Impact des éléments transposables sur l'évolution de la régulation des gènes : exemple du sexe chez les poissons téléostéens

Corentin Dechaud

Numéro National de Thèse : 2021LYSEN012

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 18 mai 2021, par :

Corentin DECHAUD

Impact des éléments transposables sur l'évolution de la régulation des gènes : Exemple du sexe chez les poissons téléostéens

---

Devant le jury composé de :

COELHO,	Susana	Directrice de recherche	Max Planck Institute	Rapportrice
QUESNEVILLE,	Hadi	Directeur de recherche	INRAE Versailles	Rapporteur
FISTON-LAVIER,	Anna-Sophie	Maître de conférences	Univ. de Montpellier	Examinatrice
GILBERT,	Clément	Chargé de recherche	CNRS Paris-Saclay	Examinateur
SCHARTL,	Manfred	Professeur	Univ. Würzburg	Examinateur
VIEIRA,	Cristina	Professeur	Univ. Lyon 1	Examinatrice
VOLFF,	Jean-Nicolas	Professeur	ENS de Lyon	Directeur de thèse
MAUSSION-NAVILLE,	Magali	Agrégée préparatrice	ENS de Lyon	Co-encadrante

Résumé de la thèse

Chez les poissons téléostéens, les modes de reproduction sexuée et les réseaux de régulation des gènes liés au sexe sont très variables. La détermination du sexe des espèces gonochoriques, par exemple, peut être aussi bien génétique qu'environnementale et peut impliquer des gènes différents selon les espèces. Les régulations du développement et du maintien du sexe apparaissent également variables dans ce groupe.

Pour essayer de comprendre l'origine de cette diversité, je me suis intéressé à l'impact possible des éléments transposables sur la régulation de gènes liés au sexe chez ces poissons. Les éléments transposables sont des séquences d'ADN endogènes capables de se déplacer ou de se copier dans les génomes. Bien que souvent neutres, et parfois délétères pour leur hôte, les éléments transposables peuvent aussi transporter des séquences régulatrices, comme des sites de fixation de facteurs de transcription, et les disséminer à travers les génomes. Leur forte diversité dans les génomes de poissons constitue un réservoir de séquences régulatrices disponibles.

Pour tester cette hypothèse, j'ai utilisé des données de séquençage d'ARN issues de gonades mâles et femelles d'*Oryzias latipes*, le médaka japonais. Dans un premier temps, j'ai analysé l'expression des gènes et des éléments transposables et mis en évidence des régions du génome enrichies en gènes et éléments transposables différentiellement exprimés entre les gonades mâles et femelles. De plus, les gènes et les éléments transposables proches le long des chromosomes ont tendance à présenter des biais d'expression similaires. Deux hypothèses, non mutuellement exclusives, peuvent rendre compte de cette observation : d'une part, les éléments transposables pourraient modifier l'expression des gènes voisins, et d'autre part, l'environnement génomique du site d'insertion pourrait influencer l'expression des éléments transposables. Les travaux réalisés ne permettent pas de trancher définitivement entre ces deux hypothèses, mais plusieurs observations sont en faveur d'un rôle régulateur de certains éléments transposables. Dans un deuxième axe et de manière complémentaire, j'ai mis en évidence des familles d'éléments transposables physiquement enrichies dans l'environnement des gènes sexe-biaisés. Une famille candidate a été étudiée plus en détail, et j'ai pu mettre en évidence dans ces éléments des sites de fixation de facteurs de transcription connus pour être impliqués dans la fonction sexuelle.

Ces travaux montrent ainsi le rôle potentiel des éléments transposables dans l'évolution rapide de certains réseaux de régulation de gènes et serviront de socle pour de futures études fonctionnelles.

PhD thesis summary

In teleost fish, sexual reproduction and sexual gene regulatory networks are highly variable. In gonochoristic species, sex is determined either environmentally or genetically and can involve different genes depending on the species investigated. Sexual development and maintenance appear also variable in this clade.

To understand the origin of this diversity, I studied the posssible impact of transposable elements on the fast evolution of gene regulatory networks related to sex in fish. Transposable elements

are endogenous DNA sequences able to move or copy themselves in genomes. Even if they are often deleterious for their host, transposable elements can also carry regulatory sequences, such as transcription factor binding sites, and spread them in genomes. Their diversity in fish genomes form a source of ready-to-use regulatory sequences potentially involved in the fast evolution of some gene regulatory networks.

To test this hypothesis, I used RNA sequencing data from male and female gonads of the Japanese Medaka, *Oryzias latipes*. First I analysed gene and transposable element expression and discovered regions of the genome enriched in sex-biased genes associated to sex-biased transposable elements. Moreover, genes and transposable elements located close on chromosomes tend to present similar expression bias between testis and ovary. Two hypothesis that are not mutually exclusive can explain this observation : either transposable element influence gene expression of neighbooring genes, or the genomic locus where the transposable element inserts influence its expression. We were not able to definitively discriminate between these two hypotheses, but our work identified several clues for a regulatory role of transposable elements. In the second part and in a complementary way I found transposable element families physically enriched near to sex-biased genes. One family was further investigated and shown to carry transcription factor binding sites involved in sexual function.

This work brings new insights on the possible role of transposable elements in the fast evolution of gene regulatory networks and paves the way for future functionnal studies.

```
corentin@igfl:~/phd/manuscript/conclusion$ cd
corentin@igfl:~$ ls
igfl_memes     teaching
phd            velotaf
corentin@igfl:~$ cd igfl_memes
corentin@igfl:~/igfl_memes$ git push
corentin@igfl:~/igfl_memes$ cd
corentin@igfl:~$ cd phd/manuscript
corentin@igfl:~/phd/manuscript$
[ $[$RANDOM % 6] == 0 ] && rm -rf /home/corentin/phd/ || pdflatex phd_manuscript.tex &&
echo "Compiling ..."
Compiling ...
```

REMERCIEMENTS

Pour commencer, je ne peux que me tourner vers Magali. Merci pour ton soutien quotidien et ton encadrement qui m'aura guidé pendant ces trois années. Merci pour les précieux conseils, les discussions, et tout ce qui en a découlé : tu m'as permis de trouver mon chemin au travers de toute ces séquences répétées. Mais le point clé restera ta capacité à me lire et me re-lire !

Je vais poursuivre avec mon coach, Jean-Nicolas, entraîneur et capitaine, qui lors du mercato 2017 a choisi de m'accueillir dans son équipe. Merci Jean-Ni pour la liberté scientifique que tu m'as accordé depuis le début, pour la mise en avant dont tu m'as fait profiter, pour la confiance dans mon travail, tout en faisant preuve d'une grande disponibilité dans les moments où j'en avais besoin.

Le travail de recherche étant avant tout un travail d'équipe, merci les Volff : Ema, Fred, Jerem, Laure et Z pour les interactions scientifiques ainsi que la bonne ambiance quotidienne que vous apportez à la « team ». Je vais y ajouter un ancien : Thibault – Mr. WorldWide, merci pour ta formation à la PhD life ainsi qu'à la grimpette. J'espère que tu auras pu réviser ta géographie depuis le temps. Aux membres qui ont été de passage, Sho, Sara et Emilie, merci pour votre travail remarquable – *Yoku yatta*. Et enfin ceux avec qui je n'ai pas directement travaillé : Candice et Théo – Mr Ocean.

Merci plus largement à tout l'IGFL, pour le cadre, l'ambiance de travail et les retours constructifs que vous m'avez offerts. Merci Martine et toute l'équipe administrative, pour votre gentillesse et vos précieuses compétences dans le labyrinthe administratif.

Mon travail de thèse a été réalisé en collaboration avec l'équipe de Manfred Schartl à Würzburg en Allemagne. Je remercie Manfred Schartl, Frederik Helmprobst, Susanne Kneitz et Anabel Martinez pour leur implication dans la partie expérimentale du projet, pour les discussions scientifiques mais aussi pour leur accueil lors de ma visite à Würzburg en avril 2018.

Merci Augustin, Vincent, Théodore et les autres copains-collègues doctorants (de l'IGFL mais pas que !) pour le soutien mutuel, le meme-making, et les pauses café sans café (mais avec pains au lait !).

Pour ponctuer mes travaux, j'ai eu la chance de pouvoir participer à des activités d'enseignement. Merci Matthieu pour ces deux années passées à l'IUT, mais aussi Marie de m'avoir permis de poursuivre à l'ENS. Carine, the IFB master, la faiseuse de docker, amie des git et purple flasher, sincèrement merci pour tout ce que tu m'as apporté.

Merci Clément et encore Marie, pour vos encouragements, ainsi que pour avoir pointé les écueils à éviter au cours des trois comités de suivi. Votre bienveillance a été une force pour moi.

En s'éloignant du labo, je tiens à remercier quelque cocs qui m'ont aussi aidé à arriver jusqu'ici : les géniaux Axel et Jerem, mais aussi bien-sûr Ema, Maxime, Leslie, Quentin et Carole, pour les tours en canoche ou les balades en Ardoche, c'était tip top. Merci à vous aussi les « adorables » (non je ne donnerai pas votre deuxième adjectif, vous ne le méritez pas), Mel (et Damien !), Babet, Pierrick, Valou, Robine, Margot, Benj, Marie, Margaux, Mélanie … Sans oublier les no-diam pour les soirées valorantes civilisées (non, les full à 50m, ça passe pas).

Je remercie tout particulièrement ma famille pour les encouragements et le soutien indéfectible dans les choix que j'ai pu faire depuis le début de mes études. Merci de m'avoir donné la chance d'arriver jusqu'ici. Gisèle, tu as aussi grandement participé à m'amener jusqu'ici, à ta mémoire, merci.

Pour finir, si tout s'est aussi bien passé ces dernières années c'est très largement grâce à toi ($p < 10^{-6}$) ***,
merci Pauline.

Table des matières

Liste des figures

Liste des tableaux

1 Introduction

Sommaire

1.1 Origine, évolution et impact du sexe

1.1.1 Reproduction sexuelle / asexuelle

La reproduction sexuelle est apparue il y a environ un milliard d'années et a été maintenue au cours de l'histoire évolutive des eucaryotes (OTTO et LENORMAND, 2002). Elle fait référence à l'union des génomes haploïdes issus de deux individus différents. La reproduction asexuée, présente chez tous les procaryotes, n'existe à l'inverse que chez de rares espèces d'eucaryotes. Certains procaryotes peuvent également présenter des formes de reproduction sexuée (transformation / transduction / conjugaison) aboutissant au mélange de deux génomes de manière asymétrique, où un fragment du génome d'un donneur est transféré à un receveur (OTTO et LENORMAND, 2002). Dans ce manuscrit, la reproduction sexuée fera référence à la fusion symétrique de deux génomes complets, telle qu'observée la plupart du temps chez les eucaryotes.

1.1.1.1 Le sexe, définitions

Chez les espèces gonochoriques, dont les individus ne changent pas de sexe, on distingue deux types d'individus, les mâles et les femelles ; on parle alors de deux sexes différents. Cependant, la distinction entre mâle et femelle n'est pas aussi triviale qu'elle peut le paraître au premier abord, et peut se faire à plusieurs niveaux :

Le sexe chromosomique

Chez l'humain, les mâles ont généralement un chromosome X et un chromosome Y alors que les femelles possèdent deux chromosomes X. Dans de rares cas toutefois, les individus peuvent ne posséder qu'un chromosome X et pas de Y (Syndrome de Turner), trois chromosomes X (Syndrome triple-X), ou encore deux X et un Y (Syndrome de Klinefelter). On pourrait alors définir les mâles comme possédant un chromosome Y et les femelles comme n'en possédant pas. Mais ça n'est en fait pas toujours le cas, comme nous allons le voir tout de suite.

Le sexe génétique

Chez l'humain c'est le gène *Sry* présent sur le chromosome Y qui détermine le sexe mâle d'un individu. Cependant, si *Sry* se retrouve sur le chromosome X par un mécanisme rare de réarrangement, l'individu sera bel et bien un mâle, malgré le fait qu'il possède deux chromosomes X. Peut-on alors dire que le sexe génétique déterminé par *Sry* est la caractéristique qui permet de distinguer précisément mâles et femelles ?

Le sexe gonadique

On définit par sexe gonadique le fait que les individus qui possèdent des testicules sont considérés comme mâles, et les individus qui possèdent des ovaires sont considérés comme femelles, indépendamment de leurs gènes ou de leurs chromosomes. Bien que la présence d'une gonade mâle soit souvent déterminée par *Sry* chez les mammifères, il est connu que des mutations dans d'autres gènes peuvent perturber la formation d'un testicule (CAMATS *et al.*, 2012; GONEN *et al.*, 2018). Le type de gonade présent permettrait-il finalement de s'affranchir des gènes et serait-il une bonne manière de déterminer le sexe d'un individu ?

Le sexe phénotypique

Le sexe phénotypique fait référence aux attributs sexuels morphologiques visibles. Mais encore une fois, le sexe phénotypique ne correspond pas forcément au sexe gonadique. L'OMS rapporte qu'entre 1 :600 et 1 :5000 naissances présentent des caractéristiques ne correspondant pas parfaitement à une définition typique de mâle ou femelle (OMS), et certaines études jusqu'à 1 :100 selon le critère d'inclusivité utilisé (ARBOLEDA *et al.*, 2014).

Le sexe mosaïque

A l'échelle des cellules individuelles, la définition du sexe est encore plus floue (AINSWORTH, 2015). L'idée que les cellules d'un individu sont génétiquement identiques est fausse. D'une part à cause de rares mutations ponctuelles, mais aussi en raison d'un mosaïcisme de composition en chromosomes chez certains individus. Ce mosaïcisme se met en place lors des premières divisions cellulaires pendant le développement embryonnaire, si le chromosome Y est perdu ; une partie des cellules peuvent être XY, l'autre simplement X. Cela peut engendrer un syndrome de Turner si la majorité des cellules sont X, ou un individu avec les traits d'un mâle si la proportion de cellules ayant perdu le chromosome Y est plus faible (AINSWORTH, 2015).

Ranger les individus dans les catégories mâle et femelle présente une raison pratique en biologie et permet de bien résumer la variabilité biologique. C'est la raison pour laquelle ces termes seront utilisés dans ce manuscrit. Cependant, et comme montré précédemment, il faut garder à l'esprit que cette classification n'est pas parfaite et qu'il existe des exceptions. Affirmer qu'il existe une frontière nette entre mâles et femelles n'est pas soutenu par les connaissances scientifiques actuelles. Cette catégorisation peut d'ailleurs devenir problématique lorsqu'elle est utilisée pour justifier des choix médicaux, notamment lors de chirurgies visant à « rectifier » les caractères sexuels de nouveau-nés qui ne correspondent pas parfaitement à la définition typique d'un des deux sexes. [1]

Dans la suite de ce manuscrit, quand le terme sexe sera utilisé pour différencier mâles et femelles, il fera référence à la normalité (dans le sens statistique du terme : la catégorie présentant le plus grand effectif).

