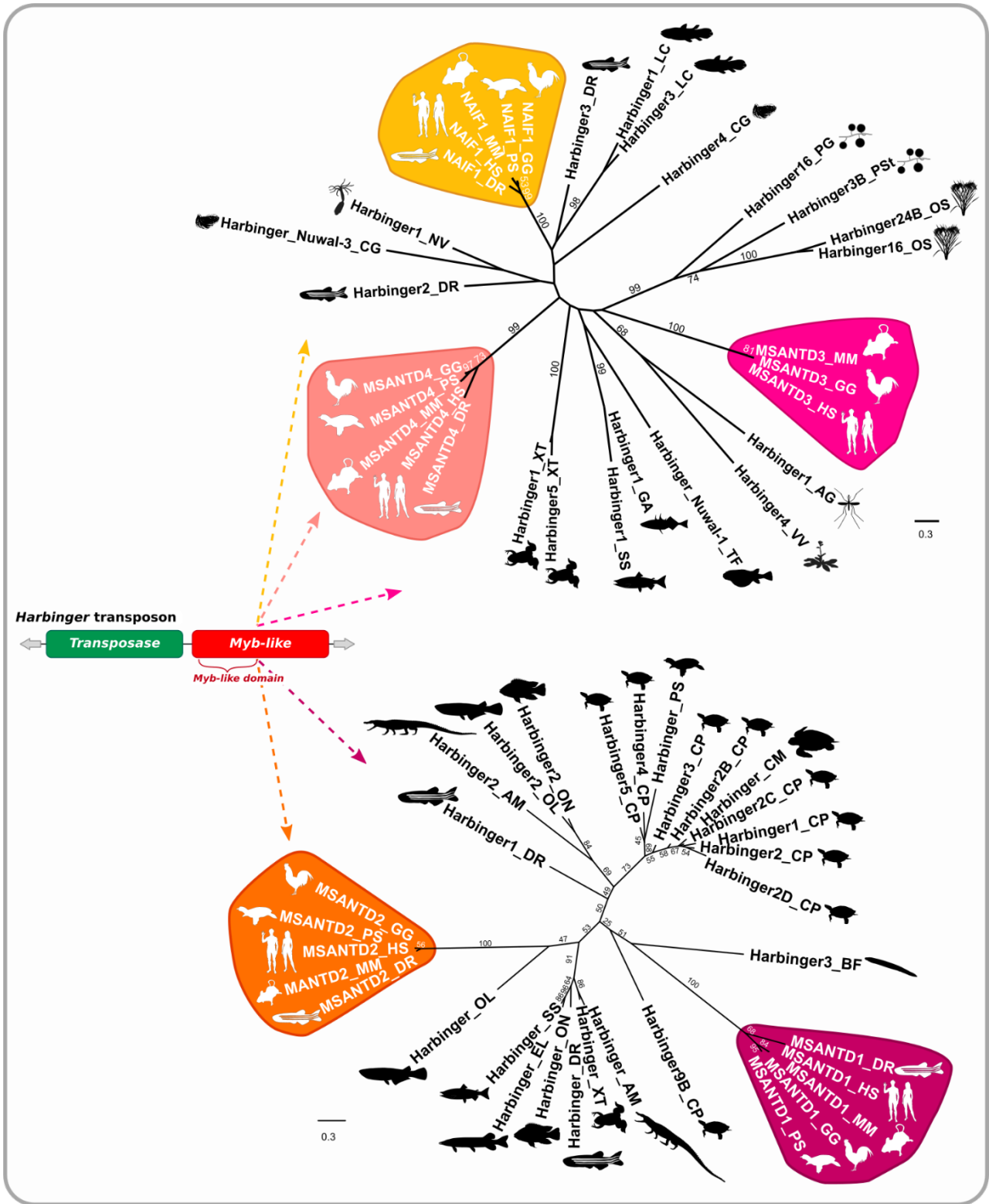


Figure S2

1211

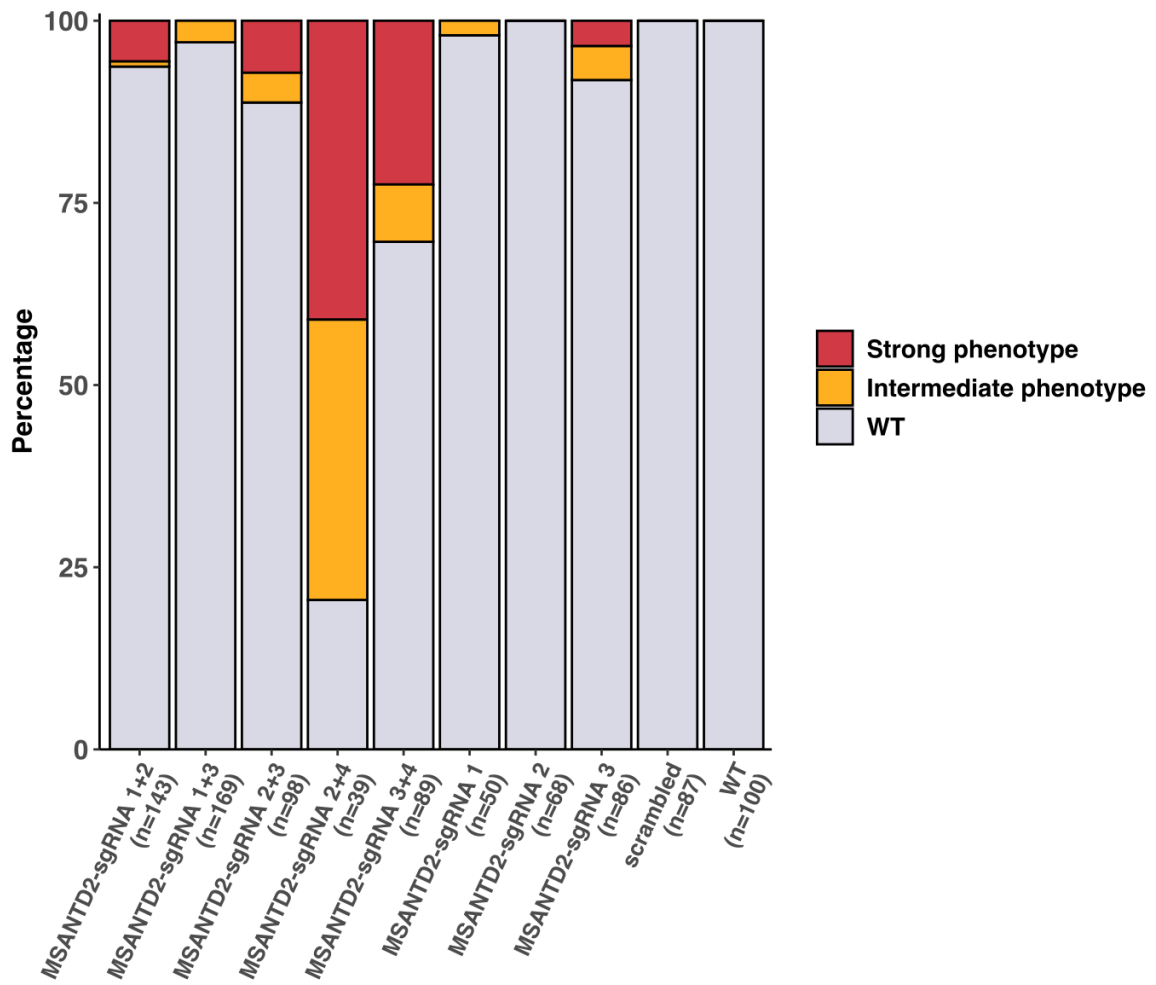


Figure S3

1212

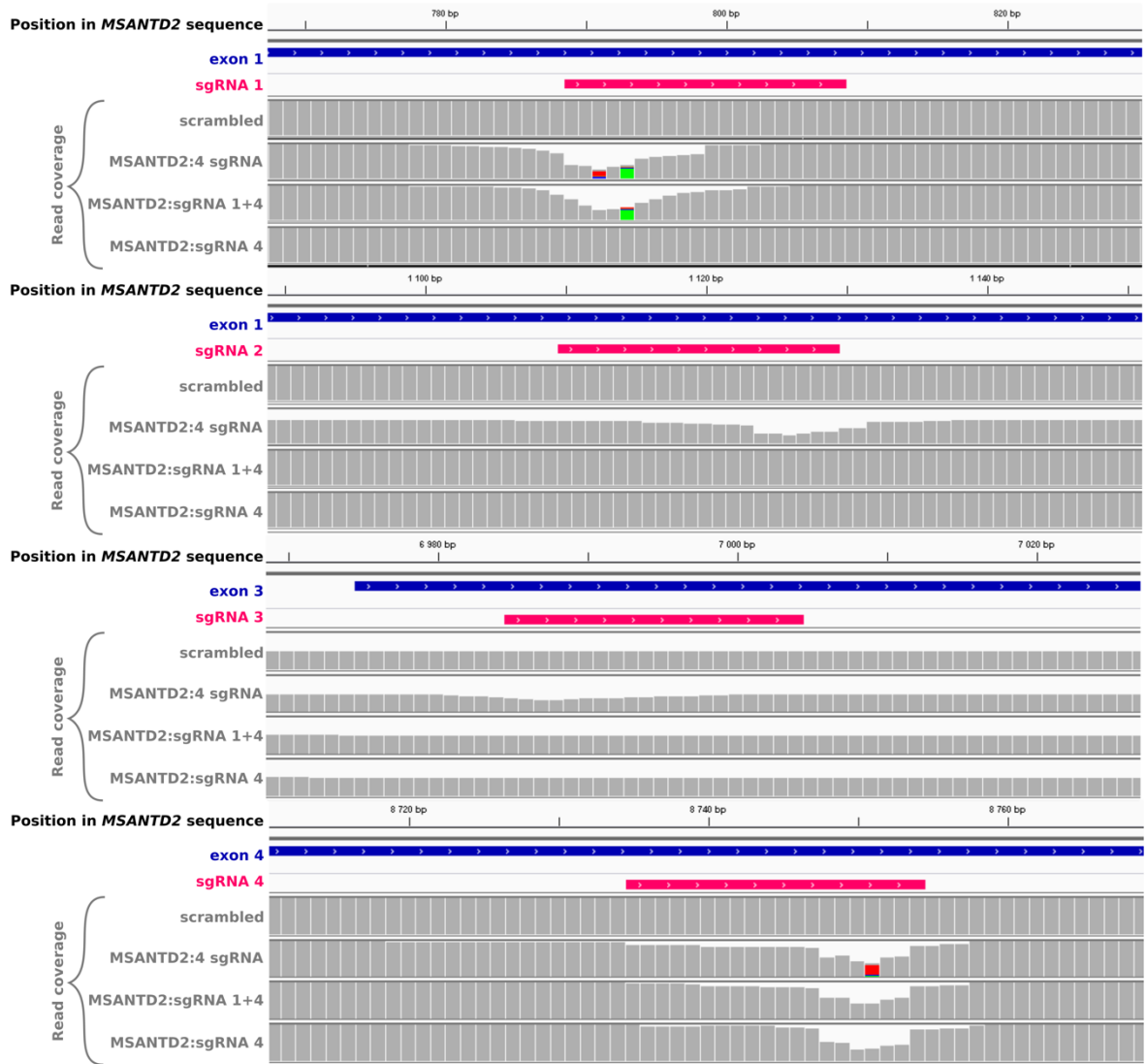


Figure S4

1213



Figure S5

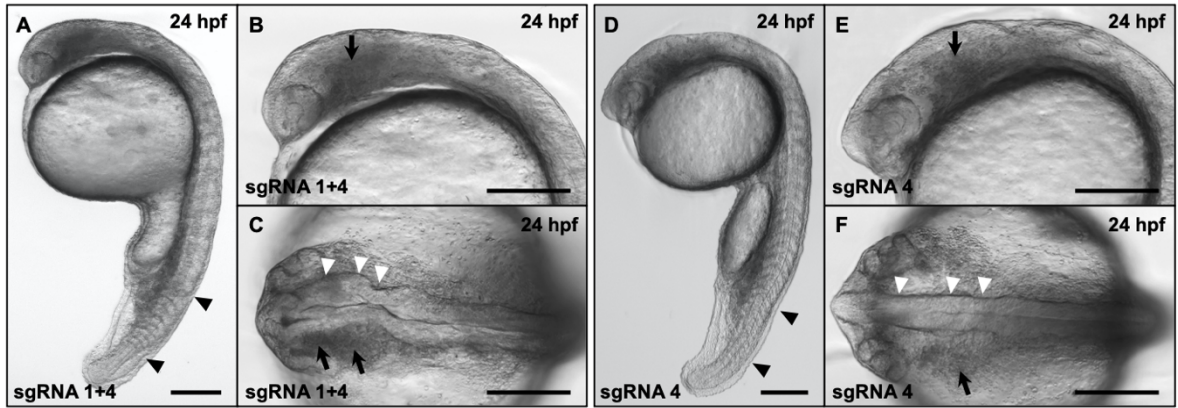


Figure S6

1215

1216

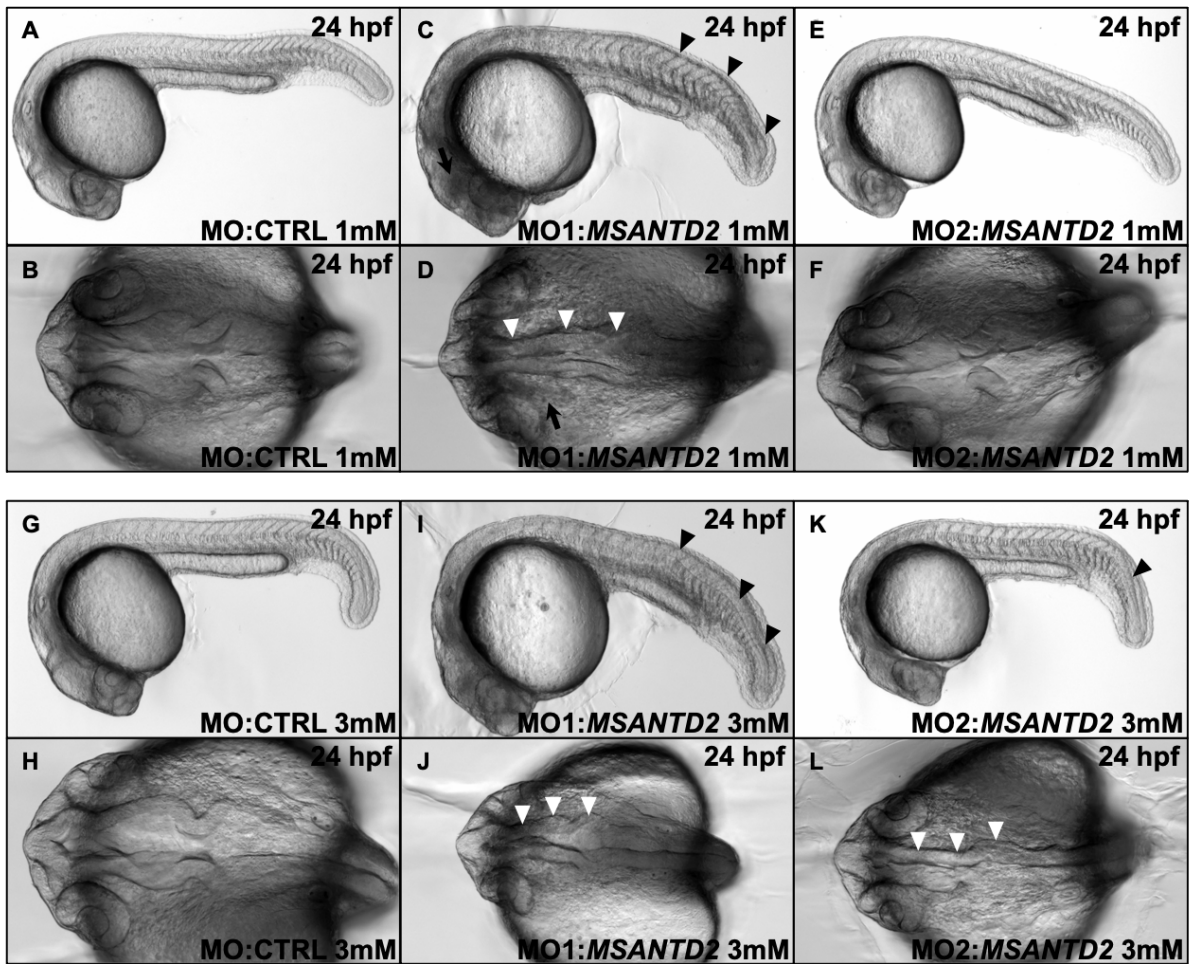


Figure S7

1217

3.6 Conclusion

Grâce à l'utilisation de cette technique de KO-direct par CRISPR/Cas9, j'ai pu étudier dans un premier criblage l'effet de l'inactivation de cinq gènes chez le poisson-zèbre, sans à avoir à générer autant de lignées mutantes. En particulier, l'inactivation du gène *MSANTD2* a produit des embryons aux phénotypes observables et reproductibles en F0 (**section 3.5**). J'ai donc pu sélectionner le gène *MSANTD2* afin d'approfondir son étude.