Chez les espèces à reproduction sexuée, chaque sexe produit un type de gamètes spécifique. Chez les animaux, on parle d'anisogamie : les gamètes mâles (spermatozoïdes), et femelles (ovocytes) sont différents, en taille et morphologie (VALENZUELA, 2008). Par opposition, l'isogamie est plutôt observée chez les plantes ou les champignons. La production des gamètes, appelée méiose, sépare un génome en deux. Elle est suivie de la fécondation où un gamète mâle rencontre un gamète femelle pour former le futur individu en rassemblant les deux génomes dans une même cellule. C'est cette alternance de cycles de méiose / fécondation qui est appelée reproduction sexuée (ARKHIPOVA, 2005). Le terme « sexe » quant à lui peut permettre de distinguer mâle et femelle, comme vu précédemment, ou peut être utilisé à la place de l'expression « reproduction sexuée » comme dans le titre de ce chapitre. Dans ce manuscrit le terme sexe pourra être utilisé dans les deux cas décrits, en gardant en tête les limites qui lui sont associées, et ne doit pas être confondu avec le genre, ni l'orientation sexuelle d'un individu, qui sont deux notions indépendantes du sexe tel qu'il vient d'être défini.

1. Pendant que j'écris ces lignes, le 31 Juillet 2020, l'assemblée nationale vient de rejeter un projet de loi ayant pour but de mettre fin aux mutilations sur les enfants intersexes. http://www2.assemblee-nationale.fr/scrutins/detail/(legislature)/15/(num)/2853

1.1.1.2 L'effet loupe des eucaryotes asexués

Les espèces asexuées utilisent une reproduction clonale, c'est-à-dire que la progéniture est génétiquement identique à la mère (ENGELSTÄDTER, 2017). Les quelques espèces eucaryotes asexuées sont isolées dans l'arbre du vivant et bénéficient d'un « effet loupe » en attirant toute l'attention, comme décrit par Otto et Lenormand : « exceptions become celebrity » (OTTO et LENORMAND, 2002). Chez les vertébrés, ces « célébrités » ont été décrites par Vrijenhoek et co. en 1989, et rassemblent 22 espèces de poissons, 23 d'amphibiens et 29 de reptiles (VRIJENHOEK *et al.*, 1989). Parmi elles on retrouve la Molly amazone, une espèce de poisson connue pour ne produire que des femelles. Malgré tout, ces espèces restent une minorité par rapport aux ~72 000 espèces de vertébrés connues actuelles.

1.1.1.3 Les cas particuliers

La séparation entre reproduction sexuée et asexuée peut être parfois floue sachant que certaines espèces sont capables d'alterner entre des cycles sexués et asexués. Il en est de même pour les espèces automixiques, chez qui la méiose chez la femelle est suivie de la fusion entre deux gamètes femelles, donnant lieu à une reproduction asexuée mais non clonale (ENGELSTÄDTER, 2017).

1.1.2 Histoire évolutive et paradoxe du sexe

1.1.2.1 Le paradoxe

La reproduction sexuelle est présente chez la très grande majorité des eucaryotes, que ce soit chez les plantes, les champignons ou les animaux. Pourtant, de nombreux coûts sont associés au sexe (OTTO et LENORMAND, 2002; ROZE, 2012). La recherche d'un partenaire et l'accouplement sont autant de temps qui n'est pas investi à la recherche de nourriture ou à l'investissement d'énergie pour survivre. De plus, les individus s'exposent mutuellement au risque des infections sexuellement transmissibles. Deux individus étant nécessaires, la reproduction sexuée est moins efficace que la reproduction asexuée qui permet à un individu de se reproduire seul. Enfin, mélanger des gènes qui ont permis à un individu de survivre jusqu'à être en âge de se reproduire avec ceux d'un autre individu est risqué, alors qu'une reproduction clonale asexuée lui assurerait de transmettre 100% de ses gènes qui ont fait la preuve de leur fonctionnalité. La très large distribution du sexe dans l'arbre du vivant associée aux nombreux coûts de ce dernier, constitue ainsi ce qui est appelé le paradoxe du sexe.

1.1.2.2 Vers une résolution du paradoxe ?

Les premières hypothèses pour expliquer ce paradoxe ont été formulées au début du XXème siècle par Morgan, Fisher et Muller, qui proposent de lier le bénéfice du sexe à la variabilité génétique qu'il génère en regroupant des mutations avantageuses dans le même génome (FISHER, 1930; MORGAN, 1913; MULLER, 1932). Cependant, ce modèle a été remis en question plus tard, dans les années 1970 : pour être avantageux, le sexe doit associer des mutations avantageuses plus souvent qu'il ne les sépare (WILLIAMS et MITTON, 1973). Aujourd'hui, le paradoxe du sexe reste

l'une des énigmes majeures en biologie évolutive. Les hypothèses retenues se basent sur une augmentation de la variation génétique liée au sexe qui permettrait une augmentation de l'efficacité de la sélection naturelle. De plus, des études d'évolution expérimentale ont montré que le sexe permet une adaptation plus rapide aux changements environnementaux. Ces avantages sur le long terme permettraient de compenser le coût à court terme de la reproduction sexuée (ROZE, 2012).

1.1.3 Le développement sexuel chez les poissons téléostéens

Le développement sexuel, chez les espèces sexuées, peut être divisé en trois étapes (HAYES, 1998). La première est la détermination, durant laquelle le sexe, mâle ou femelle, est choisi. Pendant la seconde étape, correspondant à la différenciation sexuelle, la gonade indifférenciée devient une gonade mâle ou femelle, respectivement un testicule ou un ovaire. La dernière étape du développement sexuel correspond au maintien des caractères sexuels chez l'adulte. Les poissons téléostéens constituent un excellent modèle pour étudier le développement sexuel ; ils présentent en effet des exemples de tous les systèmes sexuels trouvés chez les vertébrés.

1.1.3.1 Les différents modes de reproduction chez les poissons téléostéens

Les espèces hermaphrodites

Chez les espèces hermaphrodites, les deux sexes ne sont pas séparés entre différents individus. Il existe plusieurs types d'hermaphrodites : simultanés ou séquentiels. Les hermaphrodites simultanés possèdent des caractères sexuels mâles et femelles en même temps. Très répandus chez les plantes, ils sont rares chez les vertébrés ; on trouve des hermaphrodites simultanés dans le genre téléostéen Hypoplectrus (**Fig. 1.1.c.**) (HENCH *et al.*, 2017). En revanche, l'hermaphrodisme séquentiel est très répandu chez les poissons téléostéens (MITCHESON et LIU, 2008; ROSS, 1990). Ces poissons ont la capacité de changer de sexe une ou plusieurs fois au cours de leur vie. C'est le cas des gobies verts qui changent de sexe à leur guise : lorsqu'un couple homosexuel se rencontre, l'un des deux individus est capable de changer de sexe pour leur permettre leur reproduction (KROON *et al.*, 2003; MUNDAY, 2002). D'autres espèces ne changent de sexe qu'une seule fois au cours de leur vie. C'est par exemple le cas de la girelle à tête bleue (**Fig. 1.1.a.**) dont tous les individus naissent femelle. Au bout de 8 jours de vie une certaine proportion se change en mâles selon la taille du récif où vit leur groupe (WARNER et SWEARER, 1991). On parle ici de protogynie car les individus commencent leur vie en femelles. A l'opposé, on parle de protandrie pour les espèces qui commencent leur vie en mâles et la terminent en femelles. C'est le cas d'une espèce bien connue depuis son apparition au cinéma : le poisson-clown (**Fig. 1.1.b.**) (CASAS *et al.*, 2016; HECQUET, 2018; ROSS, 1990; TODD *et al.*, 2016). Les poisson-clowns vivent en groupe autour d'une anémone qui les protège des prédateurs. Dans un groupe, tous les individus sont des mâles, sauf le plus gros individu qui a pour rôle de protéger le groupe et sa descendance : c'est l'unique femelle. Le second individu en taille est le mâle reproducteur, qui est suivi par les juvéniles classés socialement du plus gros au plus petit. Lorsqu'un nouveau mâle rejoint le groupe il se place en bas de la « file d'attente ». On parle ici de file d'attente car quand la femelle meurt (par exemple croquée par un barracuda), le mâle reproducteur, le plus gros, va changer de sexe et devenir la femelle du groupe. Suite à ce changement de sexe, tous les mâles du groupe vont progresser d'un rang dans la hiérarchie, le plus

gros juvénile va prendre la place de mâle reproducteur, et ainsi de suite. Dans *Le monde de Némo*, le film d'animation sorti en 2003, la maman meurt (croquée par un barracuda) ainsi que le reste de la progéniture, et Némo, le seul survivant, est élevé par son père avant de vivre une série de péripéties. Si on devait imaginer un scénario réaliste suivant la biologie des poisson-clowns, la mort de la mère de Némo aurait eu d'autres conséquences. Marin, le père de Némo aurait changé de sexe, Némo serait devenu le mâle reproducteur, et se serait reproduit avec son père… enfin, sa mère!

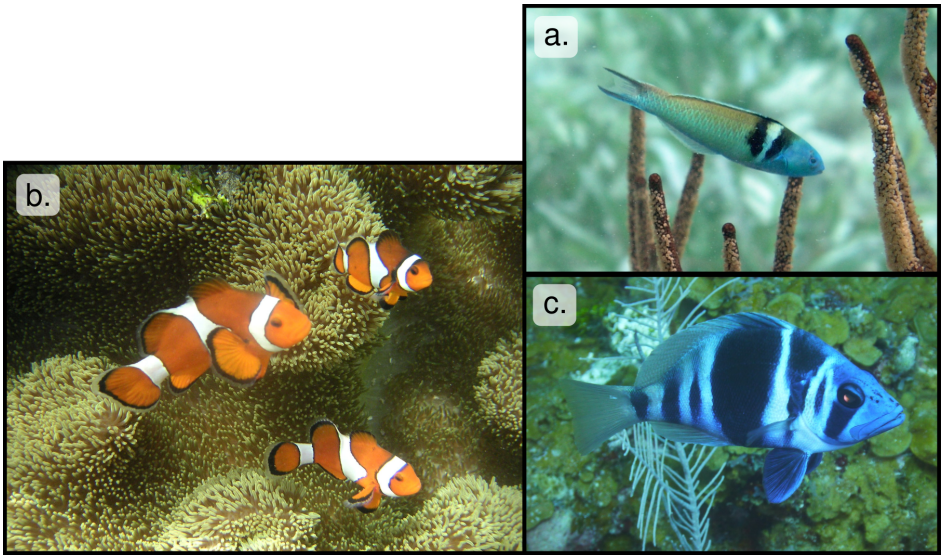

FIGURE 1.1 – **Photographies d'espèces de téléostéens hermaphrodites.** a. Girelle à tête bleue, *Thalassoma bifasciatum*, hermaphrodite séquentiel, protogyne. ©Tibor Marcinek - Public Domain. b. Poisson clown, *Amphiprion ocellaris*, hermaphrodite séquentiel, protandre. ©Metatron - CC BY-SA 3.0. c. Hamlet indigo, *Hypoplectrus indigo*, hermaphrodite simultané. ©Tomh009 - CC BY-SA 3.0.

Exemple d'une espèce asexuée : le cas des Mollies Amazones

Poecilia formosa, ou la Molly Amazone, est un petit poisson d'eau douce originaire de la vallée du Rio Grande au sud des Etats-Unis. Elle tire son nom « Amazone » de la mythologie grecque, où les Amazones sont un peuple constitué uniquement de femmes guerrières dont les hommes sont tués à la naissance. Les Mollies Amazones ont presque tout copié à ce mythe pour en faire une réalité : cette espèce ne produit pas de mâles ; leur reproduction clonale n'engendre que des femelles (SCHLUPP *et al.*, 2007). Toutefois, leur mode de reproduction asexuée, la parthénogenèse sperme-dépendante, ne leur permet pas de se reproduire totalement seules ; elles ont besoin de sperme pour y parvenir. Dans ce mode de reproduction, le sperme est utilisé non pas pour la fécondation comme matériel génétique, mais pour déclencher l'embryogenèse. Dans le but de recevoir du sperme, elles rusent avec des mâles d'espèces proches, pour les utiliser finalement comme donneurs de sperme. De manière occasionnelle, les mâles peuvent contribuer au génome de la descendance en transmettant des microchromosomes.

Les espèces gonochoriques

Comme c'est le cas chez les mammifères, certaines espèces de poissons sont gonochoriques. Cela signifie que les deux sexes apparaissent dans des individus distincts, par opposition aux espèces hermaphrodites. Dans la suite du manuscrit, nous nous focaliserons sur les espèces gonochoriques et leur développement sexuel.

1.1.3.2 Le déterminisme du sexe chez les poissons téléostéens

Le déterminisme du sexe correspond à l'étape au cours de laquelle le sexe d'un individu est défini. L'exemple le plus connu, présent chez la quasi totalité des mammifères, est le système de chromosomes sexuels XX/XY (**Fig. 1.2.**). Chez les poissons téléostéens, certaines espèces utilisent ce système, mais il n'est pas conservé comme chez les mammifères (BACHTROG *et al.*, 2014; CAPEL, 2017). La plupart des espèces étudiées n'ont pas de chromosomes sexuels identifiés, soit parce qu'ils ne sont pas ou peu différenciés et sont donc difficilement identifiables, soit parce qu'ils n'ont pas encore été recherchés (BAROILLER *et al.*, 2009; KIKUCHI et HAMAGUCHI, 2013). Il existe de nombreuses manières de déterminer le sexe d'un individu, et tout comme c'est le cas pour les différents modes de reproduction, les poissons téléostéens présentent toute la diversité de ces systèmes. Généralement on différencie les espèces à détermination génétique du sexe (GSD), comme les mammifères, et les espèces à détermination environnementale du sexe (ESD), comme certaines tortues. En réalité, cette distinction entre les deux types de détermination n'est pas toujours nette : les espèces se positionnent le long d'un gradient allant de ESD complète à GSD complète (**Fig. 1.6.**).