Au cours de ma thèse, je me suis efforcée de tester si les phénotypes observés étaient bien spécifiques de l'inactivation du gène ciblé, et non pas dus à un biais expérimental de la méthode de Wu et al. (2018). Pour ce faire, j'ai envisagé plusieurs aspects. Tout d'abord j'ai multiplié le nombre de contrôles (injection de sgRNAs scrambled, de sgRNAs ciblant *SAIYAN* et *TYR*). Aucun de ces contrôles n'a reproduit de façon aspécifique les phénotypes observés pour les crispants *MSANTD2*. De plus, les crispants des autres gènes dérivés de transposons *Harbinger* n'ont pas présenté de tels phénotypes non plus. Ces éléments suggèrent fortement que les phénotypes observés ne sont pas dus à un biais aspécifique de la technique utilisée. J'ai également pu reproduire le phénotype avec de multiples combinaisons de sgRNAs ciblant *MSANTD2*, indiquant l'absence de biais provenant d'une toxicité d'un des sgRNAs. Ensuite, j'ai reproduit ces phénotypes grâce à une autre méthode, qui est l'injection d'oligonucléotides antisens de type morpholino, qui génèrent un « knock-down » (le gène n'est pas muté mais son expression est inhibée) et pas un KO du gène comme la méthode CRISPR/Cas9. Enfin, l'expression de *MSANTD2* dans le cerveau de poisson zèbre et dans le cerveau humain au cours du développement, ainsi que l'association de *MSANTD2* à des maladies neuro-développementales humaines, convergent vers les mêmes fonctions biologiques en lien avec le développement du système nerveux. Tout ceci tend donc à suggérer que les phénotypes observés sont dus à un effet spécifique de l'inactivation du gène *MSANTD2*. La génération d'une lignée stable, avec l'observation des phénotypes chez des embryons de la génération F2, est une analyse qui reste à finaliser pour statuer définitivement sur le rôle du gène *MSANTD2* chez les vertébrés.

Pour finir, l'inactivation des autres gènes dérivés de transposons *Harbinger* ne nous a pas permis d'identifier d'indices concernant leurs fonctions biologiques. Des analyses complémentaires, comme l'utilisation de morpholinos ou la génération de lignées KO stables, permettraient de tester la reproductibilité de l'absence de phénotypes. L'absence de phénotypes observée pourrait également être expliquée par un mécanisme de compensation génique. La compensation génique est un phénomène au cours duquel un organisme avec une mutation génique ne développe pas de phénotype, dû à l'expression compensatoire d'un ou d'autres gènes qui contrebalancent fonctionnellement la perte de fonction génotypique en restaurant une fonction biologique normale. Les gènes dérivant de transposons *Harbinger* pourraient partager des fonctions communes du fait de leur parenté, ce qui pourrait être supporté par leur co-expression observée dans certains organes comme par exemple le cerveau (voir **section 3.5**). Ainsi, l'inactivation d'un de ces gènes pourrait être compensée phénotypiquement par un autre. Afin d'étudier cela, il serait intéressant d'induire chez des individus la mutation simultanée de plusieurs gènes dérivés de transposons *Harbinger*, en adaptant le protocole de Wu et al. (2018).

3.7 Annexe

Nom	Séquence
SAIYAN_sgRNA1	GAAGUACCCGGCUCGCCUCC
SAIYAN_sgRNA2	GAGAUUCUGGCACAGAUAGAG
SAIYAN_sgRNA3	UUUGGGGGCUAGUCGCUUGCC
MSANTD1_sgRNA1	AGCACAGGAGGGCUCGUAAC
MSANTD1_sgRNA2	GAUUUGUAGUACGGCCAAUC
MSANTD1_sgRNA3	GGGCUCAUAUUGGCCAUGAUC
MSANTD1_sgRNA4	GACAUGCCUCUUCGACAGCA
MSANTD2_sgRNA1	GGAGAACGCUCAGCGUUACU
MSANTD2_sgRNA2	GUGUUUUCUGGCAAGGCUCC
MSANTD2_sgRNA3	GAACUCGGGUCUUCGGAAGC
MSANTD2_sgRNA4	GGAGCACUUCCAAACGUGGA
MSANTD4_sgRNA1	UGAUCCGCGAGAUCACAAG
MSANTD4_sgRNA2	CGAAAGACUCCUCGUGCGAC
MSANTD4_sgRNA3	AGGAGAGACCACAAGCCUGC
MSANTD4_sgRNA4	GGCGCUAGAAAGGCAGCGAU
NAIF1_sgRNA1	AAAACGAACUUCUCCGAGA
NAIF1_sgRNA2	GGCGCAGGCUCGAGCCGCCA
NAIF1_sgRNA3	GGACGCGGGUACGGAGAUCG
NAIF1_sgRNA4	GAGCAGAAUCGCCUCAACU
HARBI1_sgRNA1	GACCUGCUUCUGCAUGGUCG
HARBI1_sgRNA2	CGCAGCGACUCAUGGAAGCC
HARBI1_sgRNA3	GCAGGAAUACCUAUUGUGAC
HARBI1_sgRNA4	UGCUGACUACCGCUAUAACU
scrambled_sgRNA1	GGCAGGCAAAGAAUCCCUGCC
scrambled_sgRNA2	GGUACAGUGGACCUCGGUGUC
scrambled_sgRNA3	GGCUUCAUACAAUAGACGAUG
scrambled_sgRNA4	GGUCGUUUUGCAGUAGGAUCG
TYR_sgRNA1	GGACUGGAGGACUUCUGGGG
TYR_sgRNA2	GGAUGCAUUAUUACGUGUCC
TYR_sgRNA3	GGAAAGUUACAACCUCCGCG
TYR_sgRNA4	GGUAGUGUGUGCGGGGCGGC

TABLEAU 3.1 – Séquence des sgRNAs utilisés

4

Conclusion et perspectives générales

Sommaire

4.1	Caractérisation des transposons <i>Harbinger</i> chez les poissons téléostéens	142
4.2	Chez les vertébrés les transposons <i>Harbinger</i> sont à l'origine d'une nouvelle famille de gènes .	144
4.3	Les gènes dérivés de transposons <i>Harbinger</i> chez les vertébrés pourraient être impliqués dans le développement du système nerveux	145
4.4	L'utilisation d'une nouvelle technique d'inactivation de gène pour la caractérisation fonctionnelle des gènes	147
4.5	Encore d'autres gènes dérivés de transposons <i>Harbinger</i> ?	149
4.6	Conclusion	157

L'objectif de ma thèse a donc été d'évaluer la contribution des éléments transposables comme source de nouveaux gènes à l'origine de potentielles innovations développementales chez les vertébrés. L'évolution des vertébrés est caractérisée par l'émergence de multiples innovations développementales qui ont contribué au succès évolutif de ce clade. En effet, les vertébrés présentent des os, du cartilage, des paires de membres, un système endocrinien complexe, des placodes sensorielles, une crête neurale, un système immunitaire adaptatif, ainsi qu'une importante complexification de leur système nerveux. Cependant, l'origine génétique de ces innovations n'est pas toujours caractérisée. Deux événements de duplication de génome entier ont eu lieu à la base des vertébrés (Dehal & Boore, 2005). Ainsi, Ohno proposait que de tels événements, qui permettent une extension majeure du répertoire de gènes, sont à l'origine de l'apparition massive d'innovations (Ohno, 1999). Cependant, comme en témoigne les gènes *SYNCYTIN*, impliqués dans le développement du placenta, et les gènes *RAG*, à l'origine du système immunitaire adaptatif, mes travaux de thèse suggèrent également que les nouveaux gènes formés par la domestication moléculaire des éléments transposables ont probablement contribué de façon significative à l'apparition d'innovations et à l'évolution précoce des vertébrés.

4.1 Caractérisation des transposons *Harbinger* chez les poissons téléostéens

Lors de cette thèse je me suis en particulier intéressée aux transposons *Harbinger*. Il s'agit de transposons à ADN que l'on retrouve dans de très nombreuses espèces mais qui sont malgré tout assez faiblement caractérisés. C'est ainsi que, dans un premier temps, je me suis consacrée à l'étude de ces éléments chez les poissons téléostéens. Les poissons téléostéens représentent le clade avec le plus grand nombre d'espèces chez les vertébrés, mais ils restent encore peu étudiés. La diversité de ce groupe en fait donc un modèle de choix pour permettre une compréhension plus complète de la biodiversité génomique. De plus, les génomes de ces espèces sont riches en éléments transposables, tant du point de vue quantitatif que qualitatif, ce qui rend leur étude d'autant plus adaptée dans le cadre de cette thèse. Ainsi, à travers l'étude de huit génomes de poissons téléostéens, j'ai pu montrer que les transposons *Harbinger* sont répandus chez ces espèces mais avec des contributions génomiques fortement variables. Mes résultats suggèrent que les transposons *Harbinger* ont présenté des succès évolutifs différents dans ces espèces, certainement lié à des capacités d'invasion et des mécanismes de répression variables entre les espèces.

Une des caractéristiques majeures et non-répandues chez les éléments transposables est le fait que les transposons *Harbinger* soient composés de deux ORFs, codant pour deux protéines : une transposase et une protéine Myb-like. Dans le but d'examiner l'évolution de ces transposons, j'ai étudié les relations phylogénétiques de leurs deux protéines. Les résultats suggèrent qu'elles ont généralement suivi la même histoire évolutive, indiquant une origine unique de l'association du gène de la transposase et du gène *Myb-like* dans les transposons *Harbinger*. Ceci peut être expliqué par une co-évolution des deux ORFs au sein d'un même élément, qui est nécessaire au maintien de l'interaction entre les deux protéines pour permettre la transposition de ce type d'élément. De plus, malgré l'existence de certains transposons *Harbinger* présentant un seul ORF, de tels éléments n'ont pas été détectés chez les poissons téléostéens, ni dans les génomes de vertébrés d'après la

littérature. Ceci suggère que cette structure en deux ORFs distincts est évolutivement plus efficace pour le maintien et l'expansion des éléments *Harbinger* dans les génomes de poissons. Ceci peut éventuellement aussi indiquer que la structure en un seul ORF a été plus facilement réprimée et éliminée de ces génomes pour des raisons encore inconnues. De plus, j'ai pu identifier pour la première fois l'expression des éléments *Harbinger* chez des espèces de poissons téléostéens, révélant leur activité transcriptionnelle. Les résultats montrent également l'expression de ces éléments dans les gonades du poisson médaka, témoignant d'une possible expansion de cette famille de transposons dans de futures générations.

En conclusion, ces résultats de thèse ont pu démontrer que les éléments *Harbinger* sont des transposons répandus chez les poissons téléostéens. De plus, ils se composent de deux ORFs, caractéristique singulière chez les éléments transposables, dont l'évolution reste étroitement liée, indiquant une valeur évolutive de cette structure.

On note cependant que l'étude se porte ici sur un nombre assez restreints d'espèces de poissons téléostéens. L'utilisation de séquences correspondant à des consensus de familles de transposons *Harbinger* est également une limite à prendre en compte dans cette analyse.

Grâce au continuel accroissement des données de séquençage, ainsi qu'au développement d'outils d'annotation des génomes, il serait donc important d'élargir l'étude des éléments *Harbinger* à d'autres poissons téléostéens, afin d'approfondir leur caractérisation et d'y étudier leur contribution à la structure des génomes et l'évolution des organismes. En effet, une étude récente a démontré une invasion de transposons *Harbinger* dans l'ADN de serpents de mer, ce qui a conduit à une expansion majeure de leurs génomes (Galbraith et al., 2021). Un tel événement n'a pas été identifié chez les poissons étudiés, mais au vu du nombre d'espèces composant le groupe des poissons téléostéens, l'hypothèse d'un tel événement dans ce clade reste probable. Étendre l'étude des séquences de transposons *Harbinger* aux copies individuelles issues de différents génomes de poissons plutôt qu'aux séquences consensus permettrait de mieux comprendre l'histoire évolutive de ces ETs. On pourrait par exemple plus facilement détecter des événements de transfert horizontaux, ou de recombinaison entre des éléments appartenant à différentes familles de *Harbinger*.

De plus, dans l'analyse présentée ici, on note une disparité de couverture des éléments *Harbinger* dans les génomes, dont les raisons sont encore inconnues. À notre connaissance, aucune étude n'a examiné les mécanismes de répression de ces éléments. L'identification de mécanismes impliquant des piRNAs ou des marques épigénétiques pourrait donner des indices expliquant la faible représentation des éléments *Harbinger* dans certains génomes de poissons. Enfin, le mécanisme d'insertion de ces éléments n'est pas caractérisé et manque donc pour comprendre l'évolution de ces éléments dans les génomes. Il pourrait être intéressant de comparer ce mécanisme pour les éléments ISL2EU et *Harbinger*, ainsi que les *Harbinger* à deux ou un seul ORF. L'étude de ce mécanisme pourrait également fournir des éléments de réponse expliquant la disparité des éléments *Harbinger* dans les génomes.

4.2 Chez les vertébrés les transposons *Harbinger* sont à l'origine d'une nouvelle famille de gènes

La majeure partie de mon travail de thèse a consisté en l'identification et la caractérisation fonctionnelle et évolutive de gènes dérivés de transposons *Harbinger* chez les vertébrés (voir **chapitre 3**). Pour ce faire, j'ai comparé des banques de séquences d'éléments transposables avec les séquences de protéines humaines. Ceci m'a permis d'identifier quatre nouveaux gènes potentiellement dérivés de transposons *Harbinger* : *MSANTD1*, *MSANTD2*, *MSANTD3* et *MSANTD4*. Ces gènes s'ajoutent aux gènes *HARBII* et *NAIF1*, précédemment identifiés comme dérivant de transposons *Harbinger* (Kapitonov & Jurka, 2004; Sinzelle et al., 2008), ainsi qu'à un pseudogène *MSANTD2P1*, que j'ai identifié chez les simiens. Plus particulièrement, les gènes *MSANTD1*, *MSANTD2*, *MSANTD3*, *MSANTD4* et *NAIF1* (gènes *MSANTD*) dérivent de gènes *Myb-like* alors que *HARBII* dérive du gène de la transposase. Le pseudogène *MSANTD2P1* est quant à lui issu de la rétrotransposition d'un ARNm de *MSANTD2*. Ces gènes dérivés de transposons *Harbinger* sont conservés chez les vertébrés à mâchoires, à l'exception du gène *MSANTD3* uniquement présent chez les sarcoptérygiens.

L'analyse évolutive des gènes *MSANTD* chez les vertébrés m'a permis d'établir qu'ils sont issus de trois à cinq événements de domestications moléculaires indépendants de transposons *Harbinger* : de trois ou quatre événements à la base des vertébrés à mâchoires, il y a environ 500 millions d'années, et un autre potentiel à la base des sarcoptérygiens il y a environ 430 millions d'années. L'origine de ces gènes a été étudiée selon deux méthodes d'analyses phylogénétiques. Les résultats de la méthode Bayésienne suggèrent ainsi la survenue de cinq événements de domestications indépendants à l'origine des cinq gènes de cette famille. Cependant, la méthode de maximum de vraisemblance, même si elle ne va pas totalement à l'encontre de l'hypothèse de cinq événements indépendants, ne soutient que trois des événements de façon significative. Ainsi, des analyses supplémentaires seraient nécessaires pour permettre de clarifier l'origine de ces gènes.

De plus, j'ai pu observer que la structure secondaire ancestralement présente dans les protéines *Myb-like* de transposons *Harbinger* est conservée chez les protéines *MSANTD*. La protéine *Myb-like* étant capable de se lier à l'ADN et à la transposase, cela suggère que les protéines *MSANTD* pourraient également avoir de telles capacités (Kapitonov & Jurka, 2004; Sinzelle et al., 2008). En effet, des études ont montré que la protéine *MSANTD3* est capable de se lier à l'ADN et où il recrute le complexe protéique PRC2 (polycomb repressive complex 2), ce qui permet de réguler la différenciation neuronale dans les cellules P19 de souris. De plus, une étude *in vitro* dans des cellules humaines a indiqué que la protéine *NAIF1* se fixe à l'ADN (mais non spécifiquement au TIR de transposons *Harbinger*) mais est aussi capable de se lier à la transposase des éléments *Harbinger* ainsi qu'à la protéine *HARBII* (Sinzelle et al., 2008). La capacité de *NAIF1* à se lier à la transposase pourrait suggérer un rôle des gènes *MSANTD* dans la régulation de la transposition des éléments *Harbinger*, potentiellement en empêchant l'interaction avec la protéine *Myb-like* du transposon et ainsi inhiber la transposition. Cependant, il est à noter que les gènes *MSANTD* sont présents et soumis à une pression sélective chez les vertébrés dont les mammifères, d'où les transposons *Harbinger* sont absents. Ainsi, même si les gènes *MSANTD* pourraient être recrutés pour réguler la transposition des éléments *Harbinger*, cela n'est certainement pas leur unique fonction.

Finalement, mes travaux ont permis de mettre en évidence des domestications moléculaires

récurrentes de transposons *Harbinger* au cours de l'évolution précoce des vertébrés, ayant généré cinq gènes codant pour des protéines avec potentiellement des capacités de liaison à l'ADN/à des protéines. Ainsi, ceci illustre un nouveau cas de convergence évolutive concernant la domestication moléculaire des éléments transposables, comme observé pour les SYNCYTINES chez les mammifères placentaires et les gènes CENPB chez la levure à fission et les mammifères (Casola et al., 2007; Lavalie et al., 2013).

On note cependant que les résultats présentés ici ne permettent pas de conclure quant à l'origine de tous les gènes *MSANTD*, du fait des résultats différents entre les deux méthodes de phylogénies utilisées. Les différences de résultats observées entre ces deux méthodes sont certainement liées à la grande variabilité des séquences des protéines Myb-like et à une dégénérescence des gènes après leur domestication qui conduisent à un manque de résolution. Comme on ne peut pas favoriser une des deux méthodes de construction phylogénétique, il serait pertinent d'augmenter la taille du jeu de données pour pallier le manque de résolution. En particulier, d'autres protéines Myb-like, plus proches des protéines *MSANTD* que celles présentées dans ces travaux, pourraient exister dans des transposons encore non caractérisés. L'annotation de nouvelles séquences de transposons *Harbinger* dans de nouvelles espèces pourrait permettre d'associer les protéines *MSANTD* avec un nombre plus important de protéines Myb-like de transposons, augmentant de facto la résolution de l'analyse. Cela pourrait permettre de conclure quant à l'hypothèse de cinq événements de domestications indépendants qui est suggérée par la phylogénie Bayésienne.

Enfin, il est nécessaire de tester la capacité de liaison à l'ADN de toutes les protéines *MSANTD*. En effet, notre hypothèse est que ces gènes ont conservé les capacités moléculaires initialement présentes dans les séquences ancestrales de Myb-like. Cependant, ces protéines ont probablement évolué et donc potentiellement modifié leurs interactions avec l'ADN, par exemple en ciblant d'autres séquences, permettant de réguler de nouveaux gènes. Cette question cruciale pourrait être étudiée par des expériences de ChIP-seq (Chromatin Immuno-Precipitation and sequencing) afin de tester cette capacité de liaison à l'ADN et d'identifier les séquences cibles.

4.3 Les gènes dérivés de transposons *Harbinger* chez les vertébrés pourraient être impliqués dans le développement du système nerveux

Malgré un nombre grandissant de gènes dérivés d'éléments transposables identifiés, le manque général de caractérisation fonctionnelle limite la compréhension de leur implication dans l'évolution des espèces.

Ainsi, je me suis par la suite focalisée sur l'aspect fonctionnel des gènes *MSANTD*. Pour cela, j'ai utilisé le poisson-zèbre comme espèce modèle. C'est un organisme particulièrement adapté à la réalisation de génétique inverse et à la caractérisation de phénotypes, en particulier au cours du développement embryonnaire (voir section 1.4). J'ai tout d'abord mis en évidence l'expression des gènes dérivés de transposons *Harbinger* au cours du développement embryonnaire du poisson-zèbre, particulièrement lors des étapes précoces, indiquant un effet maternel. J'ai également pu identifier une expression dans plusieurs tissus adultes de poisson-zèbre, avec particulièrement une co-expression dans le cerveau mâle. Ces gènes sont également exprimés dans de nombreux

organes humains, notamment le cerveau, en particulier lors du développement fœtal. Les données étudiées dans le cerveau humain ne m'ont pas permis de distinguer les résultats en fonction du sexe. Il serait donc intéressant d'étudier d'autres données chez l'humain ou d'autres espèces de vertébrés, afin de tester l'existence de différences d'expression des gènes *MSANTD* en fonction du sexe.

Par ailleurs, la période d'expression des gènes dérivés de transposons *Harbinger* dans le cerveau fœtal humain correspond au moment où la migration neuronale se met en place. Ces mécanismes migratoires sont essentiels au développement et au fonctionnement corrects du système nerveux, puisque leur perturbation conduit à l'apparition de troubles neuro-développementaux tels que la schizophrénie, les troubles du spectre de l'autisme et l'épilepsie (Fatemi, 2005; Guerrini & Parrini, 2010; Muraki & Tanigaki, 2015; Pan et al., 2019). De façon intéressante, l'un des gènes *MSANTD*, *MSANTD2*, a effectivement été associé à la schizophrénie et aux troubles du spectre de l'autisme chez l'Homme (Lim et al., 2017; O'Brien et al., 2018; Schizophrenia Working Group of the Psychiatric Genomics Consortium, 2014; Zhang et al., 2020). L'étude plus particulière de ce gène m'a permis de démontrer son expression chez le poisson-zèbre au cours du développement embryonnaire dans la région qui aboutira à la tête à partir de 6 heures post-fertilisation (hpf), puis effectivement dans la région du cerveau à 24 hpf. De plus, l'inactivation de ce gène par une technique de KO-direct par CRISPR/Cas9 et la méthode knock-down des morpholinos produit des embryons présentant des troubles développementaux notamment au niveau du système nerveux. En effet, j'ai pu observer des défauts de repliement du tube neural au niveau du cerveau à 24hpf. Ces malformations structurales ont été reliées à des problèmes cellulaires, puisqu'une accumulation de cellules mortes a été observée dans les embryons, ainsi que des défauts de marquage des neurones et enfin des altérations d'expression d'autres gènes impliqués dans le développement du système nerveux des vertébrés. Plus particulièrement, les patrons d'expression altérés des gènes *DLX2A* et *PAX2A* pourraient suggérer une implication de *MSANTD2* dans la migration et/ou la différenciation des cellules de la crête neurale ou des neurones.

Ainsi, les résultats de l'inactivation du gène *MSANTD2*, son patron d'expression chez le poisson-zèbre et dans le cerveau humain, ainsi que son association à des maladies neuro-développementales chez l'Homme, tendent à suggérer un rôle de *MSANTD2* dans le développement du système nerveux des vertébrés.

Il est donc nécessaire d'étudier plus en détail le rôle de *MSANTD2*, et notamment l'hypothèse de son implication dans la migration et/ou la différenciation des neurones et des cellules de la crête neurale. Pour ce faire, d'autres marqueurs caractéristiques de ces mécanismes pourraient être étudiés tels que *SOX10*, *FOXD3* et *CRESTIN* (Rocha et al., 2020). Il pourrait être particulièrement pertinent d'inactiver le gène *MSANTD2* dans une lignée de poissons transgéniques marquant par fluorescence les cellules de la crête neurale et/ou les neurones, pour suivre in vivo la migration et la différenciation de ces cellules. Ceci permettrait d'avoir plus d'indices quant au rôle de *MSANTD2* dans le développement du système nerveux des vertébrés.

Par ailleurs, l'inactivation des autres gènes dérivés de transposons *Harbinger* par la technique de KO-direct par CRISPR/Cas9 chez le poisson-zèbre ne m'a pas permis d'identifier de phénotypes remarquables qui auraient permis d'approcher leurs fonctions. Il est important de noter que l'étude des phénotypes des embryons a été réalisée à une échelle large. Cette approche a été adoptée dans

l'optique d'un criblage pour identifier l'implication éventuelle d'au moins un des gènes *MSANTD* dans des phénomènes développementaux remarquables, par des défauts morphologiques ou comportementaux majeurs. C'est un des avantages du modèle poisson zèbre qui est utilisé ici, puisque simplement en observant les embryons vivants, il est possible d'identifier des structures telles que la notochorde, le tube neural, le cœur, le cerveau, l'aorte dorsal dès 24hpf. À 5 jours post fertilisation, l'artère de la queue, l'intestin, l'estomac, le foie, la bouche, les yeux et les muscles sont clairement visibles. Ceci permet donc de balayer un éventail important de phénotypes. Mais nous ne pouvons tout de même pas exclure la présence de phénotypes non visibles à l'échelle étudiée.

Une absence complète de phénotype pourrait tout de même être expliquée par la compensation génique due à une redondance fonctionnelle avec d'autres gènes de la famille (El-Brolosy & Stainier, 2017; Tautz, 1992). L'inactivation de plusieurs gènes *MSANTD* en même temps permettrait d'étudier ce phénomène. L'absence de phénotype pourrait également être liée à un effet maternel. En effet, j'ai pu observer que ces gènes sont particulièrement exprimés avant la transition mid-blastuléenne chez le poisson-zèbre, c'est-à-dire avant l'activation de la transcription du génome zygotique, suggérant la présence d'ARNm maternels. L'inactivation par CRISPR/Cas9 des gènes *MSANTD* pourrait donc être compensée par la présence de ces ARNm. L'utilisation de la technique des morpholinos (qui n'a été réalisée que pour *MSANTD2*), bloquant la traduction des ARNm zygotiques et maternels et la génération de lignées de mutants, pourrait donc nous donner des éléments de réponse face à cette hypothèse.

4.4 L'utilisation d'une nouvelle technique d'inactivation de gène pour la caractérisation fonctionnelle des gènes

Ces dernières années l'amélioration des méthodes d'annotations et de l'augmentation de la quantité et qualité des séquences de génomes complets a permis d'accentuer l'identification des gènes dérivés d'ETs. Cependant, comme évoqué précédemment, leur caractérisation fonctionnelle reste limitée, ce qui restreint la compréhension de leur implication dans l'évolution et l'adaptation des espèces.

L'utilisation de méthodes rapides d'inactivation de gènes, en amont ou à la place de l'établissement de lignées stables nécessitant plusieurs générations, pourrait contribuer à réduire ce manque. En effet, obtenir des lignées stables est un processus chronophage et coûteux. Il peut donc être intéressant, avant de lancer ce processus pour des gènes dont les fonctions sont totalement inconnues, d'effectuer une étape de criblage à l'aide de méthodes plus rapides, pour ensuite restreindre les analyses approfondies aux gènes retournant des résultats prometteurs, puisque ces méthodes reproduisent les phénotypes observés dans des lignées KO (Buglo et al., 2020; Hoshijima et al., 2019; Kroll et al., 2021; Teboul et al., 2017; Wu et al., 2018).

C'est donc dans cette optique que j'ai adapté la méthode de KO-direct de CRISPR/Cas9 développée par Wu et al. (2018). Cette technique m'a donc permis d'éviter la génération de cinq lignées de poissons, et j'ai pu identifier plus rapidement un gène dont l'inactivation produit un phénotype remarquable sur lequel j'ai pu me focaliser. Cette technique récente et novatrice tend à se populariser puisque de plus en plus de travaux l'utilisent pour caractériser fonctionnellement

des gènes chez le poisson-zèbre (Buglo et al., 2020; Hoshijima et al., 2019; Kroll et al., 2021; Teboul et al., 2017).

En plus de sa rapidité, cette méthode permet également d'étudier les phénomènes de compensation génique. En effet, lorsqu'un gène est inactivé des phénomènes compensatoires peuvent se mettre en place, soit par la sur-expression d'autres gènes aux fonctions redondantes, soit par la mise en place de réponses plus complexes au sein de réseaux de transcription, signalisation ou métaboliques (El-Brolosy & Stainier, 2017; Nadeau, 2001; Sztal et al., 2018). Comme plusieurs sgRNAs sont utilisés dans cette technique pour cibler le même gène, il est en effet possible d'utiliser plusieurs sgRNAs ciblant deux gènes aux fonctions redondantes afin d'éliminer le phénomène compensatoire.

Au cours de cette thèse, la question de la fiabilité de la technique de KO-direct par CRISPR/Cas9 s'est posée, et en particulier celle de la spécificité des phénotypes que j'ai observés pour le gène *MSANTD2*. En effet, si Wu et al. démontrent la reproduction des phénotypes observés pour des mutants F2 directement en F0 grâce à leur technique, l'éventualité de biais et de toxicité de l'expérience ne peut être totalement écartée, particulièrement pour une technique encore peu utilisée comme celle-ci. La technique de CRISPR/Cas9 implique des mutations «off-targets», qui ne peuvent être évités, et la probabilité de ces mutations augmente avec l'augmentation du nombre de sgRNA utilisés. En plus des «off-targets», il peut également y avoir des problèmes de toxicité en cas de trop nombreuses coupures double brin de l'ADN, ce qui entraîne alors un fort taux de mort cellulaire, qui peut engendrer des défauts de développement. Au cours de ma thèse, je me suis efforcée de m'affranchir de ces biais pour statuer sur les phénotypes observés. La meilleure approche pour cela aurait été de pouvoir reproduire les phénotypes observés en F0, dans des embryons F2 issus d'une lignée stable de mutants. Je n'ai pas pu obtenir de tel résultats dans le temps imparti et je me suis donc attachée à essayer de trouver d'autres éléments de réponses à cette problématique. J'ai tout d'abord multiplié le nombre de contrôle (injection de sgRNAs scrambled, de sgRNAs ciblant *SAIYAN* et *TYR*). J'ai également réalisé cette approche pour les autres gènes *MSANTD*. Dans tous ces cas, les phénotypes observés pour les F0 *MSANTD2* n'ont pas été reproduits, soulignant la spécificité des résultats lors de l'utilisation des sgRNAs dirigés contre *MSANTD2*. En d'autres termes, ceci suggère dans un premier temps que les phénotypes observés ne sont probablement pas dû à un biais général d'expérimentation. Concernant plus spécifiquement la question des off-target, il est probable que s'ils existent, ils seront différents en fonction des sgRNAs utilisés. Les phénotypes associés devraient donc être différents selon le sgRNA. Or, j'ai pu reproduire les phénotypes avec différentes combinaisons de sgRNA et même des sgRNAs unique. Enfin, j'ai également utilisé la technique des morpholinos, qui est une technique de knock-down, avec laquelle j'ai pu reproduire les résultats. Tous ces éléments tendent à indiquer que le phénotype observé est spécifique à *MSANTD2*. De plus, cela est cohérent avec les données d'expression chez le poisson-zèbre et chez l'humain ainsi que l'association de *MSANTD2* avec des maladie neuro-développementales (Lim et al., 2017; O'Brien et al., 2018; Schizophrenia Working Group of the Psychiatric Genomics Consortium, 2014; Zhang et al., 2020).

D'autres expériences doivent être envisagé pour définitivement confirmer cette hypothèse. Il serait notamment pertinent de tester des expériences de sauvetage du phénotype. On pourrait, pour cela, co-injecter avec les sgRNAs ciblant *MSANTD2*, de l'ARNm de *MSANTD2*. Ceci pourrait

être réalisé de la même façon en co-injectant les morpholinos et de l'ARNm de *MSANTD2*. Si les phénotypes de mutants F0 *MSANTD2* sont bien spécifiques, ils devraient être au moins partiellement compensés par l'injection des ARNm. Pour finir, l'obtention d'une lignée stable de mutants *MSANTD2* permettra de conclure quant aux phénotypes présentés ici.

4.5 Encore d'autres gènes dérivés de transposons *Harbinger*?

Lors de ma thèse, j'ai donc identifié quatre nouveaux gènes dérivant de gènes *Myb-like* de transposons *Harbinger*. Dans les génomes de vertébrés, il existe de nombreux autres gènes codant pour des protéines possédant un domaine Myb (dont les domaines Myb-like font partie). Afin d'étudier un lien éventuel plus large entre transposons *Harbinger* et domaines Myb, j'ai comparé toutes les protéines humaines possédant un domaine Myb (675 protéines) avec les transposons *Harbinger* par blast (Altschul et al., 1990). Les résultats n'ont retourné de similarités significatives que pour dix gènes, dont les six gènes que j'ai déjà étudiés, suggérant plusieurs éléments : (1) tout d'abord, cela s'oppose à l'hypothèse d'une origine commune de tous les domaines Myb des protéines humaines à partir de transposons ; (2) ce faible nombre tend à supporter le fait que les gènes *MSANTD* dérivent effectivement de transposons *Harbinger* et que ce n'est pas un biais dû à quelques similarités non-homologues entre les transposons *Harbinger* et les domaines Myb ; (3) enfin, cela s'oppose également à l'hypothèse d'une acquisition d'un domaine Myb issu du génome humain par les transposons *Harbinger*, et ceci pourrait être vérifié dans d'autres espèces. Néanmoins, nous avons observé que les séquences *Myb-like* évoluent rapidement. Ainsi, des homologies plus anciennes pourraient exister, mais pourraient être aujourd'hui difficilement détectable par blast à cause de cette divergence rapide.