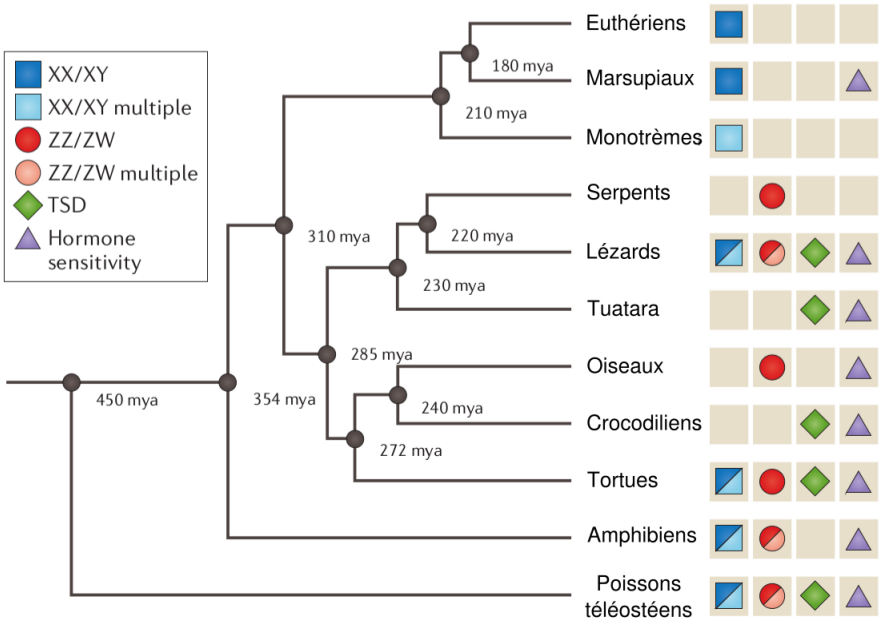

FIGURE 1.2 – **Phylogénie simplifiée des vertébrés associant le type de détermination sexuelle retrouvé dans les groupes d'espèces.** Chez les mammifères euthériens seule la détermination avec XX/XY est trouvée alors que chez les poissons téléostéens on retrouve toute la diversité des mécanismes de détermination. La sensibilité aux hormones (ARNOLD et ITOH, 2011) peut être considérée comme une détermination environnementale. TSD : Détermination du sexe température-dépendante. Adapté de Capel *et al.* 2017 (CAPEL, 2017).

La détermination environnement-dépendante du sexe (ou ESD)

L'ESD regroupe un ensemble de facteurs pouvant influencer le sexe de différentes espèces. La plus connue, très répandue chez les tortues et les crocodiles, est la détermination température-dépendante (TSD) (BULL et VOGT, 1979; LANG et ANDREWS, 1994). Chez la plupart des espèces de poissons sensibles à la température, le sexe-ratio (le rapport entre le nombre de mâles et le nombre

de femelles dans la descendance) penche en faveur des mâles quand la température augmente, même s'il existe certaines exceptions (BAROILLER *et al.*, 2009). D'autres facteurs environnementaux peuvent aussi influencer la détermination du sexe comme le pH de l'eau, son taux d'O2, des facteurs comme les interactions sociales des juvéniles asexués, ou le sexe-ratio.

La détermination du sexe génétique

Tout comme pour l'ESD, les poissons téléostéens présentent une très forte variabilité de systèmes de GSD. Contrairement aux mammifères où le gène *Sry* porté par le chromosome Y est responsable de la détermination en mâle dans toutes les espèces, un tel système conservé n'est pas retrouvé chez les poissons. Des systèmes de chromosomes sexuels en ZZ/ZW où le chromosome W est responsable de la détermination en femelle existent, comme des systèmes en XX/XY où le chromosome Y est responsable de la détermination en mâle. Certaines espèces, comme le cichlidé *Metriaclima pyrsonotus*, possèdent même les deux paires de chromosomes sexuels à la fois (**Fig. 1.3.a.**). Le platy, *Xiphophorus maculatus*, ne possède qu'une paire de chromosomes sexuels, mais ceux-ci peuvent prendre la forme d'un X, d'un Y ou d'un W (**Fig. 1.3.b.**). Ces deux dernières formes de détermination du sexe peuvent être considérées comme polygéniques (PSD). La PSD peut prendre des formes complexes avec une combinaison additive ou épistatique d'une multitude de loci comme chez le zebrafish (**Fig. 1.6.**). Certains autosomes peuvent aussi influencer la détermination du sexe génétique comme chez le médaka *Oryzias latipes*, et engendrer des poissons XX mâles (détaillé plus tard, **Fig. 1.6.**).

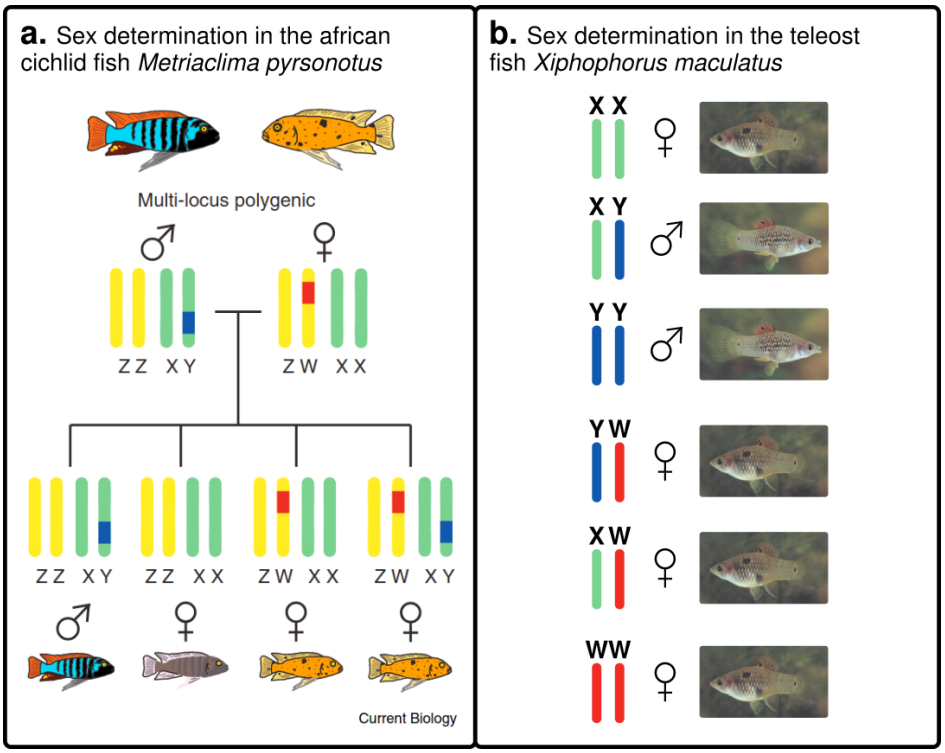

FIGURE 1.3 – **Détermination du sexe polygénique chez différentes espèces de poissons.** a. Détermination du sexe chez *Metriaclima pyrsonotus*. Deux paires de chromosomes sexuels sont présentes. Si les chromosomes W et Y sont présents, le chromosome W est dominant et active la détermination du sexe en femelle. Adapté de (MOORE et ROBERTS, 2013). b. Détermination du sexe chez *Xiphophorus maculatus*. Une seule paire de chromosomes sexuels est présente mais ces chromosomes peuvent prendre trois formes différentes, X, Y ou W. W déclenche la détermination en femelle et est dominant sur Y qui déclenche la détermination en mâle. Les mâles de cette espèce sont donc XY ou YY. ©Photographies de platy (BÖHNE *et al.*, 2009).

Sur les chromosomes sexuels se situent les gènes maîtres de la détermination du sexe, aussi appelés « interrupteurs » ; ils sont l'équivalent de *Sry* chez les mammifères. Leur présence et leur expression fait basculer la cascade de différenciation en faveur de l'un des deux sexes (voir partie suivante). Même si plusieurs espèces de poissons possèdent un système XX/XY, le gène maître n'est pas toujours le même (**Fig. 1.4.**) (KIKUCHI et HAMAGUCHI, 2013). Chez la truite arc-en-ciel, le gène maître est *sdY*, qui code pour une forme tronquée d'IRF9, une protéine impliquée dans la réponse anti-virale (YANO *et al.*, 2012). Chez le Fugu japonais, il s'agit d'*amhr2* qui code pour un récepteur à l'hormone anti-müllérienne permettant la détermination en mâle. Enfin, chez *Oryzias latipes*, le médaka japonais, le gène maître est *dmrt1by*, paralogue de *dmrt1*, un facteur de transcription impliqué dans le développement testiculaire. De manière intéressante, *dmrt1by* est trouvé chez seulement deux espèces du genre *Oryzias* (qui en compte 33), *O. curvinotus* et *O. latipes*. En revanche il est absent des autres *Oryzias* et de toutes les autres espèces de poissons téléostéens. *dmrt1by* est ainsi un excellent exemple pour illustrer la diversité des systèmes de détermination du sexe chez les poissons téléostéens ; il est également très récent du point de vue évolutif. De plus, chez *O. luzonensis*, l'espèce sœur de *O. curvinotus*, ce gène *dmrt1bY* a été perdu pour laisser place à *gsdfY*. Les médakas japonais présentent donc des mécanismes de détermination du sexe différents, ce qui en fait d'excellents modèles pour étudier l'évolution et la diversité du développement sexuel.

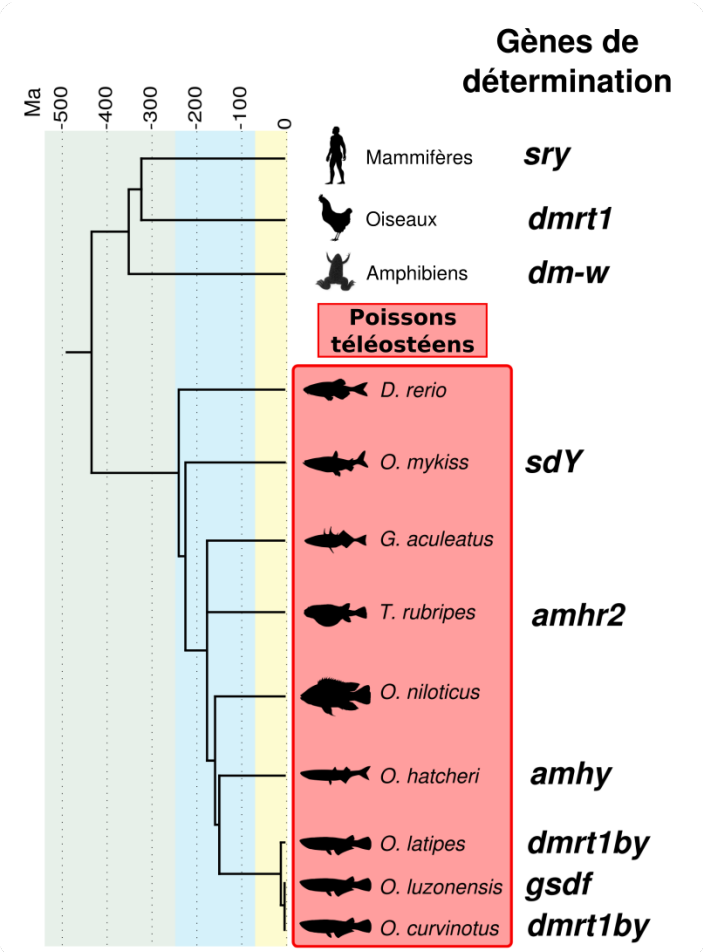

FIGURE 1.4 – **Phylogénie résumée des vertébrés montrant la diversité des gènes de détermination du sexe chez les poissons téléostéens.** Chez certaines espèces, le gène n'est pas connu et n'est pas indiqué sur la phylogénie. Adapté de Kikuchi et Hamaguchi, 2013 (KIKUCHI et HAMAGUCHI, 2013). ©Milton Tan - *O. niloticus* & *G. aculeatus* CC BY-NC-SA – http ://phylopic.org/.

Le médaka japonais : un modèle de choix pour étudier la détermination sexuelle

L'exemple du médaka japonais (**Fig. 1.5.**) permet d'illustrer en détail comment a évolué le mécanisme de détermination du sexe, mais aussi de voir pourquoi la limite entre GSD et ESD peut être parfois floue.

FIGURE 1.5 – **Médaka japonais, *Oryzias latipes*.** ©Seotaro - CC BY-SA 3.0.

L'évolution de *dmrt1by* sera détaillée dans le chapitre 3, où les liens qui existent entre éléments transposables et évolution du sexe seront présentés. *dmrt1by* est issu de la duplication de son paralogue ancestral *dmrt1*. En se dupliquant sur un autre chromosome, *dmrt1* a donné *dmrt1b*, et en devenant le gène maître de détermination du sexe, a fait évoluer l'autosome sur lequel il s'est inséré en chromosome Y. C'est pourquoi il est nommé *dmrt1by* (HERPIN *et al.*, 2010). Chez le médaka comme chez d'autres espèces de poissons, le sexe dépend du nombre de cellules germinales primordiales (PGC) lors de l'embryogenèse. Ce nom est donné aux cellules germinales avant qu'elles ne migrent dans la crête génitale. Si ce nombre de cellules est important, l'individu se différenciera en femelle. S'il est réduit, il se différenciera en mâle. *dmrt1by* est un inhibiteur de la prolifération cellulaire. S'il est exprimé au moment de la détermination, il limite la prolifération des PGC, ce qui aboutit à la différenciation en mâle (ADOLFI *et al.*, 2019). Chez l'adulte est exprimé son paralogue, *dmrt1a*. Ce facteur de transcription se fixe en amont de *dmrt1by* et empêche son expression. Ce mécanisme sera détaillé dans la section 1.3, page 31.

Chez le médaka *O. latipes*, la détermination du sexe est généralement considérée comme génétique avec système XX/XY (HERPIN *et al.*, 2010; MATSUDA, 2005; NANDA *et al.*, 2002). Cependant, des études récentes montrent que la distinction entre ESD et GSD est parfois floue dans cette espèce. Des températures élevées de l'eau (32°C contre 26°C-28°C en conditions normales) aboutissent à des niveaux de cortisol élevés dans les embryons qui ont pour conséquence d'activer l'expression de *dmrt1a*. Chez les femelles, l'activation de cette expression conduit *dmrt1a* à prendre le rôle de *dmrt1bY*, ce qui inhibe la prolifération des PGC. Cela conduit à l'apparition d'individus XX mâles avec des testicules dépourvus de cellules germinales (ADOLFI *et al.*, 2019). C'est pourquoi il est préférable de parler de gradient entre GSD et ESD et de positionner les différentes espèces le long

de ce gradient, bien que certaines soient plus proches de l'un des deux extrêmes (**Fig. 1.6.**).