Le criblage des protéines humaines à domaine Myb a donc retourné dix gènes, six qui ont été présentés dans ce manuscrit ainsi que quatre autres : *MYPOP*, *ZSCAN20*, *TSNARE1* et *PRDM11*. Ainsi, ces gènes représentent quatre autres exemples potentiels de gènes dérivés de transposons *Harbinger* chez les vertébrés. Il est à noter que le gène *TSNARE1* a également été suggéré comme étant issu de la domestication moléculaire de transposon *Harbinger* par Smith et al. (2012). *MYPOP* est un répresseur transcriptionnel avec des capacités de liaison spécifique à l'ADN et à des protéines (Lederer et al., 2005). La protéine *ZSCAN20* interagit avec plusieurs protéines et pourrait être impliquée dans la régulation transcriptionnelle (genecard). *TSNARE1* est impliqué dans le transport endosomal de protéine, grâce à des interactions protéiques, et est un facteur de susceptibilité à la schizophrénie chez l'Homme (Plooster et al., 2021). *PRDM11* se lie à l'ADN, il pourrait être impliqué dans la régulation transcriptionnelle au vu de ces sites de fixation, et son inactivation chez la souris inhibe la prolifération cellulaire et induit l'apoptose (Fog et al., 2015). Les alignements et les phylogénies comparant ces protéines aux séquences de protéines Myb-like de transposons *Harbinger* tendent à conforter cette hypothèse (**Figures 4.1, 4.2, 4.3, 4.4**). Cependant, je n'ai pu étudier ces gènes que de façon préliminaire. Des analyses supplémentaires seront donc nécessaires pour établir de façon plus certaine leur origine, en particulier concernant le gène *ZSCAN20* qui ne semble pas présenter de groupement phylogénétique préférentiel avec des séquences de transposons *Harbinger* particulières.

En conclusion, grâce à cette thèse j'ai pu mettre en évidence des domestications moléculaires

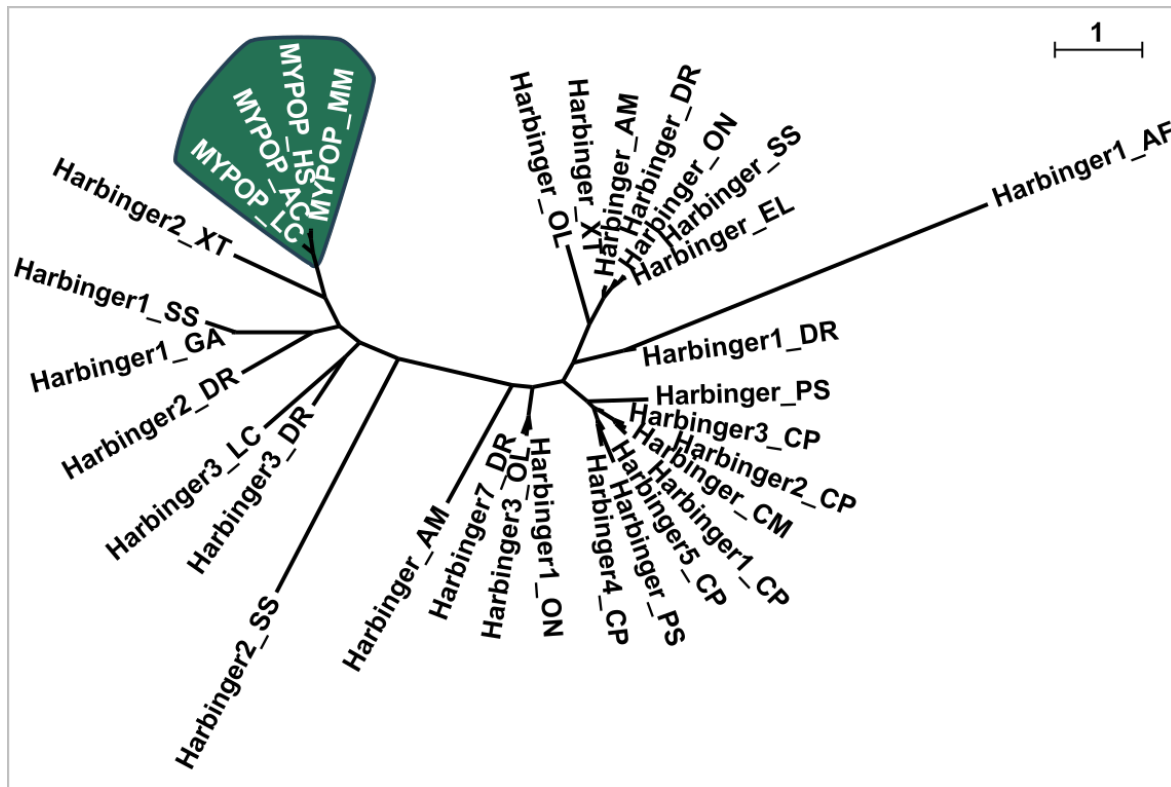


FIGURE 4.1 – Relations phylogénétiques entre les protéines Myb-like de transposons *Harbinger* et les protéines MYPOP provenant de plusieurs espèces de vertébrés. L'arbre a été construit en utilisant la méthode de maximum de vraisemblance (Guindon & Gascuel, 2003). (AC : *Anolis carolinensis*, AF : *African cichild*, AM : *Alligator mississippiensis*, CM : *Chelonia mydas*, CP : *Chrysemys picta*, DR : *Danio rerio*, EL : *Esox lucius*, GA : *Gasterosteus aculeatus*, HS : *Homo sapiens*, LC : *Latimeria chalumnae*, MM : *Mus musculus*, OL : *Oryzias latipes*, ON : *Oreochromis niloticus*, PS : *Pelodiscus sinensis*, SS : *Salmon salar*, XT : *Xenopus tropicalis*).

récurrentes de transposons *Harbinger* chez les vertébrés, ayant donné lieu à au moins six nouveaux gènes voire potentiellement dix.

Au-delà des vertébrés, les transposons *Harbinger* sont également à l'origine d'au moins 17 autres gènes. Neuf gènes dérivés de ces transposons ont été identifiés chez les plantes *Arabidopsis* (Duan et al., 2017; Liang et al., 2015; Velanis et al., 2020; Zhou et al., 2021). Casola et al. (2007) ont également décrit des domestications moléculaires récurrentes de ces transposons chez les *drosophiles*. Ils ont identifié huit gènes, sept dérivant de la transposase (DPLG1 à 7A) et un dérivant d'un gène *Myb-like* (DPLG7B) de transposons *Harbinger*.

Lors de mes analyses pour étudier la distribution des gènes *MSANTD*, j'ai pu identifier deux autres cas potentiels de domestications moléculaires de transposons *Harbinger* chez les *drosophiles*, que j'ai nommés *MSANTD1-like* et *NAIF2* en raison de leur similarité avec *MSANTD1* et *NAIF1*, respectivement. Malgré cette apparente similarité, il s'agit bien de gènes différents non-orthologues entre les vertébrés et les *drosophiles*. Ainsi, dans le cadre d'une collaboration avec le laboratoire de Cristina Vieira (LBBE, Université Claude Bernard Lyon 1), j'ai étudié un peu plus en détail les gènes issus de domestications moléculaires de transposons *Harbinger* chez les *drosophiles*. J'ai tout d'abord étudié la distribution de ces gènes, ce qui m'a permis d'étendre les résultats présentés par Casola et al. (2007) et d'y ajouter *MSANTD1-like* et *NAIF2* (Figure 4.5). Ainsi, j'ai pu observer que les gènes *DPLG1-4* sont présents chez toutes les espèces de *drosophiles* étudiées, suggérant

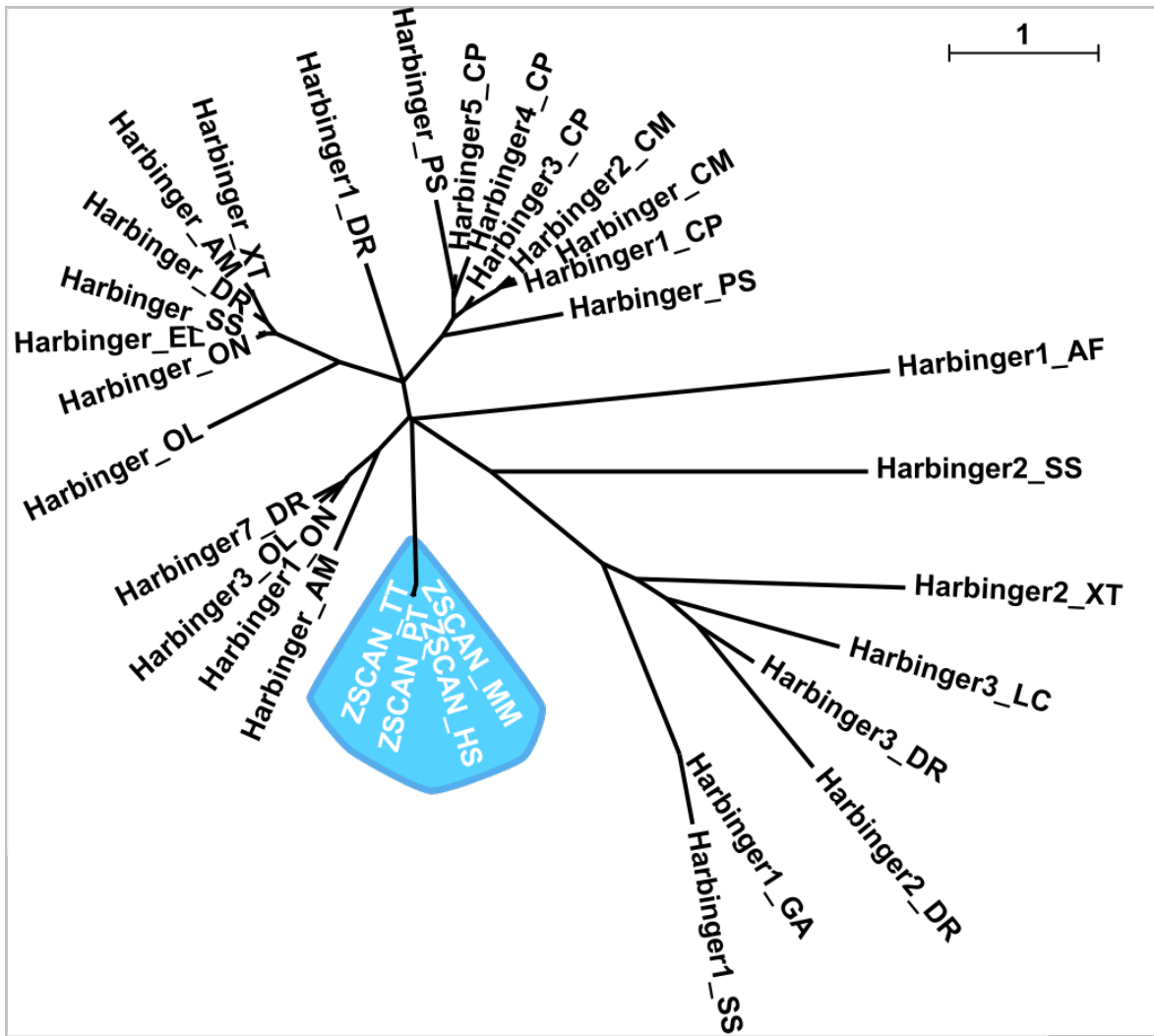


FIGURE 4.2 – Relations phylogénétiques entre les protéines Myb-like de transposons *Harbinger* et les protéines ZSCAN20 provenant de plusieurs espèces de vertébrés. L'arbre a été construit en utilisant la méthode de maximum de vraisemblance (Guindon & Gascuel, 2003). (AF : *African cichlid*, AM : *Alligator mississippiensis*, CM : *Chelonia mydas*, CP : *Chrysemys picta*, DR : *Danio rerio*, EL : *Esox lucius*, GA : *Gasterosteus aculeatus*, HS : *Homo sapiens*, LC : *Latimeria chalumnae*, MM : *Mus musculus*, OL : *Oryzias latipes*, ON : *Oreochromis niloticus*, PS : *Pelodiscus sinensis*, PT : *Pan troglodytes*, SS : *Salmon salar*, TT : *Tursiops truncatus*, XT : *Xenopus tropicalis*).

leur apparition dans l'ancêtre commun des drosophiles, soit il y a environ 120 millions d'années. *NAIF2* présente une distribution phylogénétique plus large encore suggérant sa domestication il y a environ 250 millions d'années. Les autres gènes seraient apparus plus récemment. *MSANTD1-like* pourrait provenir d'un événement de domestication ayant eu lieu il y a environ 30 millions d'années. Enfin, les gènes *DPLG5-7B* ont été identifiés dans un groupe d'espèces non monophylétique. Ceci suggère (1) que ces gènes ont été acquis dans l'ancêtre commun de ces espèces puis perdus massivement par la suite ou bien (2) que ces gènes ont été acquis dans certaines espèces par transfert horizontal. (Les datations des événements de domestication ont été estimés grâce à TimeTree (Kumar et al., 2017).)

Les alignements et les phylogénies comparant les gènes *MSANTD1-like* et *NAIF2* avec des séquences de protéines Myb-like de transposons *Harbinger* tendent à conforter les événements de domestications moléculaires (Figures 4.6 et 4.7).

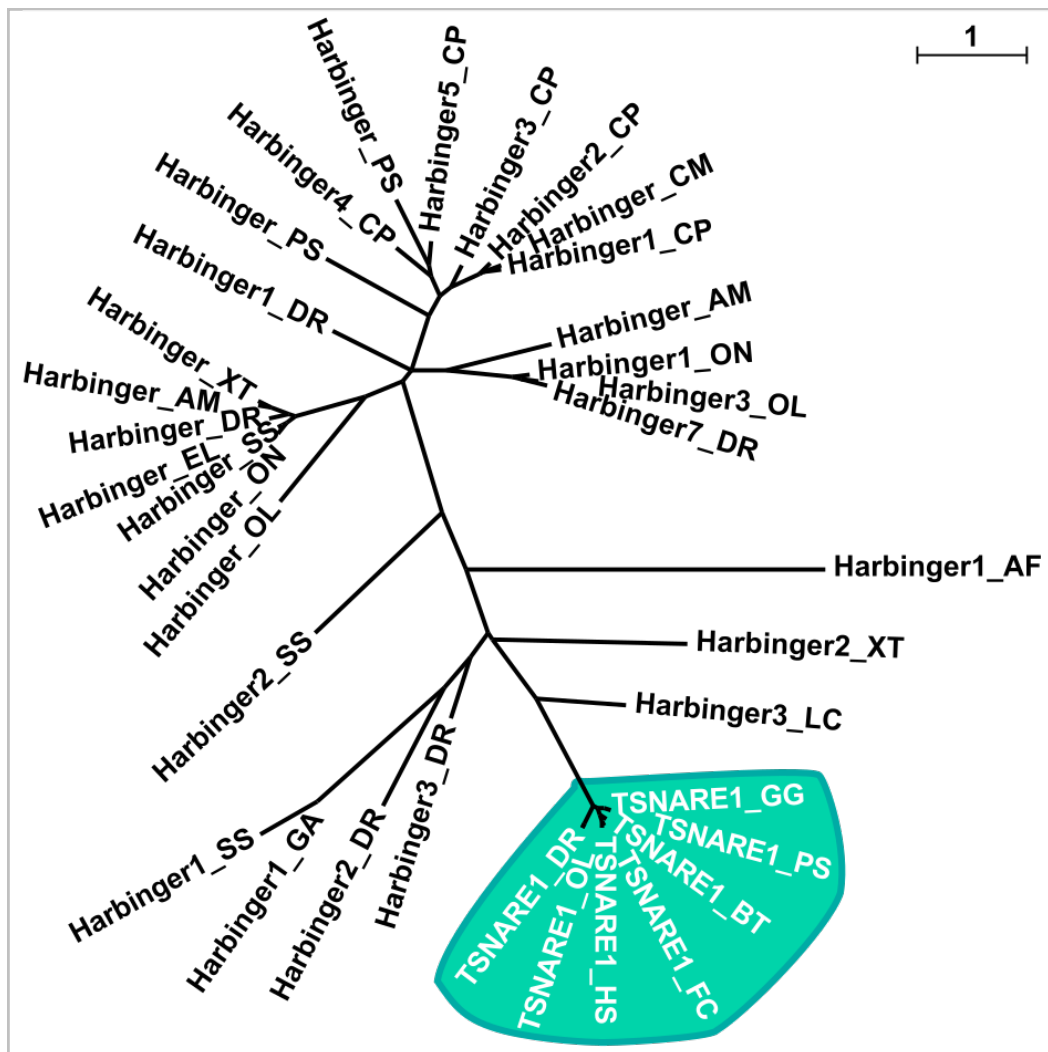


FIGURE 4.3 – Relations phylogénétiques entre les protéines Myb-like de transposons *Harbinger* et les protéines TSNARE1 provenant de plusieurs espèces de vertébrés. L'arbre a été construit en utilisant la méthode de maximum de vraisemblance (Guindon & Gascuel, 2003). (AF : *African cichlid*, AM : *Alligator mississippiensis*, BT : *Bos taurus*, CM : *Chelonia mydas*, CP : *Chrysemys picta*, DR : *Danio rerio*, EL : *Esox lucius*, FC : *Felis catus*, GA : *Gasterosteus aculeatus*, GG : *Gallus gallus*, HS : *Homo sapiens*, LC : *Latimeria chalumnae*, MM : *Mus musculus*, OL : *Oryzias latipes*, ON : *Oreochromis niloticus*, PS : *Pelodiscus sinensis*, SS : *Salmon salar*, XT : *Xenopus tropicalis*).

Ainsi, il apparaît que les éléments transposables de type *Harbinger* semblent avoir une grande propension à être recrutés en tant que nouveaux gènes dans les génomes hôtes. La caractéristique de ces transposons de posséder deux ORFs différents codant pour deux protéines aux activités différentes pouvant interagir entre elles, pourrait représenter un avantage. Ces deux ORFs codent pour des protéines avec des domaines possédant des activités moléculaires répandues et potentiellement utiles pour l'hôte. En particulier, la protéine contenant le domaine Myb-like, un domaine de liaison à l'ADN et à des protéines, pourrait être réutilisé de diverses façons pour la régulation de gènes, comme par exemple en tant que facteur de transcription ou membre d'un interactome protéique. De plus, la séparation des deux activités moléculaires du transposon (coupure/recombinaison de l'ADN et liaison à ADN/protéines) en deux ORFs indépendants n'est pas commune chez les éléments transposables. Ceci pourrait faciliter la co-optation plus spécifique d'une des deux activités moléculaires sans interférence de la deuxième.

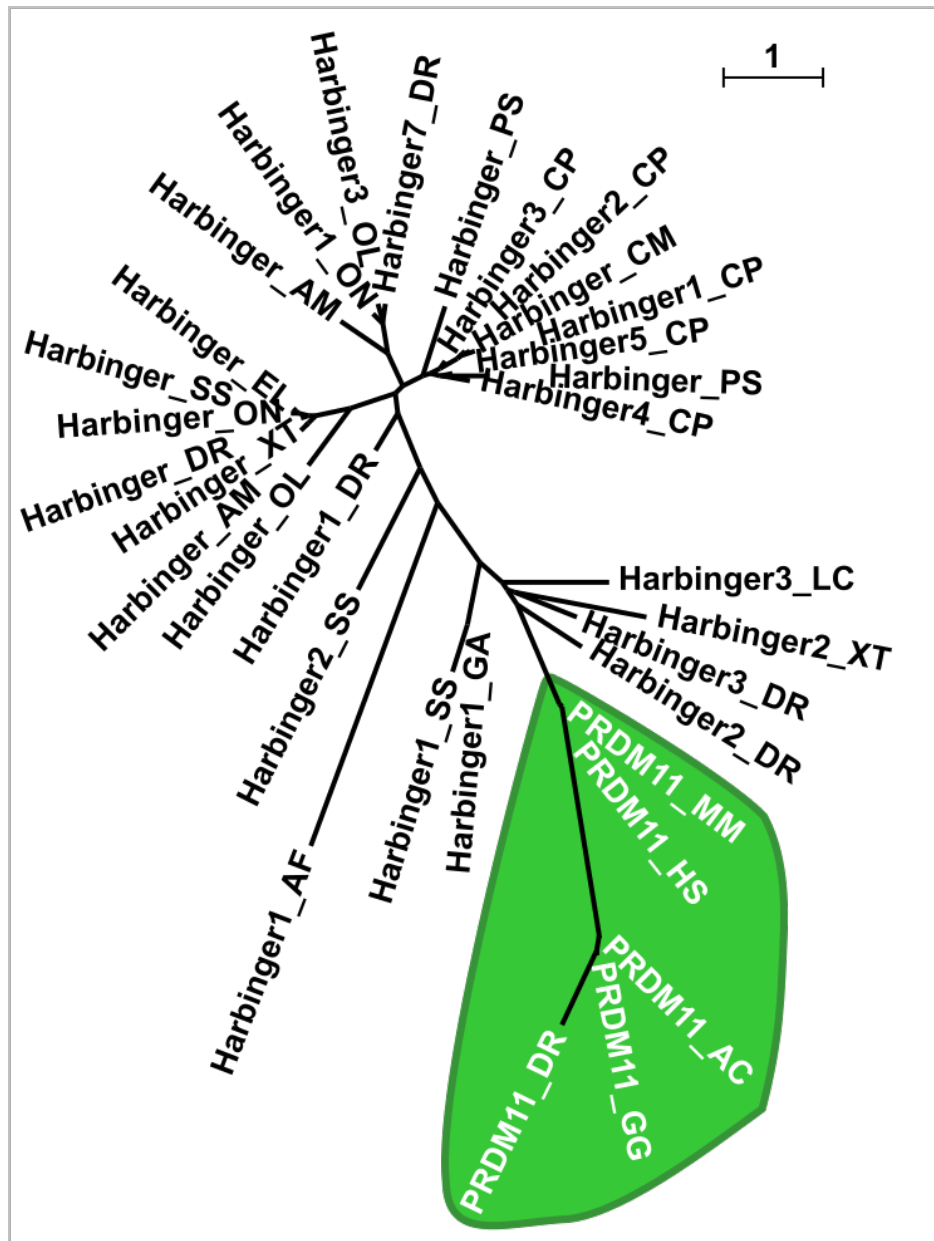


FIGURE 4.4 – Relations phylogénétiques entre les protéines Myb-like de transposons *Harbinger* et les protéines PRDM11 provenant de plusieurs espèces de vertébrés. L'arbre a été construit en utilisant la méthode de maximum de vraisemblance (Guindon & Gascuel, 2003). (AC : *Anolis carolinensis*, AF : *African cichild*, AM : *Alligator mississippiensis*, 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*, OL : *Oryzias latipes*, ON : *Oreochromis niloticus*, PS : *Pelodiscus sinensis*, SS : *Salmon salar*, XT : *Xenopus tropicalis*).

En conclusion, les résultats de cette thèse permettent de proposer les transposons *Harbinger* comme d'importants contributeurs à la formation de nouveaux gènes, et donc l'évolution et l'adaptation des espèces. En particulier, les gènes dérivés de transposons *Harbinger* ont probablement contribué à l'évolution des vertébrés, notamment via leur rôle dans le développement du système nerveux.

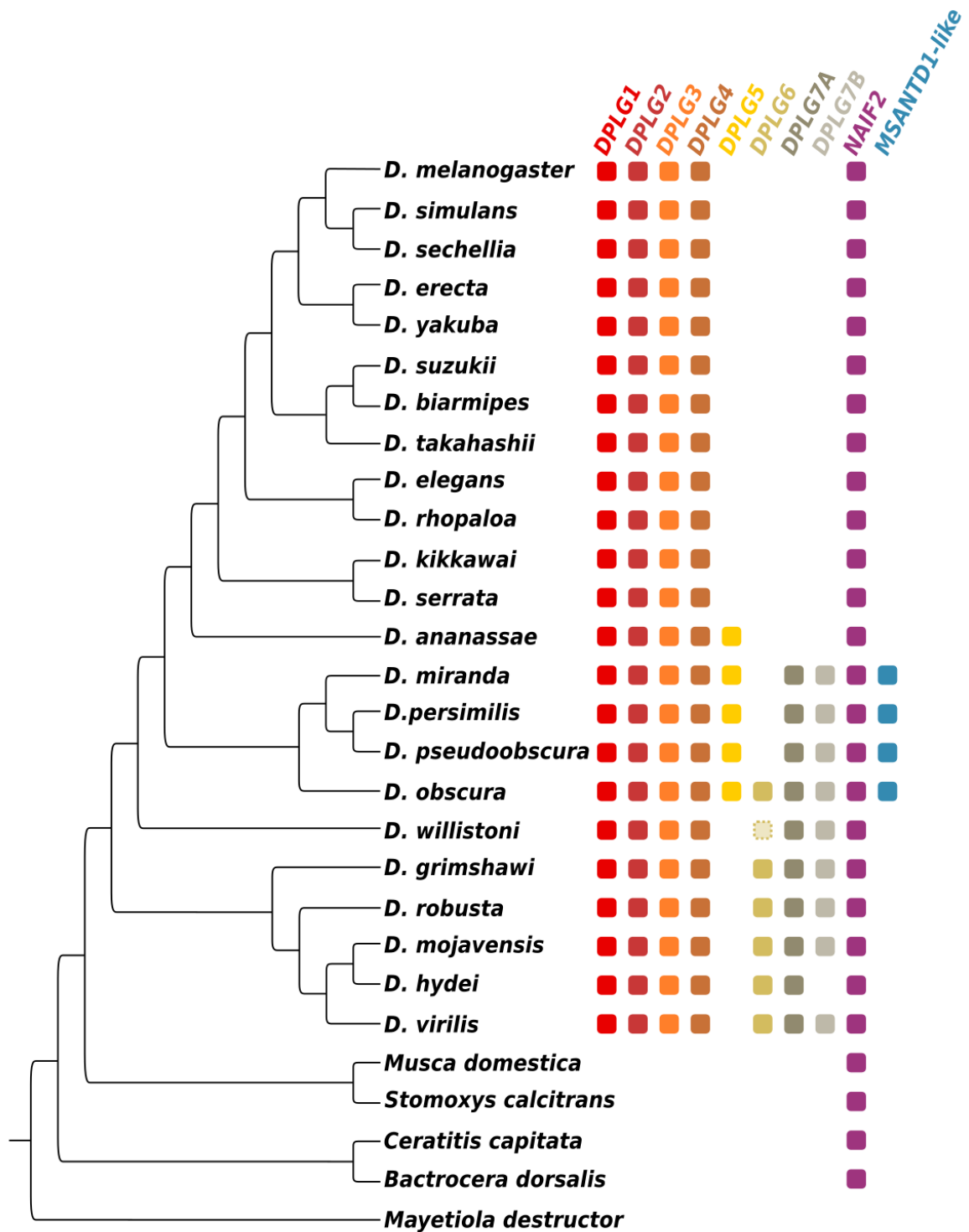


FIGURE 4.5 – Distribution phylogénétique des gènes dérivés de transposons *Harbinger* chez les drosophiles et d'autres mouches. La présence des gènes dans un taxon est représentée par un carré de couleur. Un carré au contour en pointillé indique que la séquence identifiée est dégénérée.

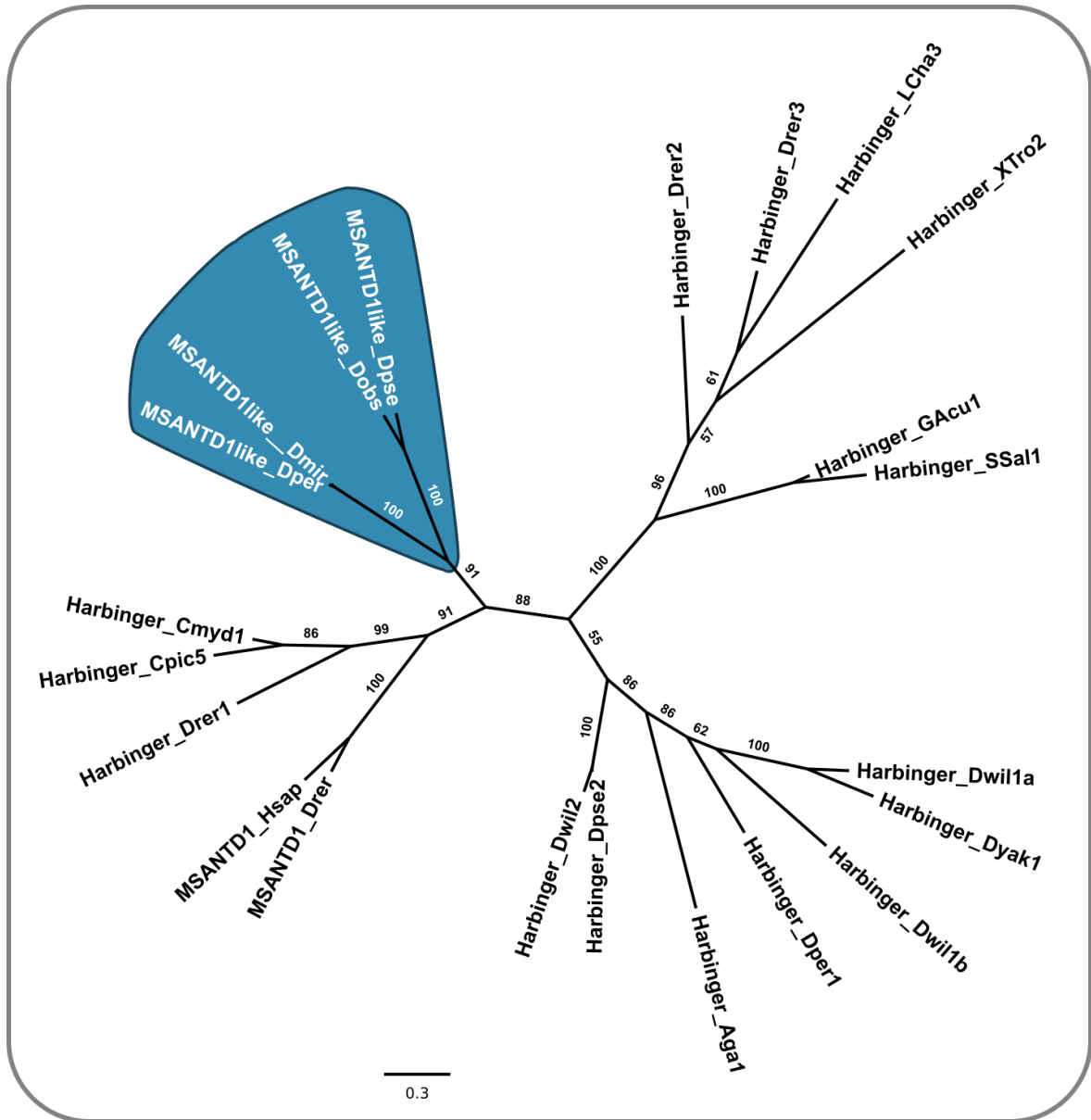


FIGURE 4.6 – Relations phylogénétiques entre les protéines Myb-like de transposons *Harbinger* et les protéines MSANTD1-like de diptères. L'arbre a été construit en utilisant la méthode Bayésienne (Huelsenbeck & Ronquist, 2001). (Aga : *Anopheles gambiae*, Cmyd : *Chelonia mydas*, Cpica : *Chrysemys picta*, Dmir : *Drosophila miranda*, Dobs : *Drosophila obscura*, Dper : *Drosophila persimilis*, Dpse : *Drosophila pseudoobscura*, Drer : *Danio rerio*, Dwil : *Drosophila willistoni*, Dyak : *Drosophila yakuba*, Gacu : *Gasterosteus aculeatus*, Hsap : *Homo sapiens*, Lcha : *Latimeria chalumnae*, Ssal : *Salmo salar*, Xtro : *Xenopus tropicalis*).