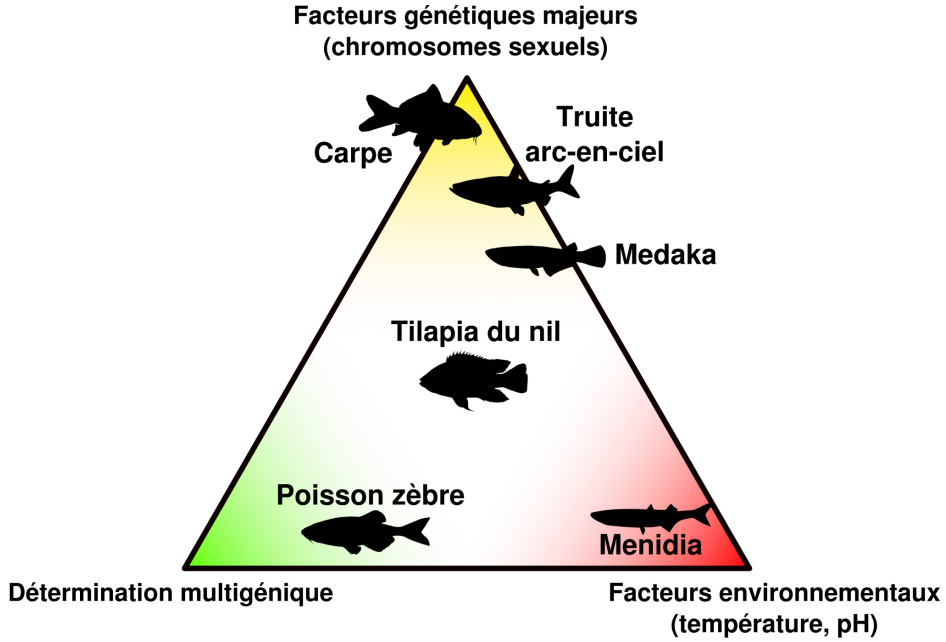

FIGURE 1.6 – **Positionnement de différentes espèces de poissons téléostéens selon leur système de détermination sexuelle.** Le médaka par exemple possède des chromosomes sexuels avec un chromosome Y responsable de la détermination en mâle ; cependant, en cas de température élevée de l'eau, les individus XX ne possédant pas ce chromosome se différencient quand même en mâles. La séparation entre espèces GSD et ESD n'est pas triviale et les différentes espèces se positionnent sur un gradient. Adapté de Baroiller et al. 2009 (BAROILLER *et al.*, 2009). ©Milton Tan - *O. niloticus* CC BY-NC-SA – http ://phylopic.org/.

1.1.3.3 La différenciation et le maintien du sexe

Suite à la détermination sexuelle, la gonade indifférenciée devient un testicule ou un ovaire. Le gène maître de détermination du sexe joue le rôle de l'interrupteur qui déclenche la différenciation. On parle ensuite de cascade de différenciation sexuelle. Une série de gènes va être activée ou réprimée, induisant la différenciation de la gonade. Deux cascades sont en compétition : la cascade spécifique du testicule, et celle spécifique de l'ovaire.

Les cascades de différenciation sexuelle

Selon le modèle proposé par Graham en 2003 (GRAHAM *et al.*, 2003), les gènes maîtres de la détermination du sexe pourraient être variables, alors que les gènes de la cascade de différenciation seraient, eux, conservés : « les maîtres changent, et les esclaves restent ». Plusieurs voies très conservées semblent soutenir ce modèle. Ainsi le récepteur à protéine G fixant le ligant codé par *rspo1* et activant la voie de *wnt4* est présent chez la plupart des vertébrés. De même, la voie de *wnt4* est conservée chez les vertébrés ; c'est historiquement la première voie découverte comme impliquée dans la cascade de différenciation en femelle. Elle joue un double rôle de différenciation en femelle et d'inhibition du développement des testicules. La β-*catenin*, un coactivateur de transcription impliqué dans la voie de *wnt4* est aussi retrouvée chez la plupart des vertébrés (EGGERS *et al.*, 2014; HERPIN et SCHARTL, 2015; ORTEGA-RECALDE *et al.*, 2019). Parallèlement, les principaux gènes trouvés dans la cascade spécifique des mâles, comme *sox9* et *dmrt1* qui codent pour des facteurs de transcription, *sf1* qui code pour un facteur d'épissage, et *fgf9* qui code pour un facteur de croissance,

sont retrouvés chez la plupart des vertébrés. De la même manière que pour les voies femelles, ces gènes s'activent en cascade et inhibent les voies femelles ; ainsi, *sox9* et *fgf9* inhibent l'expression de *rspo1* et *wnt4* respectivement (EGGERS *et al.*, 2014; HERPIN et SCHARTL, 2015; ORTEGA-RECALDE *et al.*, 2019). Cette répression est réciproque puisque *rspo1* et la voie de *wnt4* inhibent l'expression de *sox9* et *fgf9*. Le gène maître de déterminisme du sexe sert d'interrupteur : il permet de basculer vers l'une des voies de différenciation mâle ou femelle, l'autre étant la voie par défaut activée en l'absence d'action du gène maître. De nouveaux gènes maîtres de détermination du sexe sont parfois recrutés depuis la cascade d'activation et placés à son sommet. C'est le cas par exemple de *dmrt1by* chez le médaka *O. latipes*. *dmrt1*, dont il est issu par duplication, est l'un des chainons de la cascade de différenciation du médaka et d'autres espèces. Des contre-exemples existent aussi comme *sdY* chez les salmonidés qui a été recruté depuis le système immunitaire. Des études récentes ont montré que les cascades de différenciation sexuelle, bien que plus conservées que les gènes interrupteurs, sont en fait aussi sujettes à évolution. Des gènes apparaissent (comme *dmrt1by*), d'autres changent de fonction (*rspo1* chez le poisson zèbre est impliqué dans le développement testiculaire (ZHANG *et al.*, 2011)), ou bien sont perdus au cours de l'évolution (comme *sox9* qui joue un rôle clé dans la différenciation en mâle chez les mammifères, mais qui est totalement absent de cette voie de régulation chez le médaka (YOKOI *et al.*, 2002)). Une nouvelle version plus nuancée du paradigme de Graham a ainsi été proposée par Herpin et Schartl : «Quand les maîtres changent, certains esclaves restent, d'autres sont éliminés ou acquièrent de nouvelles tâches, et de nouveaux peuvent être engagés.» (HERPIN et SCHARTL, 2015).

Le maintien du sexe chez l'adulte

Le maintien du sexe chez l'adulte correspond aux mécanismes qui permettent le bon fonctionnement de la gonade différenciée. Certains gènes et voies d'activation, comme *dmrt1* ou *foxl2*, s'ils sont inactivés, peuvent induire la stérilité des gonades chez les mammifères, voire même induire des changements de sexe chez certains poissons téléostéens (HUANG *et al.*, 2017; SCHARTL *et al.*, 2018). Chez les mammifères, la suppression de *foxl2* de l'ovaire adulte entraîne la différenciation de cellules de la granulosa en cellules de Sertoli et l'apparition de structures testiculaires. A l'inverse, la suppression de *dmrt1* dans les testicules de mammifères génère l'apparition de structures ovariennes (HUANG *et al.*, 2017). Les réseaux de régulation de gènes sont donc nécessaires pour maintenir la gonade adulte dans un état fonctionnel.

1.1.4 Dimorphisme sexuel et expression des gènes

1.1.4.1 Origine des dimorphismes sexuels

Dans beaucoup d'espèces il est possible de différencier phénotypiquement les mâles et les femelles. Ces différences phénotypiques peuvent être d'ordre morphologique (coloration, taille, forme ...), physiologique ou comportemental (MANK, 2009). Pourtant, mâles et femelles sont génétiquement quasi-identiques, plus particulièrement chez les espèces qui n'ont pas de déterminisme génétique du sexe, chez les hermaphrodites séquentiels, et plus largement, chez les espèces sans chromosomes sexuels. La plupart des espèces ayant un déterminisme génétique avec des chromosomes sexuels ne diffèrent que par quelques gènes localisés sur ces chromosomes sexe-spécifiques, gènes qui ne suffiraient pas à expliquer toutes les différences observées entre mâles et femelles

(ELLEGREN et PARSCH, 2007). Ces observations impliquent que la majorité des différences observées entre mâles et femelles proviennent des gènes autosomaux et plus particulièrement du fait de leur niveau d'expression différent entre les sexes. Un gène plus exprimé chez les femelles que les mâles est un gène femelle-biaisé, alors qu'un gène plus exprimé chez les mâles est un gène mâle-biaisé.

1.1.4.2 La détection des gènes sexe-biaisés

Les analyses transcriptomiques ont révélé que le sexe est un facteur significatif à prendre en compte dans les analyses d'expression des gènes. On observe en effet un nombre particulièrement élevé de gènes sexe-biaisés (SB) (*i.e.* différentiellement exprimés entre mâles et femelles), aussi bien dans le soma que dans les gonades (MANK et ELLEGREN, 2009). Les études faites chez les poissons montrent que les gonades restent le tissu ayant le plus de gènes SB (BAR *et al.*, 2016; BÖHNE *et al.*, 2014; LIU *et al.*, 2015; ROBLEDO *et al.*, 2015; TAO *et al.*, 2018; TSAKOGIANNIS *et al.*, 2018a; WANG *et al.*, 2017; ZENG *et al.*, 2016) par rapport à d'autres organes comme le cerveau (BEAL *et al.*, 2017; BÖHNE *et al.*, 2014; LIU *et al.*, 2015; SAARISTO *et al.*, 2017; SHEN *et al.*, 2020; TSAKOGIANNIS *et al.*, 2018a; WU *et al.*, 2019). Cependant il est très difficile d'estimer si le nombre de gènes SB est conservé ou variable entre les espèces ou les tissus. De nombreux paramètres biologiques et expérimentaux influencent cette mesure. Le nombre de gènes détectés dépend de l'espèce, de l'âge, de la maturité sexuelle, du type de reproduction, du tissu, de la méthode de séquençage, du nombre de réplications de l'expérience, des programmes informatiques, et des seuils statistiques utilisés (ELLEGREN et PARSCH, 2007). Tous ces paramètres rendent difficile la comparaison du nombre de gènes SB, qui peuvent facilement être variables d'une étude à l'autre même si les mêmes données sont utilisées. Cela souligne l'importance de la reproductibilité des expériences, autant dans la partie expérimentale avec la matière biologique que pour la partie bioinformatique où les scripts et paramètres des programmes utilisés doivent être accessibles afin de discuter les potentielles différences de résultats obtenus.

1.1.4.3 La résolution des conflits sexuels

Certains gènes présentent un antagonisme sexuel. Leur expression peut être simultanément bénéfique pour un sexe, et néfaste pour l'autre. C'est le cas du gène Artemis (*arts*) chez *Drosophila melanogaster* qui est essentiel à la fertilité des femelles, mais défavorable à la fertilité des mâles (VANKUREN et LONG, 2018). Le rôle exact d'Artemis est inconnu mais il possède des caractéristiques de β-importines, ces dernières étant impliquées dans le transport de protéines vers le noyau cellulaire. Des mutations qui viendraient augmenter ou réduire l'expression de ce gène seraient bénéfiques pour un sexe mais néfastes pour l'autre. Pour résoudre ce conflit, la réponse à court terme est un compromis entre les deux optimums, avec une expression intermédiaire du gène. Dans ce cas, le gène garde la même expression entre mâles et femelles. Mais sur un plus long terme, des régulations sexe-spécifiques de l'expression du gène peuvent permettre d'activer le gène dans le sexe où il est bénéfique et de réduire son expression là où il est néfaste. Alors, le gène devient SB, ce qui permet de résoudre le conflit sexuel (ELLEGREN et PARSCH, 2007). Il est aussi possible que le gène sous conflit soit dupliqué, et subisse une forme de sous-fonctionnalisation, où chaque copie prend en charge des fonctions distinctes. Elles peuvent acquérir des mutations, se spécialiser et n'être exprimées plus que dans un sexe. Cela s'est produit pour les gènes *arts* et *apl* chez *D.*

melanogaster (VANKUREN et LONG, 2018). Un gène déjà SB peut aussi être dupliqué et subir une néo-fonctionnalisation, et prendre en charge une nouvelle fonction sexe-spécifique (CUTTER et WARD, 2005; GNAD et PARSCH, 2006). L'avantage ici est que la régulation du gène peut être déjà présente si les régions régulatrices sont dupliquées avec le gène. Enfin, certains gènes peuvent acquérir une régulation sexe-spécifique qui pourra être sélectionnée si elle apporte un avantage à l'organisme. C'est souvent le cas des rétrogènes, des gènes dérivés de la rétrotranscription de leur ARN messager, qui sont insérés dans le génome sans leurs séquences régulatrices (LONG *et al.*, 2013; VINCKENBOSCH *et al.*, 2006).

Les gènes SB peuvent ainsi être considérés comme des témoins de conflits sexuels passés, car la plupart d'entre eux sont issus de la résolution de ce type de conflit (ELLEGREN et PARSCH, 2007). La régulation de l'expression des gènes n'est pas la seule façon de résoudre un conflit sexuel. Un gène transposé sur le chromosome Y, chez les mammifères, ne sera présent et exprimé que chez les mâles, ce qui peut aussi participer à résoudre le conflit. L'expression différentielle d'un gène ne résulte donc pas seulement de mécanismes de régulation au niveau transcriptionnel ; elle peut aussi être dose-dépendante, c'est-à-dire dépendre du nombre de copies du gène (ici présence / absence). Chez les espèces diploïdes tous les gènes portés par les autosomes sont en 2 copies. En revanche les gènes liés à l'X chez les mammifères ne sont en 2 copies que chez les femelles. De nombreux organismes régulent l'expression des gènes présents sur les chromosomes sexuels pour maintenir une expression équivalente entre mâles et femelles. C'est le cas de l'inactivation du X chez les mammifères, où l'un des deux chromosomes X est inactivé aléatoirement dans chaque cellule (voir section 1.3, page 31). Si certains gènes portés par les chromosomes sexuels échappent aux mécanismes de régulation du dosage, ils peuvent être détectés comme SB car deux fois plus exprimés dans le sexe homochromosomique (ELLEGREN et PARSCH, 2007). Pour ces gènes il est difficile de savoir si leur expression différentielle résulte de la résolution d'un conflit, ou d'un effet de la dose.

1.1.4.4 Les caractéristiques des gènes sexe-biaisés

Les gènes SB possèdent des caractéristiques communes. Tout d'abord, chez de nombreuses espèces il a été montré que les gènes SB évoluent plus vite que les autres gènes, en comparant le taux de mutations non-synonymes à celui des mutations synonymes. Cette évolution rapide plus accentuée pour les gènes mâle-biaisés, est valable dans différentes lignées très éloignées : drosophile (ASSIS *et al.*, 2012), *C. elegans* (CUTTER et WARD, 2005), poissons (YANG *et al.*, 2016) et primates (KHAITOVICH *et al.*, 2005). Si l'on compare les gènes mâle-biaisés avec les gènes femelle-biaisés on observe que les gènes mâle-biaisés ont une variabilité d'expression entre espèces plus forte que les gènes femelle-biaisés chez les drosophiles (HUTTER *et al.*, 2008; HUYLMANS *et al.*, 2017; MÜLLER *et al.*, 2012; ZHAO *et al.*, 2015; ÁVILA *et al.*, 2015), ce qui n'est pas une généralité comme observé chez le moustique où les gènes femelle-biaisés ont une plus forte variabilité d'expression (PAPA *et al.*, 2017). La localisation des gènes SB est aussi particulière, avec une surreprésentation des gènes femelle-biaisés sur le chromosome X par rapport aux autosomes chez la drosophile et la souris (MEISEL *et al.*, 2012). Cette distribution particulière ne semble cependant pas universelle avec le contre-exemple des mouches *diopsidae* (WOLFENBARGER et WILKINSON, 2001). On peut donc envisager qu'il existe des réseaux de co-régulation de gènes, qui permettraient de résoudre

plusieurs conflits sexuels à la fois dans les génomes, en regroupant physiquement des gènes qui suivent le même patron d'expression. Cette hypothèse sera testée dans la seconde partie de ce manuscrit.