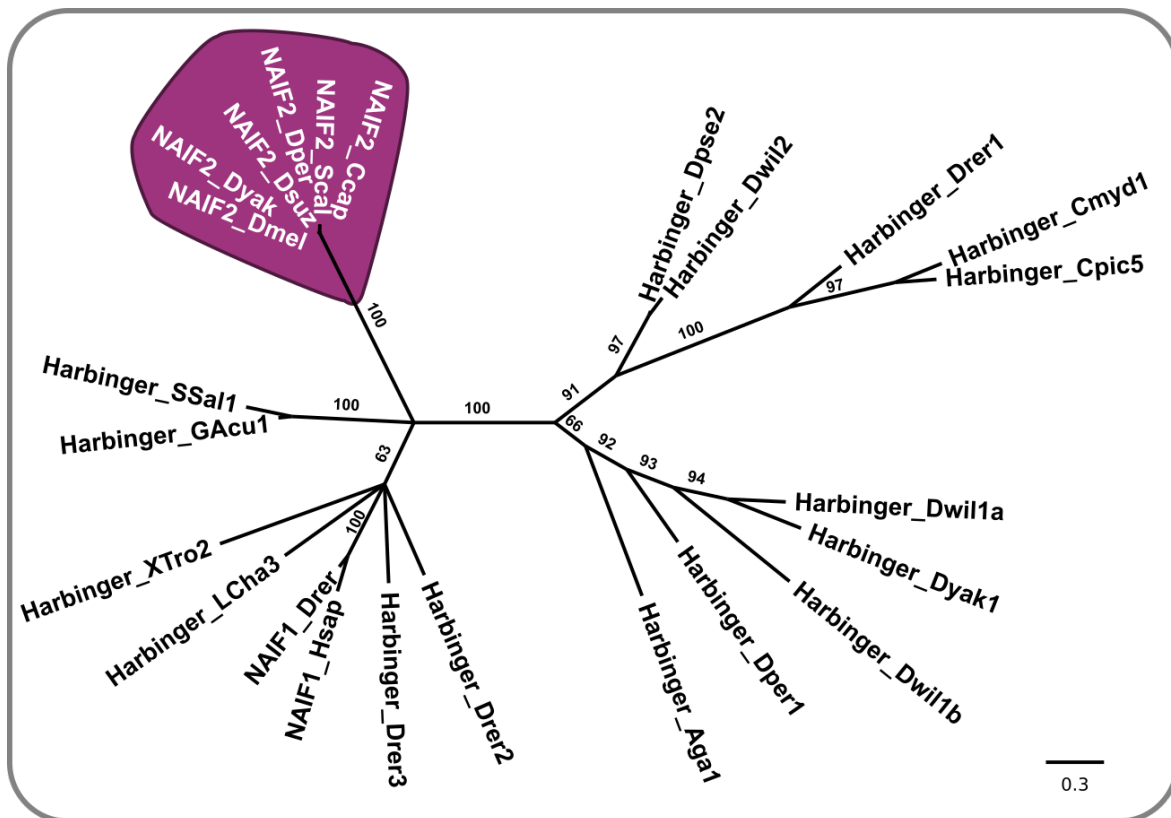


FIGURE 4.7 – Relations phylogénétiques entre les protéines Myb-like de transposons *Harbinger* et les protéines NAIF2 de diptères. L'arbre a été construit en utilisant la méthode Bayésienne (Huelsenbeck & Ronquist, 2001). (Aga : *Anopheles gambiae*, Cmyd : *Chelonia mydas*, Cpic : *Chrysemys picta*, Dmir : *Drosophila miranda*, Dobs : *Drosophila obscura*, Dper : *Drosophila persimilis*, Dpse : *Drosophila pseudoobscura*, Drer : *Danio rerio*, Dwil : *Drosophila willistoni*, Dyak : *Drosophila yakuba*, Gacu : *Gasterosteus aculeatus*, Hsap : *Homo sapiens*, Lcha : *Latimeria chalumnae*, Ssal : *Salmo salar*, Xtro : *Xenopus tropicalis*).

4.6 Conclusion

Dans le génome humain on estime actuellement à plus d'une centaine le nombre de gènes dérivés d'éléments transposables. Mes résultats de thèse indiquent que les transposons *Harbinger* auraient contribué à au moins six gènes chez les vertébrés, voire potentiellement dix au total chez l'humain. De plus, en dehors des vertébrés, d'autres gènes dérivés de transposons *Harbinger* ont été identifiés : neuf chez *Arabidopsis* et huit (voire dix) chez les *drosophiles*. Au vu de ce nombre de gène, le phénomène de domestication des éléments transposables *Harbinger* semble avoir grandement contribué à l'évolution et l'adaptation des espèces. Il paraît donc important d'étudier la possibilité d'autres événements de domestications moléculaires de transposon *Harbinger* chez d'autres espèces pour comprendre l'étendu du répertoire de gènes formés par ces éléments. De façon générale, plusieurs caractéristiques modulent la propension des éléments transposables à pouvoir être exaptés, notamment :

- des caractéristiques intrinsèques à chaque famille d'ETs, telles que la présence de promoteur, ou d'ORF
- des sites préférentiels d'insertion d'un élément modulant le contexte génomique dans lequel il est inséré
- l'activité, le nombre de copie et la diversité des éléments dans un génome.

Les caractéristiques connues des transposons *Harbinger* répondent à priori de façon efficace au premier de ces trois critères. En effet, ces transposons sont structurés en deux ORFs distincts aux activités moléculaires répandues et potentiellement utiles pour l'hôte. Mes résultats indiquent que les transposons *Harbinger* ont contribué de façon importante, à l'évolution précoce des vertébrés. Les résultats expérimentaux obtenus avec le gène *MSANTD2* corroborent cette hypothèse. La validation du rôle de *MSANTD2* dans le développement du système nerveux des vertébrés et la recherche de la fonction des autres gènes dérivés de transposons *Harbinger* permettront d'asseoir l'importance de ces éléments lors de l'évolution des vertébrés.

A

Annexes

Communications orales

Congrès National sur les Éléments Transposables (CNET) 2019

- Juillet 2019 - Lyon
- Molecular domestication of transposable elements in Vertebrates : Functional and evolutionary study of a gene family derived from Harbinger transposons

Séminaire interne de l'Institut de Génomique Fonctionnelle de Lyon

- Mai 2019 - Lyon
- Molecular domestication of transposable elements in Vertebrates : Functional and evolutionary study of a gene family derived from Harbinger transposons

Journées de l'école doctorale BMIC 2020

- Novembre 2020 - Lyon (visio-conférence)
- Molecular domestication of transposable elements in Vertebrates : Functional and evolutionary study of a gene family derived from Harbinger transposons

Séminaire interne de l'Institut de Génomique Fonctionnelle de Lyon

- Mars 2021 - Lyon (visio-conférence)
- Molecular domestication of transposable elements in Vertebrates : Functional and evolutionary study of a gene family derived from Harbinger transposons

Séminaires Talents de la génétique – Les lundis de la SFG (Société Française de Génétique)

- Mai 2021 - Paris (visio-conférence)
- Use of the Zebrafish model to study a new transposon-derived gene family involved in the development of the vertebrate nervous system

Congrès EFOR (Réseau d'Etudes Fonctionnelles chez les ORganismes modèles) meeting : zebrafish session

- Mai 2021 - Paris-Saclay (visio-conférence)
- Use of the Zebrafish model to study a new transposon-derived gene family involved in the development of the vertebrate nervous system

GDR Éléments transposables

- Septembre 2021 - Paris
- Transposable element molecular domestication in Vertebrates : Functional and evolutionary study of a gene family derived from Harbinger transposons

5th Uppsala Transposon Symposium

- Octobre 2021 - Uppsala, Suède (visio-conférence)
- Transposable element molecular domestication in Vertebrates : Functional and evolutionary study of a gene family derived from Harbinger transposons

Liste complète des références

- S. F. Altschul, W. Gish, W. Miller, E. W. Myers & D. J. Lipman (1990). Basic local alignment search tool. *Journal of Molecular Biology* **215**, 403–410. 149
- C. T. Amemiya, J. Alföldi, A. P. Lee, S. Fan, H. Philippe, I. Maccallum, I. Braasch, T. Manousaki et al. (2013). The African coelacanth genome provides insights into tetrapod evolution. *Nature* **496**, 311–316. 60, 62, 78
- S. Andreasen, S. Varma, N. Barasch, L. D. R. Thompson, M. Miettinen, L. Rooper, E. B. Stelow, T. K. Agander et al. (2019). The HTN3-MSANTD3 Fusion Gene Defines a Subset of Acinic Cell Carcinoma of the Salivary Gland. *The American Journal of Surgical Pathology* **43**, 489–496. 80
- N. Barasch, X. Gong, K. A. Kwei, S. Varma, J. Biscocho, K. Qu, N. Xiao, J. S. Lipsick et al. (2017). Recurrent rearrangements of the Myb/SANT-like DNA-binding domain containing 3 gene (MSANTD3) in salivary gland acinic cell carcinoma. *PLoS One* **12**, e0171265. 80
- A. Beauregard, M. J. Curcio & M. Belfort (2008). The take and give between retrotransposable elements and their hosts. *Annual Review of Genetics* **42**, 587–617. 14
- C. Biémont & C. Vieira (2006). Genetics : junk DNA as an evolutionary force. *Nature* **443**, 521–524. 13, 17
- G. Bourque (2009). Transposable elements in gene regulation and in the evolution of vertebrate genomes. *Current Opinion in Genetics & Development* **19**, 607–612. 17
- G. Bourque, K. H. Burns, M. Gehring, V. Gorbunova, A. Seluanov, M. Hammell, M. Imbeault, Z. Izsvák et al. (2018). Ten things you should know about transposable elements. *Genome Biology* **19**, 199. 13, 15, 17
- I. Braasch, A. R. Gehrke, J. J. Smith, K. Kawasaki, T. Manousaki, J. Pasquier, A. Amores, T. Desvignes et al. (2016). The spotted gar genome illuminates vertebrate evolution and facilitates human-teleost comparisons. *Nature Genetics* **48**, 427–437. 78
- M. Brand, M. Granato & C. Nüsslein-Volhard (2002). Keeping and raising zebrafish. In *Zebrafish*, pp. 7–37. 48
- M. D. Brazeau & M. Friedman (2015). The origin and early phylogenetic history of jawed vertebrates. *Nature* **520**, 490–497. 7
- E. Buglo, E. Sarmiento, N. B. Martuscelli, D. W. Sant, M. C. Danzi, A. J. Abrams, J. E. Dallman & S. Züchner (2020). Genetic compensation in a stable slc25a46 mutant zebrafish : A case for using F0 CRISPR mutagenesis to study phenotypes caused by inherited disease. *PLoS ONE* **15**, e0230566. 147, 148
- A. Böhne, F. Brunet, D. Galiana-Arnoux, C. Schultheis & J.-N. Volff (2008). Transposable elements as drivers of genomic and biological diversity in vertebrates. *Chromosome Research : An International Journal on the Molecular, Supramolecular and Evolutionary Aspects of Chromosome Biology* **16**, 203–215. 18
- R. K. Campbell, N. Satoh & B. M. Degnan (2004). Piecing together evolution of the vertebrate endocrine system. *Trends in genetics : TIG* **20**, 359–366. 11
- F. Carducci, M. Barucca, A. Canapa, E. Carotti & M. A. Biscotti (2020). Mobile Elements in Ray-Finned Fish Genomes. *Life (Basel, Switzerland)* **10**, E221. 18, 58
- C. Casola, A. M. Lawing, E. Betrán & C. Feschotte (2007). PIF-like transposons are common in drosophila and have been repeatedly domesticated to generate new host genes. *Molecular Biology and Evolution* **24**, 1872–1888. 16, 58, 71, 73, 145, 150
- V. Cazalis, M. Di Marco, S. H. M. Butchart, H. R. Akçakaya, M. González-Suárez, C. Meyer, V. Clausnitzer, M. Böhm et al. (2022). Bridging the research-implementation gap in IUCN Red List assessments. *Trends in Ecology & Evolution* **37**, 359–370. 3
- D. Chalopin, M. Naville, F. Plard, D. Galiana & J.-N. Volff (2015). Comparative analysis of transposable elements highlights mobilome diversity and evolution in vertebrates. *Genome Biology and Evolution* **7**, 567–580. 18, 19, 20, 58, 60, 62
- N. Chang, C. Sun, L. Gao, D. Zhu, X. Xu, X. Zhu, J.-W. Xiong & J. J. Xi (2013). Genome editing with RNA-guided Cas9 nuclease in Zebrafish embryos. *Cell Research* **23**, 465–472. 49
- B. Chenais (2015). Transposable elements in cancer and other human diseases. *Current Cancer Drug Targets* **15**, 227–242. 17
- E. B. Chuong, N. C. Elde & C. Feschotte (2016). Regulatory evolution of innate immunity through co-option of endogenous retroviruses. *Science* **351**, 1083–1087. 58
- E. B. Chuong, N. C. Elde & C. Feschotte (2017). Regulatory activities of transposable elements : from conflicts to benefits. *Nature Reviews. Genetics* **18**, 71–86. 58, 73
- J. M. Coffin (1992). Structure and Classification of Retroviruses. In *The Retroviridae* (J. A. Levy, ed.), The Viruses, pp. 19–49. Springer US, Boston, MA. 14
- R. L. Cosby, N.-C. Chang & C. Feschotte (2019). Host-transposon interactions : conflict, cooperation, and cooption. *Genes & Development* **33**, 1098–1116. 73
- R. L. Cosby, J. Judd, R. Zhang, A. Zhong, N. Garry, E. J. Pritham & C. Feschotte (2021). Recurrent evolution of vertebrate transcription factors by transposase capture. *Science (New York, N.Y.)* **371**, eabc6405. 73

- M. J. Curcio & K. M. Derbyshire (2003). The outs and ins of transposition : from mu to kangaroo. *Nature Reviews Molecular Cell Biology* **4**, 865–877. 14, 16
- R. Dahm (2002). Atlas of embryonic stages of development in the zebrafish. In *Zebrafish*, pp. 219–236. 48, 51, 52
- D. Darriba, G. L. Taboada, R. Doallo & D. Posada (2011). Prot-Test 3 : fast selection of best-fit models of protein evolution. *Bioinformatics (Oxford, England)* **27**, 1164–1165. 61
- E. M. De Robertis, J. Larraín, M. Oelgeschläger & O. Wessely (2000). The establishment of Spemann's organizer and patterning of the vertebrate embryo. *Nature Reviews. Genetics* **1**, 171–181. 6
- D. M. de Vienne, T. Giraud & O. C. Martin (2007). A congruence index for testing topological similarity between trees. *Bioinformatics (Oxford, England)* **23**, 3119–3124. 68
- C. Dechaud, S. Miyake, A. Martinez-Bengochea, M. Scharl, J.-N. Volf & M. Naville (2021). Clustering of Sex-Biased Genes and Transposable Elements in the Genome of the Medaka Fish *Oryzias latipes*. *Genome Biology and Evolution* **13**, evab230. 61
- P. Dehal & J. L. Boore (2005). Two rounds of whole genome duplication in the ancestral vertebrate. *PLoS biology* **3**, e314. 12, 142
- P. L. Deininger, J. V. Moran, M. A. Batzer & H. H. Kazazian (2003). Mobile elements and mammalian genome evolution. *Current Opinion in Genetics & Development* **13**, 651–658. 18
- J. Deneweth, Y. Van de Peer & V. Vermeirssen (2022). Nearby transposable elements impact plant stress gene regulatory networks : a meta-analysis in *A. thaliana* and *S. lycopersicum*. *BMC genomics* **23**, 18. 73
- Deniz, J. M. Frost & M. R. Branco (2019). Regulation of transposable elements by DNA modifications. *Nature Reviews Genetics* . 15
- M. Dewannieux, C. Esnault & T. Heidmann (2003). LINE-mediated retrotransposition of marked Alu sequences. *Nature Genetics* **35**, 41–48. 16
- M. R. Dietrich, R. A. Ankeny & P. M. Chen (2014). Publication Trends in Model Organism Research. *Genetics* **198**, 787–794. 50
- W. F. Doolittle & C. Sapienza (1980). Selfish genes, the phenotype paradigm and genome evolution. *Nature* **284**, 601–603. 13, 17
- L. C. Drickamer (1981). Selection for age of sexual maturation in mice and the consequences for population regulation. *Behavioral and Neural Biology* **31**, 82–89. 48
- C.-G. Duan, X. Wang, S. Xie, L. Pan, D. Miki, K. Tang, C.-C. Hsu, M. Lei et al. (2017). A pair of transposon-derived proteins function in a histone acetyltransferase complex for active DNA demethylation. *Cell Research* **27**, 226–240. 73, 150
- C. Duchamp, J. L. Rouanet & H. Barré (2002). Ontogeny of thermoregulatory mechanisms in king penguin chicks (*Aptenodytes patagonicus*). *Comparative Biochemistry and Physiology. Part A, Molecular & Integrative Physiology* **131**, 765–773. 3
- M. A. El-Brolosy & D. Y. R. Stainier (2017). Genetic compensation : A phenomenon in search of mechanisms. *PLoS genetics* **13**, e1006780. 147, 148
- A. Elewa, H. Wang, C. Talavera-López, A. Joven, G. Brito, A. Kumar, L. S. Hameed, M. Penrad-Mobayed et al. (2017). Reading and editing the *Pleurodeles waltl* genome reveals novel features of tetrapod regeneration. *Nature Communications* **8**, 2286. 59, 69
- E. Etchegaray, M. Naville, J.-N. Volf & Z. Haftek-Terreau (2021). Transposable element-derived sequences in vertebrate development. *Mobile DNA* **12**, 1. 21, 73
- S. H. Fatemi (2005). Reelin glycoprotein : structure, biology and roles in health and disease. *Molecular Psychiatry* **10**, 251–257. 146
- N. V. Fedoroff (2012). Transposable Elements, Epigenetics, and Genome Evolution. *Science* **338**, 758–767. 17
- C. Feschotte & E. J. Pritham (2007). DNA Transposons and the Evolution of Eukaryotic Genomes. *Annual review of genetics* **41**, 331–368. 16, 18
- C. K. Fog, F. Asmar, C. Côme, K. T. Jensen, J. V. Johansen, T. B. Kheir, L. Jacobsen, C. Friis et al. (2015). Loss of PRDM11 promotes MYC-driven lymphomagenesis. *Blood* **125**, 1272–1281. 149
- R. Freitas, G. Zhang & M. J. Cohn (2006). Evidence that mechanisms of fin development evolved in the midline of early vertebrates. *Nature* **442**, 1033–1037. 11
- R. Fricke, W. N. Eschmeyer & R. Van der Laan (2022). *Eschmeyer's Catalog of Fishes : Genera, Species, References*. Electronic version 2022 edn. 58
- B. Fritsch & R. G. Northcutt (1993). Cranial and spinal nerve organization in amphioxus and lampreys : evidence for an ancestral craniate pattern. *Acta Anatomica* **148**, 96–109. 11
- Y. Fu & E. Cao (2015). MicroRNA-125a-5p regulates cancer cell proliferation and migration through NAIF1 in prostate carcinoma. *OncoTargets and Therapy* **8**, 3827–3835. 80
- J. D. Galbraith, A. J. Ludington, K. L. Sanders, A. Suh & D. L. Adelson (2021). Horizontal transfer and subsequent explosive expansion of a DNA transposon in sea kraits (*Laticauda*). *Biology Letters* **17**, 20210342. 73, 143
- C. Gans & R. G. Northcutt (1983). Neural crest and the origin of vertebrates : a new head. *Science (New York, N.Y.)* **220**, 268–273. 7
- A. Generous, M. Thorson, J. Barcus, J. Jacher, M. Busch & H. Sleister (2014). Identification of putative interactions between swine and human influenza A virus nucleoprotein and human host proteins. *Virology Journal* **11**, 228. 80
- E. Glaw, J. Köhler, O. Hawlitschek, F. M. Ratsoavina, A. Rakotoarison, M. D. Scherz & M. Vences (2021). Extreme miniaturization of a new amniote vertebrate and insights into the evolution of genital size in chameleons. *Scientific Reports* **11**, 2522. 3
- J. L. Goodier (2016). Restricting retrotransposons : a review. *Mobile DNA* **7**, 16. 14

- I. Grabundzija, S. A. Messing, J. Thomas, R. L. Cosby, I. Bilic, C. Miskey, A. Gogol-Döring, V. Kapitonov et al. (2016). A Helitron transposon reconstructed from bats reveals a novel mechanism of genome shuffling in eukaryotes. *Nature Communications* **7**, 10716. 16
- T. Grocott, M. Tambalo & A. Streit (2012). The peripheral sensory nervous system in the vertebrate head : a gene regulatory perspective. *Developmental Biology* **370**, 3–23. 10
- D. Grzebelus, S. Lasota, T. Gambin, G. Kucherov & A. Gambin (2007). Diversity and structure of PIF/Harbinger-like elements in the genome of *Medicago truncatula*. *BMC genomics* **8**, 409. 16, 58
- R. Guerrini & E. Parrini (2010). Neuronal migration disorders. *Neurobiology of Disease* **38**, 154–166. 146
- S. Guindon & O. Gascuel (2003). A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Systematic Biology* **52**, 696–704. 60, 64, 68, 75, 150, 151, 152, 153
- L. Guio & J. González (2019). New Insights on the Evolution of Genome Content : Population Dynamics of Transposable Elements in Flies and Humans. *Methods in Molecular Biology (Clifton, N.J.)* **1910**, 505–530. 13
- B. K. Hall (2008). *The Neural Crest and Neural Crest Cells in Vertebrate Development and Evolution*. Springer Science & Business Media. Google-Books-ID : BfGzkeohtuEC. 7
- M.-J. Han, H.-E. Xu, H.-H. Zhang, C. Feschotte & Z. Zhang (2014). Spy : A New Group of Eukaryotic DNA Transposons without Target Site Duplications. *Genome Biology and Evolution* **6**, 1748–1757. 59, 72
- M.-J. Han, C.-L. Xiong, H.-B. Zhang, M.-Q. Zhang, H.-H. Zhang & Z. Zhang (2015). The diversification of PHIS transposon superfamily in eukaryotes. *Mobile DNA* **6**, 12. 16, 58, 59, 62, 64, 71, 72
- C. N. Hancock, F. Zhang & S. R. Wessler (2010). Transposition of the Tourist-MITE mPing in yeast : an assay that retains key features of catalysis by the class 2 PIF/Harbinger superfamily. *Mobile DNA* **1**, 5. 59
- T. Hirasawa & S. Kuratani (2015). Evolution of the vertebrate skeleton : morphology, embryology, and development. *Zoological Letters* **1**, 2. 11
- S. A. Hocknull, M. Wilkinson, R. A. Lawrence, V. Konstantinov, S. Mackenzie & R. Mackenzie (2021). A new giant sauropod, *Australotitan cooperensis* gen. et sp. nov., from the mid-Cretaceous of Australia. *PeerJ* **9**, e11317. 3
- L. Z. Holland (2009). Chordate roots of the vertebrate nervous system : expanding the molecular toolkit. *Nature Reviews. Neuroscience* **10**, 736–746. 10
- L. Z. Holland & N. D. Holland (2001). Evolution of neural crest and placodes : amphioxus as a model for the ancestral vertebrate? *Journal of Anatomy* **199**, 85–98. 10
- P. W. Holland, J. Garcia-Fernández, N. A. Williams & A. Sidow (1994). Gene duplications and the origins of vertebrate development. *Development (Cambridge, England). Supplement* pp. 125–133. 12
- N. Hopwood (2007). A history of normal plates, tables and stages in vertebrate embryology. *The International journal of developmental biology* **51**, 1–26. 5
- N. Hopwood (2011). Approaches and Species in the History of Vertebrate Embryology. In *Vertebrate Embryogenesis : Embryological, Cellular, and Genetic Methods* (F. J. Pelegri, ed.), Methods in Molecular Biology, pp. 1–20. Humana Press, Totowa, NJ. 5
- K. Hoshijima, M. J. Juryec, D. Klatt Shaw, A. M. Jacobi, M. A. Behlke & D. J. Grunwald (2019). Highly Efficient CRISPR-Cas9-Based Methods for Generating Deletion Mutations and F0 Embryos that Lack Gene Function in Zebrafish. *Developmental Cell* **51**, 645–657.e4. 85, 147, 148
- K. Howe, M. D. Clark, C. F. Torroja, J. Torrance, C. Berthelot, M. Mufato, J. E. Collins, S. Humphray et al. (2013). The zebrafish reference genome sequence and its relationship to the human genome. *Nature* **496**, 498–503. 48
- J. P. Huelsenbeck & F. Ronquist (2001). MRBAYES : Bayesian inference of phylogenetic trees. *Bioinformatics (Oxford, England)* **17**, 754–755. 60, 64, 68, 155, 156
- W. Y. Hwang, Y. Fu, D. Reyon, M. L. Maeder, P. Kaini, J. D. Sander, J. K. Joung, R. T. Peterson et al. (2013a). Heritable and Precise Zebrafish Genome Editing Using a CRISPR-Cas System. *PLoS ONE* **8**, e68708. 49
- W. Y. Hwang, Y. Fu, D. Reyon, M. L. Maeder, S. Q. Tsai, J. D. Sander, R. T. Peterson, J.-R. J. Yeh et al. (2013b). Efficient genome editing in zebrafish using a CRISPR-Cas system. *Nature Biotechnology* **31**, 227–229. 49, 80
- A. V. Igoshin, A. A. Yurchenko, N. M. Belonogova, D. V. Petrovsky, R. B. Aitnazarov, V. A. Soloshenko, N. S. Yudin & D. M. Larkin (2019). Genome-wide association study and scan for signatures of selection point to candidate genes for body temperature maintenance under the cold stress in Siberian cattle populations. *BMC genetics* **20**, 26. 80
- Y. W. Iwasaki, M. C. Siomi & H. Siomi (2015). PIWI-Interacting RNA : Its Biogenesis and Functions. *Annual Review of Biochemistry* **84**, 405–433. 72
- D. Jangam, C. Feschotte & E. Betrán (2017). Transposable Element Domestication As an Adaptation to Evolutionary Conflicts. *Trends in Genetics* **33**, 817–831. 58
- P. Janvier (2011). Comparative anatomy : all vertebrates do have vertebrae. *Current biology : CB* **21**, R661–663. 3, 7
- A. R. Jeffs, S. M. Benjes, T. L. Smith, S. J. Sowerby & C. M. Morris (1998). The BCR gene recombines preferentially with Alu elements in complex BCR-ABL translocations of chronic myeloid leukaemia. *Human Molecular Genetics* **7**, 767–776. 17
- N. Jiang, Z. Bao, X. Zhang, H. Hirochika, S. R. Eddy, S. R. McCouch & S. R. Wessler (2003). An active DNA transposon family in rice. *Nature* **421**, 163–167. 16, 58, 71
- J. Jurka & V. V. Kapitonov (2001). PIFs meet Tourists and Harbingers : A superfamily reunion. *Proceedings of the National Academy of Sciences of the United States of America* **98**, 12315–12316. 16, 58

- H. Kaessmann (2010). Origins, evolution, and phenotypic impact of new genes. *Genome Research* **20**, 1313–1326. 12
- R. Kalendar, C. M. Vicient, O. Peleg, K. Anamthawat-Jonsson, A. Bolshoy & A. H. Schulman (2004). Large retrotransposon derivatives : abundant, conserved but nonautonomous retroelements of barley and related genomes. *Genetics* **166**, 1437–1450. 15
- V. V. Kapitonov & J. Jurka (1999). Molecular paleontology of transposable elements from *Arabidopsis thaliana*. *Genetica* **107**, 27–37. 16, 58
- V. V. Kapitonov & J. Jurka (2004). Harbinger transposons and an ancient HARBI1 gene derived from a transposase. *DNA and cell biology* **23**, 311–324. 16, 58, 59, 62, 64, 71, 73, 79, 144
- V. V. Kapitonov & J. Jurka (2006). Self-synthesizing DNA transposons in eukaryotes. *Proceedings of the National Academy of Sciences* **103**, 4540–4545. 16
- V. V. Kapitonov & J. Jurka (2007). Helitrons on a roll : eukaryotic rolling-circle transposons. *Trends in Genetics* **23**, 521–529. 16
- V. V. Kapitonov & J. Jurka (2008). A universal classification of eukaryotic transposable elements implemented in Repbase. *Nature Reviews. Genetics* **9**, 411–412; author reply 414. 14, 16, 60
- K. Katoh, K. Misawa, K.-i. Kuma & T. Miyata (2002). MAFFT : a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Research* **30**, 3059–3066. 60
- M. Kaucka & I. Adameyko (2019). Evolution and development of the cartilaginous skull : From a lancelet towards a human face. *Seminars in Cell & Developmental Biology* **91**, 2–12. 11
- M. H. Kaufman (1992). *Atlas of mouse development*. Academic press edn. 48
- H. H. Kazazian (2004). Mobile elements : drivers of genome evolution. *Science (New York, N.Y.)* **303**, 1626–1632. 13, 18
- O. Khaner (2007). Evolutionary innovations of the vertebrates. *Integrative Zoology* **2**, 60–67. 7
- M. G. Kidwell & D. R. Lisch (2000). Transposable elements and host genome evolution. *Trends in Ecology & Evolution* **15**, 95–99. 58
- K. Kikuchi, K. Terauchi, M. Wada & H.-Y. Hirano (2003). The plant MITE mPing is mobilized in anther culture. *Nature* **421**, 167–170. 16, 58
- C. B. Kimmel, W. W. Ballard, S. R. Kimmel, B. Ullmann & T. F. Schilling (1995). Stages of embryonic development of the zebrafish. *Developmental Dynamics : An Official Publication of the American Association of Anatomists* **203**, 253–310. 48, 51, 87
- S. Kneitz, R. R. Mishra, D. Chalopin, J. Postlethwait, W. C. Warren, R. B. Walter & M. Schartl (2016). Germ cell and tumor associated piRNAs in the medaka and *Xiphophorus* melanoma models. *BMC genomics* **17**, 357. 72
- K. K. Kojima & J. Jurka (2011). Crypton transposons : identification of new diverse families and ancient domestication events. *Mobile DNA* **2**, 12. 16
- D. Kong & Z. Zhang (2018). NAIF1 suppresses osteosarcoma progression and is regulated by miR-128. *Cell Biochemistry and Function* **36**, 443–449. 80
- D. Kordis (2009). Transposable elements in reptilian and avian (sauropsida) genomes. *Cytogenetic and Genome Research* **127**, 94–111. 18
- M. E. Kossack & B. W. Draper (2019). Genetic regulation of sex determination and maintenance in zebrafish (*Danio rerio*). *Current topics in developmental biology* **134**, 119–149. 52
- M. Kottelat, R. Britz, T. H. Hui & K.-E. Witte (2006). *Paedocypris*, a new genus of Southeast Asian cyprinid fish with a remarkable sexual dimorphism, comprises the world's smallest vertebrate. *Proceedings of the Royal Society B : Biological Sciences* **273**, 895–899. 3
- F. Kroll, G. T. Powell, M. Ghosh, G. Gestri, P. Antinucci, T. J. Hearn, H. Tunbak, S. Lim et al. (2021). A simple and effective F0 knockout method for rapid screening of behaviour and other complex phenotypes. *eLife* **10**, e59683. 85, 147, 148
- M. Krupovic & E. V. Koonin (2015). Polintons : a hotbed of eukaryotic virus, transposon and plasmid evolution. *Nature Reviews Microbiology* **13**, 105–115. 16
- A. Kumar & J. L. Bennetzen (1999). Plant retrotransposons. *Annual Review of Genetics* **33**, 479–532. 14
- S. Kumar, G. Stecher, M. Suleski & S. B. Hedges (2017). TimeTree : A Resource for Timelines, Timetrees, and Divergence Times. *Molecular Biology and Evolution* **34**, 1812–1819. 63, 151
- E. S. Lander, L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody, J. Baldwin, K. Devon, K. Dewar et al. (2001). Initial sequencing and analysis of the human genome. *Nature* **409**. 12, 13, 17
- C. Lavialle, G. Cornelis, A. Dupressoir, C. Esnault, O. Heidmann, C. Vernochet & T. Heidmann (2013). Paleovirology of 'syncytins', retroviral *env* genes exapted for a role in placentation. *Philosophical Transactions of the Royal Society B : Biological Sciences* **368**, 20120507. 145
- M. Lederer, B. M. Jockusch & M. Rothkegel (2005). Profilin regulates the activity of p42POP, a novel Myb-related transcription factor. *Journal of Cell Science* **118**, 331–341. 149
- D. Y. Lee, K. J. Brayer, Y. Mitani, E. A. Burns, P. H. Rao, D. Bell, M. D. Williams, R. Ferrarotto et al. (2020). Oncogenic Orphan Nuclear Receptor NR4A3 Interacts and Cooperates with MYB in Acinic Cell Carcinoma. *Cancers* **12**, E2433. 80
- M. Li, L. Zhao, P. S. Page-McCaw & W. Chen (2016). Zebrafish Genome Engineering Using the CRISPR-Cas9 System. *Trends in genetics : TIG* **32**, 815–827. 80, 81
- S. C. Liang, B. Hartwig, P. Perera, S. Mora-García, E. de Leau, H. Thornton, F. de Lima Alves, F. L. de Alves 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 genetics* **11**, e1005660. 73, 150