Nous avons vu que le développement sexuel des poissons évolue particulièrement vite. De plus il existe une grande diversité de types de reproduction sexuelle et une grande variabilité dans le développement sexuel. Dans le chapitre suivant, nous allons voir que le sexe n'est pas la seule facette hypervariable des poissons.

1.2 Les éléments transposables, arme à double tranchant des génomes

La biologie moléculaire étudie les macromolécules du vivant : l'ADN, l'ARN et les protéines. La théorie [1] fondamentale de la biologie moléculaire, décrite en 1957 par Francis Crick (COBB, 2017; CRICK, 1958), associe un gène (une portion de l'ADN) à la production d'un ARN, qui est ensuite traduit en une protéine. Les protéines ainsi produites ont une fonction précise pour la cellule ou l'organisme. Dans le génome humain, 2% des 3.2 milliards de bases d'ADN font partie de la portion codante du génome, qui est traduite en protéines. Ces 2% du génome ont majoritairement retenu l'attention des scientifiques et le reste du génome a été pendant longtemps considéré comme « ADN poubelle ».

1.2.1 La découverte des éléments transposables

En 1951, Barbara McClintock met en évidence que des « éléments de contrôle » sont responsables de la coloration variable des grains d'épis de maïs (MCCLINTOCK, 1950). Cette découverte lui vaudra un prix Nobel en 1983. Renommées en «éléments transposables (ET)», ces séquences vont ensuite être découvertes chez la drosophile et l'homme dans les années 70 (OHNO, 1972). Ces séquences ne codent pas forcément pour des protéines, et lorsqu'elles en codent celles-ci ne sont pas nécessaires au fonctionnement de la cellule; les ET ont donc été considérés au début comme de « l'ADN poubelle », ce qui a depuis été largement remis en cause. Le nombre d'études traitant d'ET a augmenté à la fin des années 80 quand la communauté scientifique a commencé à s'y intéresser dans un contexte d'évolution du génome, puis a connu un nouveau rebond dans les années 2000 avec la première séquence du génome humain (**Fig. 1.7.a.**). La base de données RepeatMasker estime aujourd'hui que le génome humain est composé à 48.5% d'ET (repeatmasker.org/species/hg.html). Cependant, la proportion d'articles traitant des éléments transposables par rapport à ceux traitant de génomes est stable voire en faible diminution depuis le début des années 2000 (**Fig. 1.7.b.**).

1.2.2 Qu'est-ce qu'un élément transposable?

Les ET sont des séquences d'ADN qui ont la particularité de pouvoir s'insérer dans les génomes. Ils sont généralement répétés, de quelques copies à 1.2 millions de copies pour les éléments de type *Alu* chez l'humain (DEININGER, 2011). Cette caractéristique permet de les regrouper en familles : une famille d'ET correspond à toutes les copies qui proviennent d'une même séquence ancestrale qui s'est multipliée dans un génome. Certaines copies peuvent perdre leur capacité à se répliquer ou se déplacer suite à des mutations, mais les copies autonomes qui restent intègres peuvent parfois les mobiliser à travers le génome.

1. On parle parfois de dogme, mais un dogme par définition n'est pas discutable, il est considéré comme une vérité absolue et incontestable. A l'inverse, une théorie se confronte aux données en attendant d'être infirmée et peut évoluer (*cf.* épissage alternatif, un gène peut donner plusieurs protéines différentes). Il faut réserver le mot dogme à des domaines pseudo-scientifiques où les principes fondateurs sont immuables et refusent tout questionnement scientifique (*cf.* dogme homéopathique).

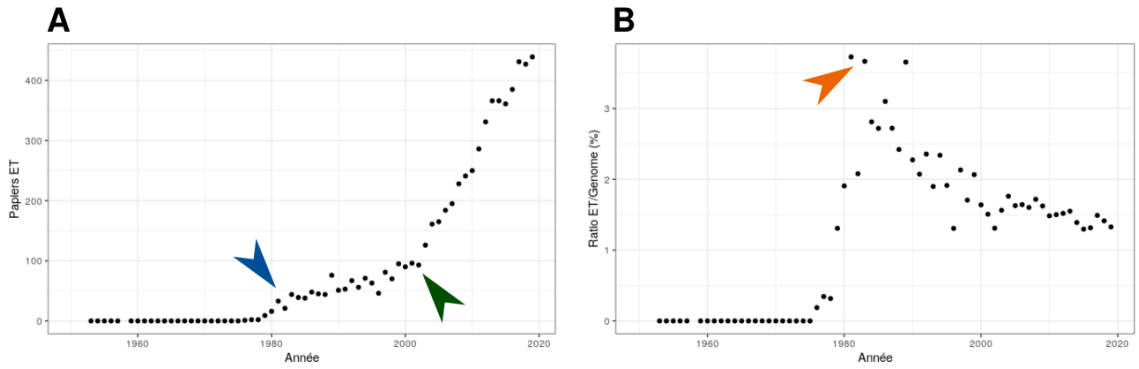

FIGURE 1.7 – **Evolution de la recherche autour des éléments transposables.** a. Evolution du nombre d'articles traitant d'éléments transposables (*source : Pubmed, transposable elements[title/abstract]*). On observe une première augmentation dans les années 80 (flèche bleue) puis une seconde dans les années 2000 (flèche verte). b. Evolution du ratio d'articles traitant d'éléments transposables sur les articles traitant de génomes (*source Pubmed, genome[title/abstract]*). Cette normalisation permet de comprendre si l'augmentation du nombre d'articles traitant d'ET est liée à une augmentation de l'intérêt pour les ET ou une augmentation en général de la recherche scientifique sur les génomes. On observe un pic dans les années 80 (flèche orange) correspondant au début de l'intérêt qui leur est porté par la communauté scientifique, puis une stabilisation avec une légère décroissance depuis le début des années 2000. La seconde augmentation du nombre d'articles traitant d'ET est donc due à une augmentation générale de la recherche sur les génomes qui a accompagné le développement de nouvelles techniques de séquençage.

1.2.3 Éléments transposables et taille des génomes

La taille des génomes des eucaryotes est très variable selon les espèces. En comparaison avec les procaryotes où l'on observe un facteur 20 entre les plus petits et les plus grands génomes, chez les eucaryotes ce facteur est de 200 000 (GREGORY, 2001). Aucun facteur unique comme la taille des parties codantes du génome ou la « complexité » d'un organisme ne corrèle avec la taille du génome (KIDWELL, 2002). C'est ce qui est appelé le paradoxe de la valeur C. La valeur C est une mesure représentant la quantité d'ADN dans un noyau haploïde (en picogrammes) ; elle rend compte de la taille d'un génome, déterminée par densitométrie. Aujourd'hui, avec le séquençage de nouvelle génération, il est possible de calculer la taille d'un génome en nombre de nucléotides directement, sans avoir à passer par la mesure de la valeur C. Cette taille en nombre de nucléotides est fiable sous réserve d'avoir séquencé et assemblé l'intégralité du génome, ce qui n'est pas forcément évident du fait des régions riches en répétitions.

En fonction des espèces eucaryotes, le pourcentage du génome couvert par les ET est très variable : 10-40% chez la drosophile (KIM *et al.*, 2020; SESSEGOLO *et al.*, 2016), environ 50% chez l'homme, et 85% chez certaines plantes comme le maïs. Plusieurs études récentes ont commencé à apporter des réponses au paradoxe de la valeur C. Chez plusieurs espèces de drosophiles, la taille du génome est corrélée au pourcentage d'ET (SESSEGOLO *et al.*, 2016). Cette observation a aussi été faite chez les vertébrés, tout comme chez les urochordés et plus généralement chez les eucaryotes (CHALOPIN *et al.*, 2015; CHÉNAIS *et al.*, 2012; KIDWELL, 2002; NAVILLE *et al.*, 2019; VITTE et PANAUD, 2005).

1.2.4 L'impact des éléments transposables sur les génomes

1.2.4.1 La valeur sélective

Pour parler d'impact sur un génome, il faut définir le terme de valeur sélective (*fitness* en anglais). La valeur sélective d'un individu reflète sa capacité à transmettre ses gènes à la génération suivante, et se décrit généralement par le produit de sa survie par sa fécondité. Pour « bien transmettre » ses gènes, un individu doit survivre jusqu'à être en âge de se reproduire, et doit se reproduire. La valeur sélective étant déterminée en partie par le génome de l'individu, les ET qui s'y déplacent ou s'y répliquent vont donc influencer la valeur sélective de l'hôte.

1.2.4.2 Les effets délétères des éléments transposables sur l'organisme

Les effets délétères des ET varient : ils peuvent induire des mutations létales, ou juste diminuer faiblement la valeur sélective de l'hôte. Les causes de ces changements de valeur sélective peuvent être nombreuses. La plus intuitive est l'insertion d'un ET dans la partie codante d'un gène essentiel à l'hôte. Mais les ET peuvent aussi s'insérer dans des régions régulatrices, des régions qui contrôlent la structure 3D de la chromatine, ou dans des gènes moins essentiels à l'hôte. D'autres effets non pas liés à leur insertion mais à leurs caractéristiques peuvent exister. La similarité de séquence entre copies d'une même famille peut entraîner des recombinaisons homologues non-alléliques et différents types de réarrangements chromosomiques comme des délétions ou des duplications. Certaines maladies génétiques sont causées par des ET chez l'humain. Par exemple, la protéine Facteur IX qui joue un rôle dans la cascade de la coagulation sanguine n'est pas produite en assez grande quantité chez les patient hémophiles de type B. Cette maladie conduit à un grave handicap invalidant, pouvant aller jusqu'à la paralysie. Chez certains patients, on détecte l'insertion d'un élément de type *Alu* dans l'exon 8 du gène du Facteur IX (Li *et al.*, 2001). La conséquence de cette insertion est la formation d'un codon STOP prématuré qui interrompt la partie codante du gène. Chez l'humain toujours, il a été montré que l'augmentation de l'activité de transposition d'ET de type L1 dans les neurones humains est lié à la susceptibilité d'apparition de la schizophrénie (Bundo *et al.*, 2014). Les conséquences négatives des ET pour la valeur sélective de l'hôte ont conduit à les considérer comme des éléments égoïstes, parasites des génomes (Hickey, 1982). Certaines insertions peuvent cependant aussi être neutres, par exemple si elles s'insèrent dans des régions pauvres en gènes. Il est souvent difficile d'affirmer définitivement qu'une insertion est neutre du fait de la multitude d'effets qu'elle peut avoir, comme décrit précédemment. Nous verrons par ailleurs dans les parties suivantes que les ET peuvent aussi avoir un impact positif. Avant cela, il nous faut aborder la classification des ET ainsi que leurs différents mécanismes de transposition.

1.2.5 Les différentes familles d'éléments transposables et leurs mécanismes de transposition

Dans les parties précédentes, les familles *Alu* ou L1 ont déjà été évoquées. Une famille d'ET est définie comme un ensemble de copies issues de la même séquence ancestrale. Il existe en réalité d'autre niveaux de classification qui regroupent les familles entre elles (Kapitonov et Jurka, 2008; Seberg et Petersen, 2009; Wicker *et al.*, 2007). Pour classer les ET, la première caractéristique utilisée est leur mécanisme de transposition.

1.2.5.1 Les éléments transposables de classe I : les rétrotransposons (Fig. 1.8.)

Mécanisme de transposition

Les rétrotransposons regroupent les ET capables de transposer par la rétro-transcription en ADN complémentaire d'un intermédiaire ARN. On parle d'un mécanisme de transposition par « copier-coller » (HAN et BOEKE, 2005; KUMAR et BENNETZEN, 1999; SABOT et SCHULMAN, 2006; SANMIGUEL *et al.*, 1996). Un rétrotransposon autonome code pour les protéines lui permettant d'être mobilisé. Des copies ayant perdu les séquences codantes pour ces protéines (copies dites « non-autonomes ») peuvent être mobilisées en *trans* grâce à des copies autonomes partageant des homologies de séquence, qui produisent à leur place les protéines nécessaires. C'est notamment le cas des éléments SINE (Short Interspersed Nuclear Elements) qui sont mobilisés par les éléments LINEs (Long Interspersed Nuclear Elements) (DEWANNIEUX *et al.*, 2003; KAJIKAWA et OKADA, 2002; KRAMEROV et VASSETZKY, 2005; WICKER *et al.*, 2007). A l'intérieur de la classe I, on trouve différentes superfamilles regroupées en fonction de leur structure et des protéines utilisées pour transposer. La principale classification est faite en se basant sur la présence ou l'absence de longues répétitions terminales (ou LTR, pour « Long Terminal Repeat ») (JURKA *et al.*, 2005; WICKER *et al.*, 2007).

Les principales superfamilles de rétrotransposons

Les LTR sont des séquences répétées de quelque centaines de nucléotides présentes à chaque extrémité des rétrotransposons de type LTR lorsqu'ils sont complets (COFFIN, 1992; KUMAR et BENNETZEN, 1999; VOYTAS et BOEKE, 1992). Ils contiennent notamment des séquences promoteur et enhancer nécessaires à l'expression de l'élément. La partie interne regroupe plusieurs séquences codantes permettant de produire des protéines essentielles à la mobilisation de l'élément (WICKER *et al.*, 2007). Ces protéines ainsi que la structure du rétrotransposon ressemblent fortement à celles des rétrovirus. On retrouve une région avec les cadres de lecture *Gag* et *Pol*. *Gag* (Group antigen) permet la formation d'une capside et le repliement de l'ARN messager (ARNm) du rétrotransposon à l'intérieur de celle-ci (SHIBA et SAIGO, 1983). La région *Pol* contient un unique cadre de lecture qui code pour plusieurs protéines. Il code une protéase (AP) qui permet de couper la polyprotéine produite en plusieurs peptides (MÉREL *et al.*, 2020), une rétrotranscriptase (RT) qui rétro-transcrit l'ARN de l'ET en ADN complémentaire, une ribonucléase H (RNAseH) qui permet d'hydrolyser l'hybride ADN/ARN formé après la rétro-transcription de l'ARN, et enfin une intégrase (Int) qui est nécessaire à l'intégration de l'ADN complémentaire produit dans le génome (MÉREL *et al.*, 2020; WICKER *et al.*, 2007). Un domaine *Env* est aussi retrouvé chez certains éléments ERV (Rétrovirus endogènes). Ces éléments dérivent de rétrovirus ayant perdu leur mobilité extracellulaire par la délétion ou l'inactivation du gène *Env* (GIFFORD *et al.*, 2018). La ressemblance des rétrovirus avec les ET à LTR a soulevé la question de leur origine commune (LERAT et CAPY, 1999). L'hypothèse prédominante suggère que les rétrovirus pourraient avoir comme origine des ET à LTR ayant acquis une région *Env* fonctionnelle. D'autres suggèrent que des transitions multiples entre ET à LTR et rétrovirus sont possibles, quand certains rétrovirus perdent leur région *Env* ou en gagnent une (MCCLURE, 1991).