- E. T. Lim, M. Uddin, S. De Rubeis, Y. Chan, A. S. Kamumbu, X. Zhang, A. M. D’Gama, S. N. Kim et al. (2017). Rates, distribution and implications of postzygotic mosaic mutations in autism spectrum disorder. *Nature Neuroscience* **20**, 1217–1224. 79, 146, 148
- T. Lorin (2018). Les cellules des crêtes neurales : « la seule chose intéressante » chez les Vertébrés? *Planet-Vie* . 4, 8, 9
- Q. Luo, M. Zhao, J. Zhong, Y. Ma, G. Deng, J. Liu, J. Wang, X. Yuan et al. (2011). NAIF1 is down-regulated in gastric cancer and promotes apoptosis through the caspase-9 pathway in human MKN45 cells. *Oncology Reports* **25**, 1117–1123. 80
- J. Mariette, C. Noirot, I. Nabihoudine, P. Bardou, C. Hoede, A. Djari, C. Cabau & C. Klopp (2014). RNAbrowse : RNA-Seq de novo assembly results browser. *PLoS One* **9**, e96821. 61
- D. N. Markova & R. J. Mason-Gamer (2015). Diversity, abundance, and evolutionary dynamics of Pong-like transposable elements in Triticeae. *Molecular Phylogenetics and Evolution* **93**, 318–330. 16, 58, 64, 71
- D. N. Markova & R. J. Mason-Gamer (2017). Transcriptional activity of PIF and Pong-like Class II transposable elements in Triticeae. *BMC evolutionary biology* **17**, 178. 59, 69
- B. McClintock (1956). Controlling elements and the gene. *Cold Spring Harbor Symposia on Quantitative Biology* **21**, 197–216. 12
- S. E. McCaugh, J. B. Gross, B. Aken, M. Blin, R. Borowsky, D. Chalopin, H. Hinaux, W. R. Jeffery et al. (2014). The cavefish genome reveals candidate genes for eye loss. *Nature Communications* **5**, 5307. 78
- B. Morse, P. G. Rotherg, V. J. South, J. M. Spandorfer & S. M. Astrin (1988). Insertional mutagenesis of the myc locus by a LINE-1 sequence in a human breast carcinoma. *Nature* **333**, 87–90. 17
- K. Muraki & K. Tanigaki (2015). Neuronal migration abnormalities and its possible implications for schizophrenia. *Frontiers in Neuroscience* **9**, 74. 146
- J. H. Nadeau (2001). Modifier genes in mice and humans. *Nature Reviews. Genetics* **2**, 165–174. 148
- N. Narita, H. Nishio, Y. Kitoh, Y. Ishikawa, Y. Ishikawa, R. Minami, H. Nakamura & M. Matsuo (1993). Insertion of a 5’ truncated L1 element into the 3’ end of exon 44 of the dystrophin gene resulted in skipping of the exon during splicing in a case of Duchenne muscular dystrophy. *Journal of Clinical Investigation* **91**, 1862–1867. 17
- A. Nasevicius & S. C. Ekker (2000). Effective targeted gene ‘knock-down’ in zebrafish. *Nature Genetics* **26**, 216–220. 49
- M. Naville, S. Henriot, I. Warren, S. Sumic, M. Reeve, J.-N. Volff & D. Chourrout (2019). Massive Changes of Genome Size Driven by Expansions of Non-autonomous Transposable Elements. *Current biology : CB* **29**, 1161–1168.e6. 18
- D. O. Norris & J. A. Carr (2020). *Vertebrate Endocrinology*. Academic Press. Google-Books-ID : L9PVDwAAQBAJ. 11
- R. G. Northcutt (1984). Evolution of the Vertebrate Central Nervous System : Patterns and Processes I. *American Zoologist* **24**, 701–716. 10
- R. G. Northcutt (1995). The forebrain of gnathostomes : in search of a morphotype. *Brain, Behavior and Evolution* **46**, 275–318. 11
- R. G. Northcutt (1996). The Agnathan ark : the origin of craniate brains. *Brain, Behavior and Evolution* **48**, 237–247. 11
- C. Nüsslein-Volhard, D. T. Gilmour & R. Dahm (2002). Introduction : zebrafish as a system to study development and organogenesis. In *Zebrafish*, pp. 1–5. 48
- H. E. O’Brien, E. Hannon, M. J. Hill, C. C. Toste, M. J. Robertson, J. E. Morgan, G. McLaughlin, C. M. Lewis et al. (2018). Expression quantitative trait loci in the developing human brain and their enrichment in neuropsychiatric disorders. *Genome Biology* **19**, 194. 79, 146, 148
- S. Ohno (1999). Gene duplication and the uniqueness of vertebrate genomes circa 1970–1999. *Seminars in Cell & Developmental Biology* **10**, 517–522. 12, 54, 78, 142
- J. O’Neil, J. Tchinda, A. Gutierrez, L. Moreau, R. S. Maser, K.-K. Wong, W. Li, K. McKenna et al. (2007). Alu elements mediate MYB gene tandem duplication in human T-ALL. *Journal of Experimental Medicine* **204**, 3059–3066. 17
- L. E. Orgel & F. H. Crick (1980). Selfish DNA : the ultimate parasite. *Nature* **284**, 604–607. 13, 17
- Y.-H. Pan, N. Wu & X.-B. Yuan (2019). Toward a Better Understanding of Neuronal Migration Deficits in Autism Spectrum Disorders. *Frontiers in Cell and Developmental Biology* **7**, 205. 146
- D. M. Parichy (2015). Advancing biology through a deeper understanding of zebrafish ecology and evolution. *eLife* **4**. 47
- J. Pasquier, C. Cabau, T. Nguyen, E. Jouanno, D. Severac, I. Braasch, L. Journot, P. Pontarotti et al. (2016). Gene evolution and gene expression after whole genome duplication in fish : the PhyloFish database. *BMC genomics* **17**, 368. 61, 69, 72
- L. M. Payer & K. H. Burns (2019). Transposable elements in human genetic disease. *Nature Reviews. Genetics* **20**, 760–772. 17
- J. W. Pendleton, B. K. Nagai, M. T. Murtha & F. H. Ruddle (1993). Expansion of the Hox gene family and the evolution of chordates. *Proceedings of the National Academy of Sciences of the United States of America* **90**, 6300–6304. 12
- J. F. Pereira, A. P. M. M. Almeida, J. Cota, J. A. Pamphile, G. Ferreira da Silva, E. F. de Araújo, K. P. Gramacho, S. H. Brommonschenkel et al. (2013). Boto, a class II transposon in *Moniliophthora perniciosa*, is the first representative of the PIF/Harbinger superfamily in a phytopathogenic fungus. *Microbiology (Reading, England)* **159**, 112–125. 16, 58, 71
- M. Plooster, G. Rossi, M. S. Farrell, J. C. McAfee, J. L. Bell, M. Ye, G. H. Diering, H. Won et al. (2021). Schizophrenia-Linked Protein tSNARE1 Regulates Endosomal Trafficking in Cortical Neurons. *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience* **41**, 9466–9481. 149

- H. Qi, Y.-B. Zhang, L. Sun, C. Chen, B. Xu, F. Xu, J.-W. Liu, J.-C. Liu et al. (2017). Discovery of susceptibility loci associated with tuberculosis in Han Chinese. *Human Molecular Genetics* **26**, 4752–4763. 79
- R. Rebollo, M. T. Romanish & D. L. Mager (2012). Transposable elements : an abundant and natural source of regulatory sequences for host genes. *Annual Review of Genetics* **46**, 21–42. 72
- S. R. Richardson, A. J. Doucet, H. C. Kopera, J. B. Moldovan, J. L. Garcia-Perez & J. V. Moran (2015). The influence of LINE-1 and SINE retrotransposons on mammalian genomes. *Microbiology Spectrum* **3**, MDNA3-0061–2014. 16
- M. Rocha, N. Singh, K. Ahsan, A. Beiriger & V. E. Prince (2020). Neural crest development : Insights from the zebrafish. *Developmental dynamics : an official publication of the American Association of Anatomists* **249**, 88–111. 146
- A. S. L. Rodrigues, J. D. Pilgrim, J. F. Lamoreux, M. Hoffmann & T. M. Brooks (2006). The value of the IUCN Red List for conservation. *Trends in Ecology & Evolution* **21**, 71–76. 3
- F. Sabot, P. Sourdille, N. Chantret & M. Bernard (2006). Morgane, a new LTR retrotransposon group, and its subfamilies in wheats. *Genetica* **128**, 439–447. 15
- P. Salis, T. Lorin, V. Lewis, C. Rey, A. Marcionetti, M.-L. Escande, N. Roux, L. Besseau et al. (2019). Developmental and comparative transcriptomic identification of iridophore contribution to white barring in clownfish. *Pigment Cell & Melanoma Research* **32**, 391–402. 81, 82
- A. Sarkar, J.-N. Volf & C. Vaury (2017). piRNAs and their diverse roles : a transposable element-driven tactic for gene regulation? *The FASEB Journal* **31**, 436–446. 72
- T. F. Schilling (2002). The morphology of larval and adult zebrafish. In *Zebrafish*, pp. 59–94. 52
- Schizophrenia Working Group of the Psychiatric Genomics Consortium (2014). Biological insights from 108 schizophrenia-associated genetic loci. *Nature* **511**, 421–427. 79, 146, 148
- E. C. Scott, E. J. Gardner, A. Masood, N. T. Chuang, P. M. Vertino & S. E. Devine (2016). A hot L1 retrotransposon evades somatic repression and initiates human colorectal cancer. *Genome Research* **26**, 745–755. 17
- R. Sears & W. F. Perrin (2009). Blue Whale. In *Encyclopedia of Marine Mammals*, pp. 120–124. Elsevier. 3
- S. M. Shimeld & P. W. Holland (2000). Vertebrate innovations. *Proceedings of the National Academy of Sciences of the United States of America* **97**, 4449–4452. 7, 10
- C. Singleman & N. G. Holtzman (2014). Growth and Maturation in the Zebrafish, Danio Rerio : A Staging Tool for Teaching and Research. *Zebrafish* **11**, 396–406. 48
- L. Sinzelle, V. V. Kapitonov, D. P. Grzela, T. Jursch, J. Jurka, Z. Izsvák & Z. Ivics (2008). Transposition of a reconstructed Harbinger element in human cells and functional homology with two transposon-derived cellular genes. *Proceedings of the National Academy of Sciences of the United States of America* **105**, 4715–4720. 17, 59, 73, 79, 144
- J. J. Smith, K. Sumiyama & C. T. Amemiya (2012). A living fossil in the genome of a living fossil : Harbinger transposons in the coelacanth genome. *Molecular Biology and Evolution* **29**, 985–993. 59, 69, 149
- G. Streisinger, C. Walker, N. Dower, D. Knauber & F. Singer (1981). Production of clones of homozygous diploid zebra fish (*Brachydanio rerio*). *Nature* **291**, 293–296. 48
- F. Sugahara, Y. Murakami, J. Pascual-Anaya & S. Kuratani (2017). Reconstructing the ancestral vertebrate brain. *Development, Growth & Differentiation* **59**, 163–174. 10
- T. E. Sztal, E. A. McKaige, C. Williams, A. A. Ruparella & R. J. Bryson-Richardson (2018). Genetic compensation triggered by actin mutation prevents the muscle damage caused by loss of actin protein. *PLoS genetics* **14**, e1007212. 148
- D. Tautz (1992). Redundancies, development and the flow of information. *BioEssays : News and Reviews in Molecular, Cellular and Developmental Biology* **14**, 263–266. 147
- L. Teboul, S. A. Murray & P. M. Nolan (2017). Phenotyping first-generation genome editing mutants : a new standard? *Mammalian Genome* **28**, 377–382. 147, 148
- J. Thomas & E. J. Pritham (2015). Helitrons, the eukaryotic rolling-circle transposable elements. *Microbiology Spectrum* **3**. 16
- P. van der Harst & N. Verweij (2018). Identification of 64 Novel Genetic Loci Provides an Expanded View on the Genetic Architecture of Coronary Artery Disease. *Circulation Research* **122**, 433–443. 79
- R. Van Der Laan, W. N. Eschmeyer & R. Fricke (2014). Family-group names of Recent fishes. *Zootaxa* **3882**, 1–230. 58
- C. N. Velanis, P. Perera, B. Thomson, E. de Leau, S. C. Liang, B. Hartwig, A. Förderer, H. Thornton et al. (2020). The domesticated transposase ALP2 mediates formation of a novel Polycomb protein complex by direct interaction with MSI1, a core subunit of Polycomb Repressive Complex 2 (PRC2). *PLoS genetics* **16**, e1008681. 73, 150
- J.-N. Volf (2005). Genome evolution and biodiversity in teleost fish. *Heredity* **94**, 280–294. 58
- D. F. Voytas & J. D. Boeke (1992). Yeast retrotransposon revealed. *Nature* **358**, 717. 14
- M. R. Wallace, L. B. Andersen, A. M. Saulino, P. E. Gregory, T. W. Glover & F. S. Collins (1991). A de novo Alu insertion results in neurofibromatosis type 1. *Nature* **353**, 864–866. 17
- K. Wang, J. Wang, C. Zhu, L. Yang, Y. Ren, J. Ruan, G. Fan, J. Hu et al. (2021). African lungfish genome sheds light on the vertebrate water-to-land transition. *Cell* **184**, 1362–1376.e18. 3
- T. Wicker, F. Sabot, A. Hua-Van, J. L. Bennetzen, P. Capy, B. Chalhoub, A. Flavell, P. Leroy et al. (2007). A unified classification system for eukaryotic transposable elements. *Nature Reviews Genetics* **8**, 973–982. 14, 16, 58

- J. B. Williams, A. Muñoz-Garcia, S. Ostrowski & B. I. Tieleman (2004). A phylogenetic analysis of basal metabolism, total evaporative water loss, and life-history among foxes from desert and mesic regions. *Journal of Comparative Physiology. B, Biochemical, Systemic, and Environmental Physiology* **174**, 29–39. 3
- C. P. Witte, Q. H. Le, T. Bureau & A. Kumar (2001). Terminal-repeat retrotransposons in miniature (TRIM) are involved in restructuring plant genomes. *Proceedings of the National Academy of Sciences of the United States of America* **98**, 13778–13783. 15
- R. S. Wu, I. I. Lam, H. Clay, D. N. Duong, R. C. Deo & S. R. Coughlin (2018). A Rapid Method for Directed Gene Knockout for Screening in G0 Zebrafish. *Developmental Cell* **46**, 112–125.e4. 79, 80, 81, 82, 83, 85, 87, 139, 147
- M. Yang, J. Zhong, M. Zhao, J. Wang, Y. Gu, X. Yuan, J. Sang & C. Huang (2014). Overexpression of nuclear apoptosis-inducing factor 1 altered the proteomic profile of human gastric cancer cell MKN45 and induced cell cycle arrest at G1/S phase. *PLoS One* **9**, e100216. 80
- W. R. Yang, D. Ardeljan, C. N. Pacyna, L. M. Payer & K. H. Burns (2019). SQuIRE reveals locus-specific regulation of interspersed repeat expression. *Nucleic Acids Research* **47**, e27. 61
- J. R. York & D. W. McCauley (2020). The origin and evolution of vertebrate neural crest cells. *Open Biology* **10**, 190285. 7
- J.-K. Yu, D. Meulemans, S. J. McKeown & M. Bronner-Fraser (2008). Insights from the amphioxus genome on the origin of vertebrate neural crest. *Genome Research* **18**, 1127–1132. 7
- Y.-W. Yuan & S. R. Wessler (2011). The catalytic domain of all eukaryotic cut-and-paste transposase superfamilies. *Proceedings of the National Academy of Sciences of the United States of America* **108**, 7884–7889. 16, 58, 59, 71
- J. Zempleni, Y. C. Chew, B. Bao, V. Pestinger & S. S. K. Wijeratne (2009). Repression of transposable elements by histone biotinylation. *The Journal of Nutrition* **139**, 2389–2392. 72
- L. Zhang, L. Yan, J. Jiang, Y. Wang, Y. Jiang, T. Yan & Y. Cao (2014). The structure and retrotransposition mechanism of LTR-retrotransposons in the asexual yeast *Candida albicans*. *Virulence* **5**, 655–664. 14
- X. Zhang, C. Feschotte, Q. Zhang, N. Jiang, W. B. Eggleston & S. R. Wessler (2001). P instability factor : An active maize transposon system associated with the amplification of Tourist-like MITEs and a new superfamily of transposases. *Proceedings of the National Academy of Sciences* **98**, 12572–12577. 16, 58, 71
- X. Zhang, N. Jiang, C. Feschotte & S. R. Wessler (2004). PIF- and Pong-like transposable elements : distribution, evolution and relationship with Tourist-like miniature inverted-repeat transposable elements. *Genetics* **166**, 971–986. 16, 58, 71
- Y. Zhang, X. You, S. Li, Q. Long, Y. Zhu, Z. Teng & Y. Zeng (2020). Peripheral Blood Leukocyte RNA-Seq Identifies a Set of Genes Related to Abnormal Psychomotor Behavior Characteristics in Patients with Schizophrenia. *Medical Science Monitor : International Medical Journal of Experimental and Clinical Research* **26**, e922426. 79, 146, 148
- G. Zhao, L. Liu, T. Zhao, S. Jin, S. Jiang, S. Cao, J. Han, Y. Xin et al. (2015). Upregulation of miR-24 promotes cell proliferation by targeting NAIF1 in non-small cell lung cancer. *Tumour Biology : The Journal of the International Society for Oncodevelopmental Biology and Medicine* **36**, 3693–3701. 80
- X. Zhou, J. He, C. N. Velanis, Y. Zhu, Y. He, K. Tang, M. Zhu, L. Graser et al. (2021). A domesticated Harbinger transposase forms a complex with HDA6 and promotes histone H3 deacetylation at genes but not TEs in Arabidopsis. *Journal of Integrative Plant Biology* **63**, 1462–1474. 73, 150
- C. Zimmer (2000). Evolution. In search of vertebrate origins : beyond brain and bone. *Science (New York, N.Y.)* **287**, 1576–1579. 3, 7, 12