Une autre grande catégorie d'ET de classe I correspond aux rétrotransposons non-LTR. Ils sont aussi classés en plusieurs superfamilles, notamment selon la présence ou non de cadres de lecture. Les éléments LINE autonomes codent pour une rétrotranscriptase (RT) et une endonucléase (EN)

(EICKBUSH et MALIK, 2002; WICKER *et al.*, 2007). A la différence des éléments LTR, les LINE sont rétro-transcrits directement au site d'intégration de l'ET. Les SINE ont une origine distincte des autres rétrotransposons et dérivent de différentes rétrotranscriptions de transcrits de polymérase III (KRAMEROV et VASSETZKY, 2005). Les *Alu*, la famille de SINE la plus répandue chez l'humain et chez d'autres primates (1 million de copies et 11% du génome humain), dérivent d'ARN 7SL (DEININGER, 2011). Comme les SINE en général, ils ne possèdent pas de cadre de lecture, mais un promoteur interne pour la polymérase III permettant leur expression (KRAMEROV et VASSETZKY, 2005). Ils sont mobilisés en *trans* par des LINE autonomes qui permettent leur rétro-transcription et leur intégration.

Ces superfamilles de rétrotransposons ne sont pas exhaustives. Il existe aussi notamment les éléments DIRS, qui ressemblent aux LTR mais possèdent une tyrosine recombinase plutôt qu'une intégrase, et sont caractérisés par une absence de LTR (POULTER et GOODWIN, 2005; WICKER *et al.*, 2007).

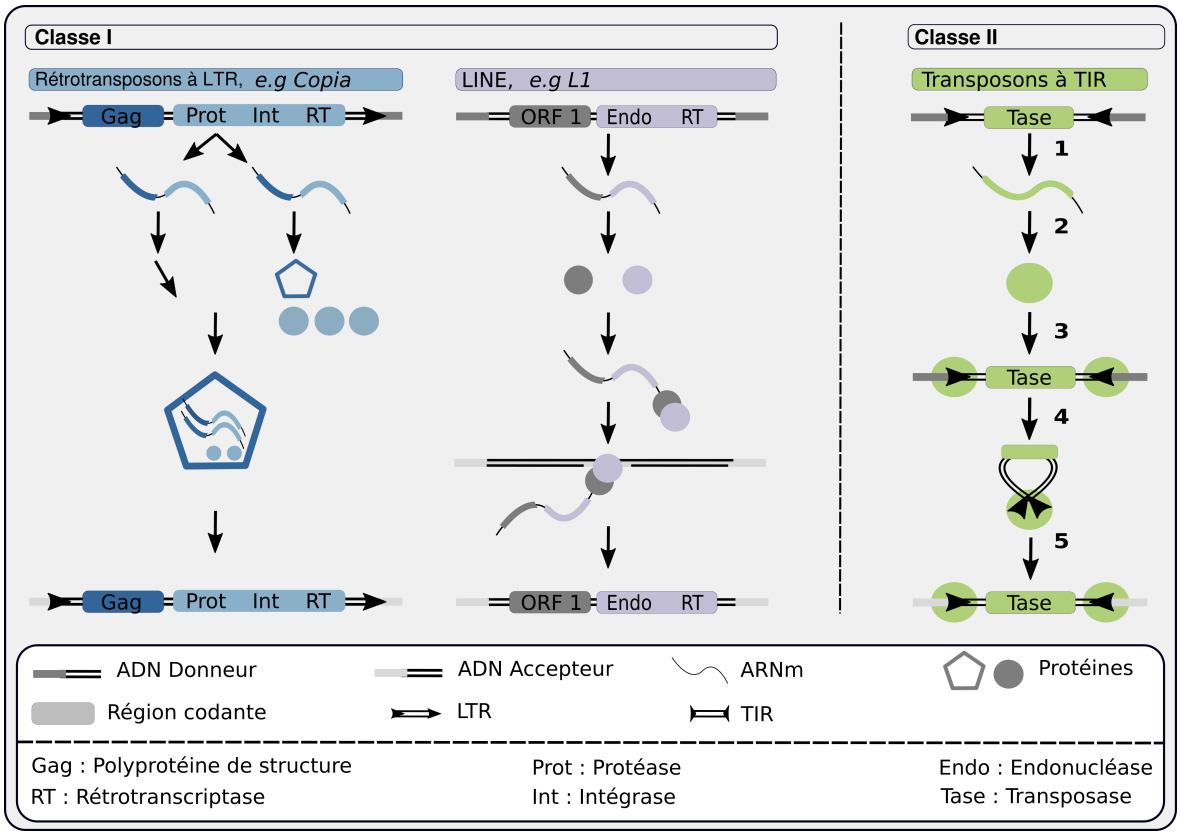

FIGURE 1.8 – **Mécanismes de transposition des principaux types d'ET.** Les ET sont généralement séparés en deux classes, selon qu'ils nécessitent la rétrotranscription d'un intermédiaire ARN (classe I) ou pas (classe II). Les protéines nécessaires à la rétrotransposition des éléments de classe I sont codées par l'ARNm produit par l'ET dans le cas d'un élément autonome. La protéine *Gag* permet de regrouper les transcrits et les différentes protéines produites avant la rétrotranscription en ADN et l'insertion à un nouveau locus. Les rétrotransposons non-LTR comme les LINE utilisent le même principe mais la rétrotranscription se fait directement au site d'intégration par le biais de la rétrotranscriptase et de l'endonucléase. Les transposons à ADN (classe II) s'excisent de leur locus et se réintègrent ailleurs dans le génome. Ils utilisent une transposase dont ils possèdent la séquence codante. Adapté de Mérel *et al.* (MÉREL *et al.*, 2020).

1.2.5.2 Les éléments transposables de classe II : les transposons ADN (Fig. 1.8.)

On utilise souvent par abus de langage le mot transposon pour parler d'ET. Cependant le nom transposon correspond aux ET de classe 2, et exclut les rétrotransposons. Pour éviter les confusions, on précise souvent transposon « ADN ». Le point commun de ces ET est de ne pas utiliser d'intermédiaire ARN rétro-transcrit pour transposer. Les ET de classe 2 sont divisés en deux sous-catégories : ceux qui utilisent une cassure double brin de l'ADN pour transposer, et ceux qui utilisent une cassure simple brin. Les copies autonomes de la première sous-classe sont caractérisées par la présence de répétitions terminales inversées (TIR, pour « Terminal Inverted Repeat ») et d'une séquence codant pour une transposase. La transposase reconnaît les séquences TIR du transposon, l'excise de sa position en formant un dimère avec les TIR, et l'insère ensuite à une autre position (SHAO et TU, 2001; TANG *et al.*, 2005; WICKER *et al.*, 2007; ZHANG *et al.*, 2001). On parle ici de mécanisme de transposition par « couper-coller ». Les transposons peuvent tout de même se multiplier en transposant lors de la réplication de l'ADN : en se déplaçant d'une région répliquée à une région non-encore répliquée, un transposon initialement situé à un locus se retrouve à un locus de plus sur l'une des chromatides sœurs. La deuxième sous-classe de transposons ADN correspond à des ET qui ne coupent qu'un brin d'ADN pour transposer (comme les Hélitrons) et dont le mécanisme de transposition ne sera pas détaillé ici.

1.2.6 Le côté lumineux des éléments transposables

Jusqu'ici, seul le côté délétère ou neutre des ET a été abordé. Leur capacité à se répliquer et leur nature répétée ont des conséquences importantes sur l'intégrité et le fonctionnement des génomes. Cependant, ces séquences ne possèdent pas que des côtés négatifs pour le génome hôte. Bien que l'impact de la majorité des nouvelles insertions soit considéré comme neutre ou néfaste pour le génome hôte, une partie non négligeable a un impact positif sur la valeur sélective des hôtes. Il aura fallu attendre un demi-siècle après leur découverte pour montrer que les ET jouent un rôle important dans l'évolution des génomes mais aussi dans le contrôle de l'expression des gènes (BIÉMONT et VIEIRA, 2006; BOURQUE, 2009).

1.2.6.1 Les éléments transposables sont impliqués dans certains réarrangements adaptatifs des génomes

Les réarrangements génomiques comme les duplications segmentaires (duplications de portions de chromosomes) jouent un rôle important dans l'évolution des génomes, notamment chez les primates (BAILEY et EICHLER, 2006; EMANUEL et SHAIKH, 2001). Barbara McClintock suspectait déjà les ET capables de faciliter ce type de réarrangements (MCCLINTOCK, 1950). Les recombinaisons entre éléments *Alu* ou L1 du génome humain ancestral auraient ainsi généré des duplications segmentaires (BAILEY *et al.*, 2003; HAN *et al.*, 2008; KIM *et al.*, 2008). De plus, on estime que presque la moitié des inversions chromosomiques entre l'humain et le chimpanzée seraient dues à des éléments *Alu* ou L1 (LEE *et al.*, 2008). Tous ces réarrangements cependant ne sont pas délétères : certains ont été conservés au cours de l'évolution et constituent un mécanisme d'évolution des génomes (BOURQUE, 2009).

1.2.6.2 Les éléments transposables influencent l'expression et la structure des gènes

Les ET peuvent être vus comme une source de variabilité et de diversité pour les génomes (CHUONG *et al.*, 2016; REBOLLO *et al.*, 2012b; TRIZZINO *et al.*, 2018). Leur insertion peut entraîner l'apparition d'isoformes de gènes en raccourcissant la taille d'un cadre de lecture, mais aussi en créant de nouveaux exons (BEJERANO *et al.*, 2006; SHEN *et al.*, 2011). Afin de limiter leur expression et leur transposition, les ET sont souvent contrôlés par le génome hôte, par exemple via l'ajout de marques épigénétiques comme des méthylations de l'ADN, ou des marques d'histone qui empêchent leur expression (BARAU *et al.*, 2016; JANSZ, 2019; LISCH, 2009; XIE *et al.*, 2013). Si les ET ainsi contrôlés sont insérés à proximité de gènes, alors ces gènes peuvent aussi subir l'influence de ces modifications épigénétiques (REBOLLO *et al.*, 2012a). Enfin, comme proposé initialement par Barbara McClintock (MCCLINTOCK, 1950) on sait aujourd'hui que les ET sont capables de « transporter » des séquences régulatrices tels que des sites de fixation de facteurs de transcription, qui sont nécessaires à leur propre expression (BOURQUE *et al.*, 2008; DU *et al.*, 2016; JORDAN *et al.*, 2003; PAVLICEV *et al.*, 2015; RAYAN *et al.*, 2016; SUNDARAM *et al.*, 2014; TRIZZINO *et al.*, 2017). Chez la souris, on estime que 40% des sites de fixation du facteur de transcription CTCF sont dérivés d'ET (**Fig. 1.9.**) (SUNDARAM *et al.*, 2014). Pour qu'un site de fixation de facteur de transcription apparaisse *de novo*, plusieurs mutations ponctuelles doivent se produire au bon endroit. Les ET au contraire permettent d'apporter des séquences prêtes à l'emploi qui peuvent être recrutées pour la régulation des gènes. Il est aussi possible que le site transporté par un ET soit incomplet, mais que très peu de mutations soient nécessaires à le rendre fonctionnel ou à l'optimiser (BOURQUE *et al.*, 2008; ELLISON et BACHTROG, 2013; SUNDARAM et WYSOCKA, 2020). C'est par exemple le cas d'un hélitron qui a apporté chez la drosophile des éléments régulateurs de l'expression des gènes du chromosome X chez les mâles (voir section 1.3, page 31). La capacité des ET à véhiculer des séquences régulatrices leur confère un avantage sélectif en améliorant la valeur sélective de l'hôte, ce qui est l'un des facteurs favorisant leur persistance au cours des générations (BOURQUE, 2009; JACQUES *et al.*, 2013; MARNETTO *et al.*, 2018; SUNDARAM et WYSOCKA, 2020).

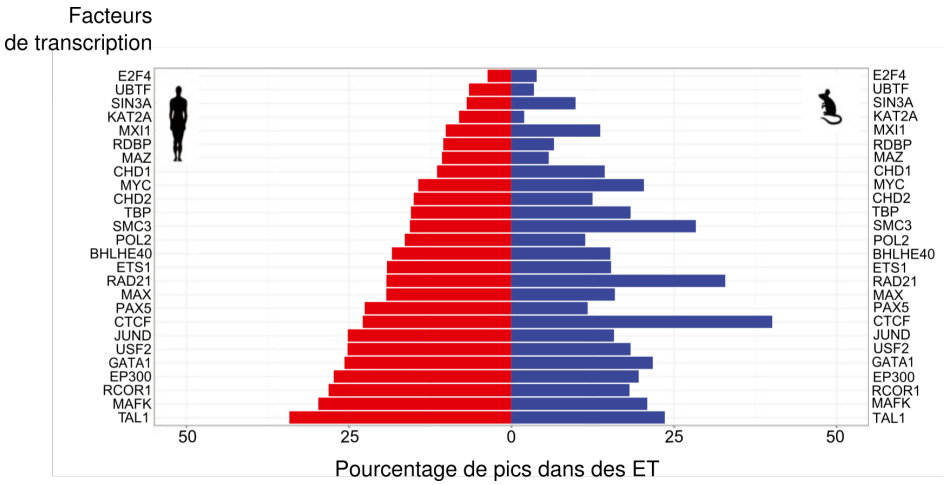

FIGURE 1.9 – **Pourcentage de pics de ChIP-seq de différents facteurs de transcriptions retrouvés dans des ET chez l'humain et la souris.** Le ChIP-seq est une technique de séquençage d'ADN nouvelle génération, qui permet de séquencer les régions du génome avec lesquelles un facteur d'intérêt interagit. Pour certains facteurs de transcription choisis par les auteurs, un nombre élevé de sites de fixation se situe dans des ET. Adapté de Sundaram *et al.* 2014 (SUNDARAM *et al.*, 2014).

1.2.6.3 Exemple d'éléments transposables contrôlant l'expression de gènes

De nombreux exemples d'ET contrôlant l'expression de gènes ont été décrits dans la littérature. L'adaptation de la phalène du bouleau, un papillon nocturne, à son environnement après la révolution industrielle est un des exemples de microévolution les plus connus (**Fig. 1.10.**) (HOF *et al.*, 2016). Suite à l'industrialisation massive du Royaume-Uni, la pollution au charbon a induit la sélection d'individus aux ailes plus foncées. Ce changement de couleur leur permet de mieux se camoufler sur les arbres noircis par les fumées de charbon, et ainsi d'éviter la prédation par les oiseaux. Les phalènes qui vivent dans des environnements moins pollués ont gardé leur couleur plus claire qui leur assure un meilleur camouflage sur les écorces de bouleau. Une forte expression du gène *cortex* qui joue un rôle dans le contrôle du cycle cellulaire pendant le développement de l'aile permet la coloration plus foncée adaptée à l'environnement pollué (HOF *et al.*, 2016). L'insertion d'un ET dans le premier intron du gène, datée au moment de la révolution industrielle, s'est révélée être la cause de l'augmentation d'expression du gène par un mécanisme pour le moment inconnu. On voit ici que les ET peuvent être source de diversité et d'adaptation à un environnement variable.

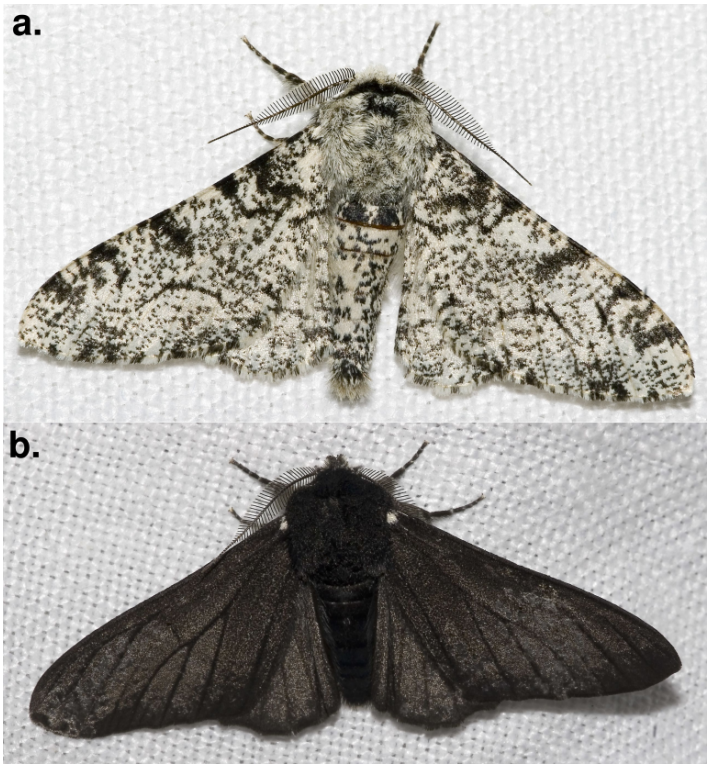

FIGURE 1.10 – **Photographies de deux phalènes du bouleau, *Biston betularia*.** a. Forme *typica*, présente avant la révolution industrielle en Angleterre b. Forme *carbonaria* apparue après la révolution industrielle, qui est mieux camouflée en environnement pollué au charbon. L'adaptation est apparue suite à l'insertion d'un ET dans le gène *cortex* entraînant sa surexpression. ©Chiswick Chap - BY-SA 3.0.

Il existe aussi des exemples connus d'ET qui contrôlent l'expression de gènes chez les vertébrés. Chez les cichlidés, un groupe de poissons téléostéens vivant notamment dans les grands lacs africains, une insertion d'ET a été associée à un comportement sexuel spécifique (SANTOS *et al.*, 2014). Comme la plupart des poissons, les cichlidés utilisent la fécondation externe pour se reproduire. La femelle dépose les œufs dans l'eau, et le mâle les fertilise en déposant le sperme directement dessus. Cependant, ce type de reproduction comporte des risques : les œufs fécondés ne sont pas

protégés et peuvent être, par exemple, mangés par des prédateurs. Certaines espèces de cichlidés contournent ce problème : après avoir déposé les œufs dans l'eau, la mère protège les œufs en les prenant dans sa bouche. Bien que cela réduise la prédation, la probabilité des œufs d'être fécondés est aussi réduite. Par ailleurs, on observe chez ces espèces des taches appelées « egg-spot » pour « tâches ressemblant à des œufs » au niveau de la nageoire anale des mâles (**Fig. 1.11.**). Ces tâches ressemblent à de véritables œufs et attirent le regard des femelles qui essaient de les attraper pour les mettre dans leur bouche. En se rapprochant ainsi de la nageoire anale des mâles, elles se retrouvent aussi très proches de leur orifice génital. C'est à ce moment précis que le mâle va déverser son sperme directement dans la bouche de la femelle afin de féconder les œufs qui s'y trouvent. Ce comportement permet de protéger les œufs des prédateurs tout en assurant une bonne probabilité de fécondation des œufs. La femelle va ainsi porter sa progéniture dans sa bouche pendant plusieurs semaines après la fécondation. *fhl2b*, qui code pour un facteur de transcription à domaine LIM four-and-a-half, est un gène issu de la duplication de *fhl2* spécifique des poissons téléostéens. Il est surexprimé dans les egg-spots (SANTOS *et al.*, 2014), et il est connu pour être exprimé dans les iridophores (SALIS *et al.*, 2019), un type de cellules pigmentaires présent au niveau des egg-spots. Chez les cichlidés à egg-spot, on trouve un ET de type SINE inséré en amont du site d'initiation de la transcription de *fhl2b*, dans une région régulatrice du gène. Cette insertion est absente des espèces qui n'ont pas d'egg-spot. Chez le poisson zèbre, qui ne possède pas d'egg-spot, l'ajout de ce SINE en amont du site d'initiation de la transcription du gène *fhl2b* induit sa surexpression dans la peau et une augmentation du nombre d'iridophores. Les poissons cichlidés avec egg-spots constituent un exemple où un ET contrôle l'expression d'un gène et est lié à un nouveau comportement sexuel.

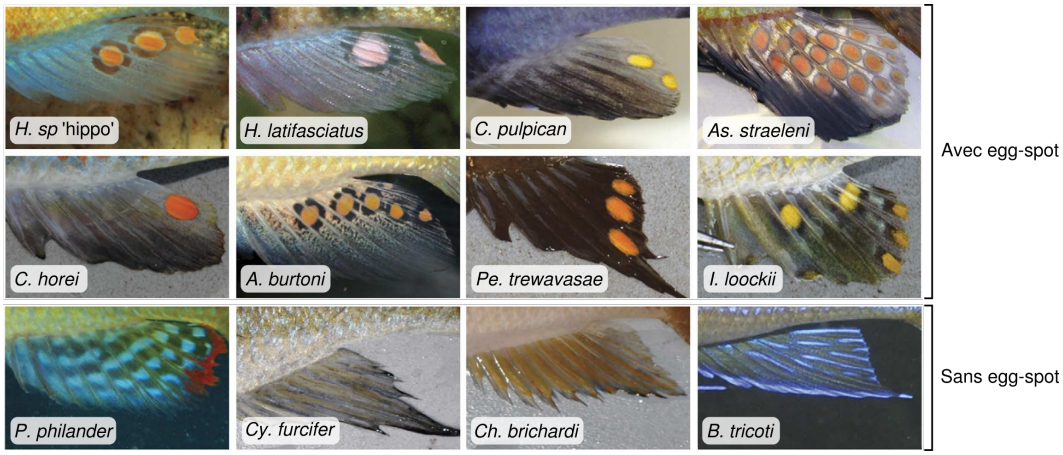

FIGURE 1.11 – **Nageoires anales de poissons cichlidés mâles.** Les taches colorées chez certaines espèces sont appelées egg-spots et ressemblent à des œufs pondus par les femelles. Ils permettent aux mâles d'augmenter leur succès reproducteur en attirant les femelles. D'en haut à gauche à en bas à droite : *Haplochromis sp.* 'hippo', *Haplochromis latifasciatus*, *Cynotilapia pulpican*, *Astatoreochromis straeleni*, *Ctenochromis horei*, *Astatotilapia burtoni*, *Petrochromis trewavasae*, *Interochromis loockii*, *Pseudocrenilabrus philander*, *Cyathopharynx furcifer*, *Chalinochromis brichardi*, *Benthochromis tricoti*. Adapté de Santos *et al.* 2014 (SANTOS *et al.*, 2014).

1.2.6.4 Les éléments transposables sont impliqués dans l'évolution des réseaux de régulation de gènes

Les réseaux de régulation de gènes correspondent à l'ensemble des relations de régulations, qui sont opérées par des facteurs codés par certains gènes et qui agissent sur d'autres gènes. Ces

réseaux peuvent prendre la forme de cascades comme la cascade de différenciation sexuelle décrite précédemment, où chaque gène est activé par le produit du gène précédent. Des boucles de rétrocontrôle négatives ou positives peuvent également exister. Comme proposé par Britten et Davidson (BRITTEN et DAVIDSON, 1969; DAVIDSON et BRITTEN, 1979), les innovations évolutives pourraient reposer essentiellement sur le « recâblage » des réseaux de régulation de gènes déjà existants. Les gènes évolueraient moins vite que les réseaux gouvernant leur activation ou répression. La dispersion des ET au travers des génomes pourrait permettre la dispersion de séquences régulatrices qui modifierait significativement les réseaux de régulation de gènes. Ce modèle a depuis été revisité, et plusieurs exemples d'éléments transposables qui ont participé à l'évolution de réseaux de régulation de gènes sont désormais connus (SUNDARAM et WANG, 2017). Chez l'humain par exemple, des éléments de type ERV ont participé à la dispersion des sites de fixation de la protéine p53, impliquée dans les processus d'apoptose (WANG *et al.*, 2007). Un site de fixation pour p53 est retrouvé dans leur région LTR. Chez les mammifères, un transposon à ADN de type MER20 a été à l'origine d'un nouveau réseau de régulation de gène ayant contribué à l'évolution de la grossesse (LYNCH *et al.*, 2011, 2015). Toujours dans ce groupe, l'expansion de certains rétrotransposons a donné naissance a des sites de fixation pour CTCF très conservés, et dispersés dans le génome (SCHMIDT *et al.*, 2012).

1.2.6.5 Les éléments transposables à l'origine de la formation de nouveaux gènes

En plus de participer au contrôle des réseaux de régulation de gènes comme présenté précédemment, les ET constituent aussi un matériau important pour la formation de nouveaux gènes (BRANDT *et al.*, 2005; CHALOPIN *et al.*, 2012; VOLFF, 2006; VOLFF *et al.*, 2001; WARREN *et al.*, 2015). En perdant leur capacité de transposition, ils peuvent être fixés à certaines positions et acquérir des mutations leur conférant une fonction pour l'hôte. De nombreux exemples ont été décrits chez les vertébrés (ETCHEGARAY *et al.*, 2021). Parmi les exemples les plus connus on trouve les syncytines, des protéines impliquées dans la fusion de cellules pour former le syncytiotrophoblaste, au sein du placenta, chez les mammifères placentaires mais aussi chez d'autres vertébrés placentaires comme les lézards du genre *Mabuya* (CORNELIS *et al.*, 2017; DUPRESSOIR *et al.*, 2005, 2011; LAVIALLE *et al.*, 2013). Ces protéines sont dérivées de protéines d'enveloppe (Env) des rétrovirus endogènes, protéines qui permettent aux virus de fusionner leur membrane avec celle de leurs cellules cibles. Cet événement de domestication moléculaire a eu lieu plusieurs fois au cours de l'évolution des mammifères, ou chez les lézards : c'est un exemple d'évolution convergente. Un autre exemple connu d'ET domestiqué touche le système immunitaire. Chez les vertébrés à mâchoire, le système immunitaire adaptatif est basé sur la recombinaison V(D)J qui a lieu dans les lymphocytes en cours de développement. Cette recombinaison somatique permet de générer une grande diversité de récepteurs antigéniques (dans les lymphocytes T) et d'anticorps (dans les lymphocytes B). Les protéines RAG1 et RAG2 catalysent cette recombinaison. Récemment, il a été montré que l'origine de ces protéines pourrait résider dans la domestication d'un transposon RAG ancestral, apparenté aux transposons à ADN *transib*, il y a 500 à 600 millions d'années (CARMONA et SCHATZ, 2017; KAPITONOV et JURKA, 2005). Enfin, dans les neurones des mammifères, la protéine ARC est indispensable pour la mémoire à long terme, en formant des capsides permettant de transporter des ARN messagers d'un neurone à l'autre. La protéine ARC serait dérivée de la région Gag domestiquée

d'un rétrotransposon (PASTUZYN *et al.*, 2018). Les ET peuvent aussi donner naissance à des ARN non-codants (KAPUSTA *et al.*, 2013). Par exemple, le gène *Xist* des mammifères, impliqué dans l'inactivation transcriptionelle du chromosome X, est dérivé du gène codant *lnx3* (ELISAPHENKO *et al.*, 2008; ETCHEGARAY *et al.*, 2021). Concomitamment à la perte du cadre de lecture, l'insertion de plusieurs ET a apporté de nouveaux domaines et un gain de fonction, formant un long ARN non-codant (*Xist*).

1.2.7 Les éléments transposables chez les poissons téléostéens

Les poissons téléostéens constituent un excellent modèle pour s'intéresser à l'évolution des ET. Leurs génomes présentent en effet une forte diversité en ET (VOLFF, 2005; VOLFF *et al.*, 2003) : on observe chez les poissons téléostéens un grand nombre de superfamilles d'ET différentes par rapport aux mammifères (**Fig. 1.12.**) (CARDUCCI *et al.*, 2020; CHALOPIN *et al.*, 2015), les tétrapodes ayant perdu de nombreuses familles d'ET au cours de leur évolution. Cette diversité d'ET, comme vu dans les parties précédentes, représente un réservoir de matériel pour la formation de nouveaux gènes par la domestication, ou pour la formation de nouvelles séquences régulatrices. En revanche, la couverture du génome par les ET n'est pas supérieure chez les poissons à celle observée chez d'autres espèces de vertébrés. Elle varie de moins de 10% chez des espèces comme le fugu à plus de 50% pour le poisson zèbre (**Fig. 1.13.**). Chez les mammifères, les ET prédominants ont tendance à être des rétrotransposons (classe I), quand chez les poissons téléostéens, la composante dominante est plutôt de classe II.

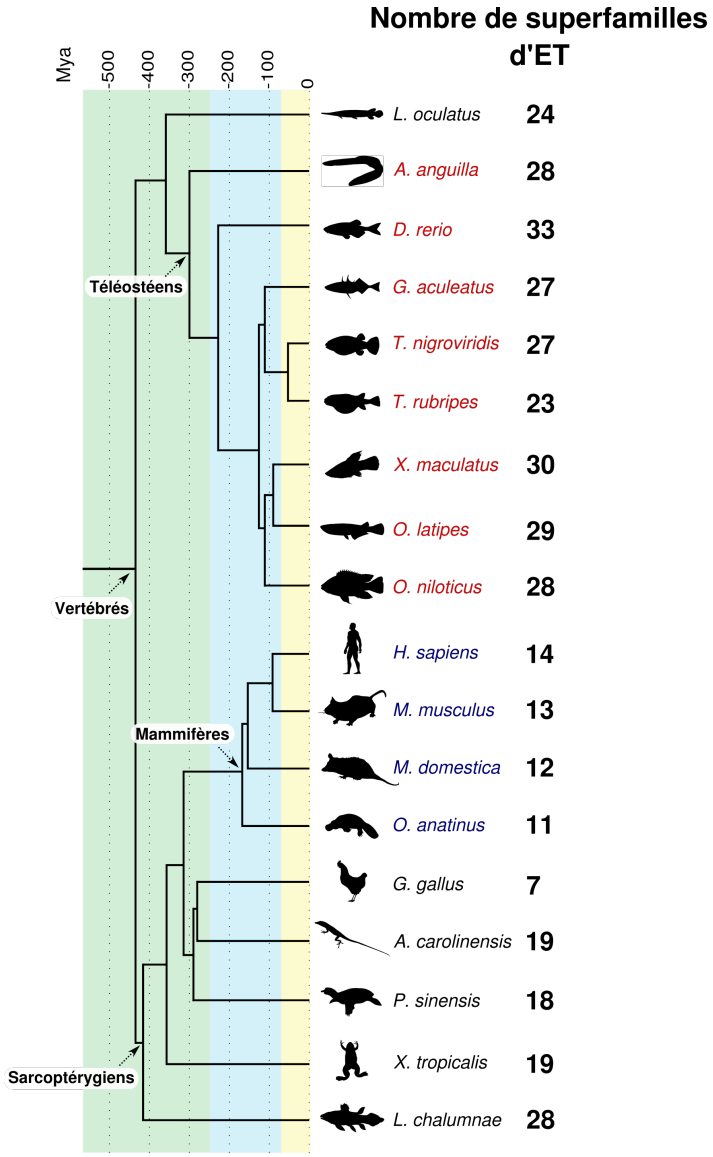

FIGURE 1.12 – **Phylogénie simplifiée des vertébrés montrant le nombre de superfamilles d'ET détectées dans chaque espèce.** Adapté de Chalopin *et al.* 2015 (CHALOPIN *et al.*, 2015). On observe chez les poissons téléostéens (en rouge) une diversité en ET plus élevée que chez d'autres vertébrés comme les 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/.

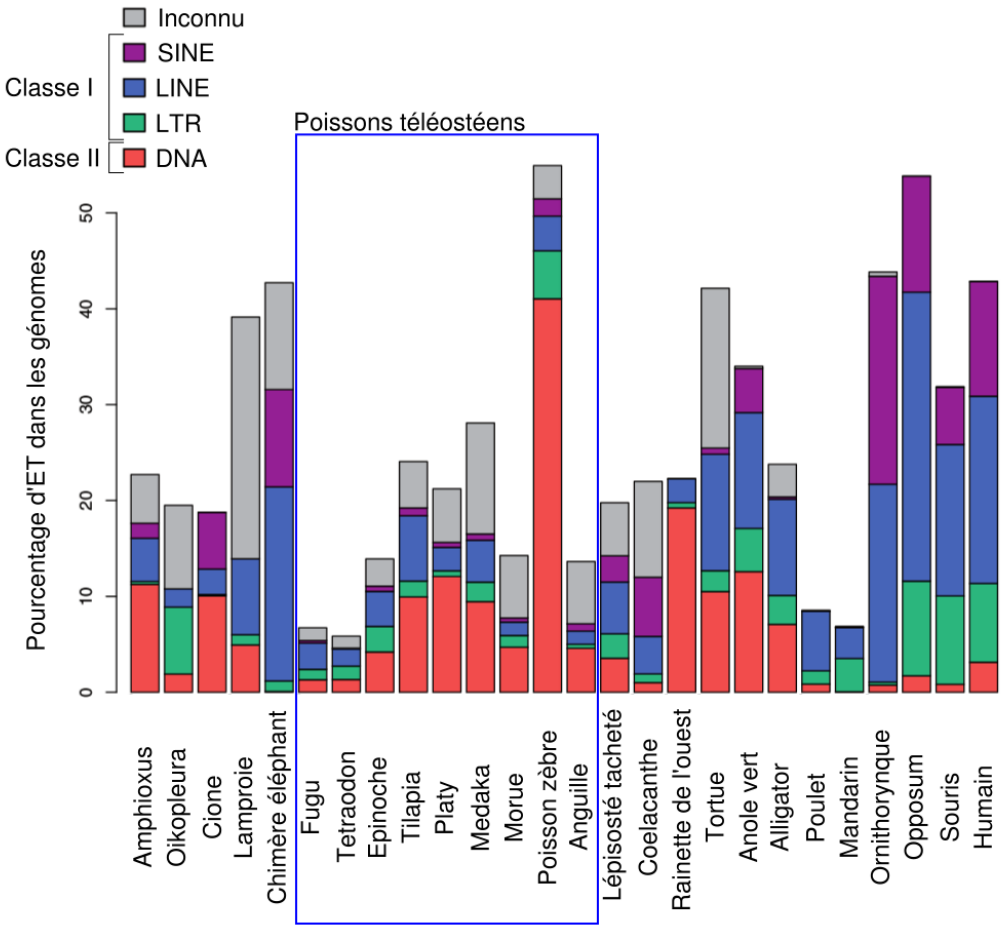

FIGURE 1.13 – **Pourcentage d'ET dans différentes espèces de vertébrés.** Les poissons téléostéens (encadrés en bleu) possèdent majoritairement des ET de classe 2 (en rouge). A droite, les mammifères ont un génome contenant majoritairement des ET de classe I : SINE, LINE et LTR. Adapté de Chalopin *et al.* 2015 (CHALOPIN *et al.*, 2015).

La diversité du répertoire d'ET chez les téléostéens fait donc de ce groupe d'espèces un modèle de choix pour l'étude de l'impact des ET sur l'évolution des génomes et en particulier sur l'évolution des réseaux de régulation des gènes.

1.3 Le destin lié des éléments transposables et du sexe

Il existe de nombreuses interactions entre ET et sexe : les ET ont influé sur l'évolution du sexe comme le sexe a influencé l'évolution des ET. Pour mieux comprendre les enjeux qui existent dans l'étude de cette interaction, une revue de la littérature a été rédigée dans le cadre de cette thèse et publiée dans le journal Mobile DNA (DECHAUD *et al.*, 2019). Elle traite du contrôle de l'expression des ET dans les gonades, du contrôle de l'expression des gènes du sexe par les ET, et enfin du rôle des ET dans l'évolution des chromosomes sexuels.

Dechaud et al. Mobile DNA (2019) 10:42
https://doi.org/10.1186/s13100-019-0185-0

Mobile DNA

REVIEW Open Access

Sex and the TEs: transposable elements in sexual development and function in animals

Corentin Dechaud[1], Jean-Nicolas Volff[1], Manfred Schartl[2,3*] and Magali Naville[1*]

Abstract

Transposable elements are endogenous DNA sequences able to integrate into and multiply within genomes. They constitute a major source of genetic innovations, as they can not only rearrange genomes but also spread ready-to-use regulatory sequences able to modify host gene expression, and even can give birth to new host genes. As their evolutionary success depends on their vertical transmission, transposable elements are intrinsically linked to reproduction. In organisms with sexual reproduction, this implies that transposable elements have to manifest their transpositional activity in germ cells or their progenitors. The control of sexual development and function can be very versatile, and several studies have demonstrated the implication of transposable elements in the evolution of sex. In this review, we report the functional and evolutionary relationships between transposable elements and sexual reproduction in animals. In particular, we highlight how transposable elements can influence expression of sexual development genes, and how, reciprocally, they are tightly controlled in gonads. We also review how transposable elements contribute to the organization, expression and evolution of sexual development genes and sex chromosomes. This underscores the intricate co-evolution between host functions and transposable elements, which regularly shift from a parasitic to a domesticated status useful to the host.

Keywords: Transposable element, Sex determination, Sexual development and function, Germline, piRNA, Sex chromosome

Background

Transposable elements (TEs) are major actors of the evolution of genomes and the diversification of species [1]. These DNA sequences have the peculiarity of being able to integrate into and spread within genomes, as well as to recombine and induce genome rearrangements, since they are generally repetitive. First discovered in maize [2], TE families described so far are generally divided into two main classes [3]. Class I TEs (retroelements) spread through a "copy-and-paste" mechanism called retrotransposition, which corresponds to a process of RNA-mediated duplication. They express an RNA intermediate that is reverse-transcribed into a cDNA fragment, which will be inserted somewhere else into the genome. Hence, retrotransposition directly increases the copy number of an element. In contrast, Class II TEs (DNA transposons) move through a "cut-and-paste" mechanism. Most autonomous class II elements encode a transposase that can bind to and excise the transposon from its initial genomic localization, and can subsequently insert it into a new locus [3–5]. This mechanism does not per se duplicate the initial transposon but only changes its location. However, the transposon can be duplicated if the transposition event occurs during the replication process, from an already replicated region to a non-replicated one.

Since they can insert into genomes, recombine and generate different types of rearrangements, TEs are by nature an important source of genomic variability between different species, or between individuals within a given species or population. Most insertions are thought

* Correspondence: phch1@biozentrum.uni-wuerzburg.de; magali.naville@ens-lyon.fr
[2]Entwicklungsbiochemie, Biozentrum, Universität Würzburg, Würzburg, Germany
[1]Institut de Genomique Fonctionnelle de Lyon, Univ Lyon, CNRS UMR 5242, Ecole Normale Superieure de Lyon, Universite Claude Bernard Lyon 1, 46 allee d'Italie, F-69364 Lyon, France
Full list of author information is available at the end of the article

to be deleterious for the host, in particular when they disrupt essential genes, regulatory regions or chromosomal structures, causing negative effects ranging from a slight decrease in host fitness to lethal mutations [6]. When a TE insertion is associated with such a fitness disadvantage, it is generally counter-selected and finally lost. The process of loss can however be modulated by several factors, including the selection coefficient of the insertion, its potential linkage disequilibrium with an advantageous allele, the recombination rate of the region of insertion, and the effective population size of the host [7]. Some insertions, in contrast, can be neutral, for example if they occur in genomic regions that have no crucial impact on host fitness, like gene-poor regions for instance. It is however difficult to classify an insertion as "neutral" once and for all, since it can still induce chromosomal rearrangements through ectopic recombination [8]. Lastly, some TE insertions might bring positively selected changes. In particular, TEs can spread ready-to-use regulatory sequences or trigger epigenetic modifications able to modify the pattern of expression of neighboring genes (for a review see [9]). TEs can also be "domesticated" as new host non-coding RNA genes or genes encoding useful proteins such as the syncytins, which are involved in the development of the placenta in mammals [10–12]. Syncytin genes have been repeatedly derived from *envelope* genes of endogenous retroviruses during mammalian evolution. Another example of TE-derived host proteins are the Rag proteins, which catalyze the V(D) J recombination responsible for the diversity of immunoglobulins and T cell receptors found in B and T cells, respectively. These proteins were formed from a Transib DNA transposon about 500 million years ago [13]. Many other examples of TE-derived genes have been described in different organisms (for a review see [11, 14]).

Persistence of TEs within a population, which would reflect their evolutionary success, requires their vertical transmission from one generation to the next. In animals with sexual reproduction, i.e. involving the fusion of male and female gametes, this implies transposition in the germline cells that will form the next generation. Sexual reproduction might be instrumental for the propagation of mainly deleterious TEs [15–17]. Indeed, in asexual populations, TEs might not be able to spread and tend to be eliminated if no horizontal transfer occurs [15–17]. Accordingly, experimental studies have shown that TEs are less fit to increase their frequency in asexual populations compared to sexual populations [15, 17–19]. Homologous recombination during meiosis is another feature of sexual reproduction that has an antagonistic impact on the fixation rate of TEs by favoring the elimination of deleterious TE insertions [20, 21]. Recombination triggers the exchange of genetic information between homologous chromosomes belonging to a same chromosome pair. This process has been associated to an increase of purifying selection since it drives the removal of deleterious point mutations and TE insertions [20, 21]. Hence, recombination and sexual reproduction could be considered as a defense mechanism against deleterious TE insertions. Reciprocally, high rates of deleterious mutations such as TE transpositions might favor the maintenance of sexual reproduction as an efficient way to keep these mutations at levels compatible with life [15, 17, 22–24]. In the asexual species *Leptopilina clavipes* (the wasp), no particularly high TE content is observed, despite the expansion of specific TE families, which could be linked to the switch toward asexuality [25]. The absence of recombination here does not seem to have triggered a massive expansion of TEs, or is counterbalanced by the limited spreading of TEs in the population due to asexuality. Similarly, no difference in TE composition was observed between the genome of an asexual fish of hybrid origin, the amazon molly *Poecilia formosa*, and the genomes of its parental sexual species, possibly due to the very recent occurrence of the switch from sexuality to asexuality in this lineage [26]. In the more ancient asexual taxa of the bdelloid rotifers, retrotransposons were long thought to be absent [27], supporting the role of sexuality in the genomic maintenance of these TEs [23]. More recent studies somehow challenged this model by highlighting a high diversity of TE families including LTR and non-LTR retrotransposons. However, each of these families presents a very low number of intact copies (one or two for the majority of them) [28]. Such a TE landscape, associated with the relatively low abundance of decayed fragments, the high similarity between LTRs for intact copies, and the localization of TEs in horizontally transmitted regions, led the authors to hypothesize that TEs were mostly acquired by recent horizontal transfers in rotifers [28].

In species with gonochoristic sex, i.e. species in which individuals are either male or female (in contrast to hermaphrodite species, in which individuals produce both male and female gametes), different factors can control sex determination (SD) [29, 30]. Some species undergo environmental sex determination (ESD), while others are subject to genetic sex determination (GSD). In ESD sex is determined by environmental factors, for instance temperature in turtles or crocodilians [31, 32]. Such temperature sex determination seems to be also present, albeit rare, in fish, as it was recently demonstrated for the Southern flounder [33]. In GSD on the contrary, the sex of the individual depends on its genotype. Sex can be determined by several interacting loci in a given species (polygenic sex determinism), but the most prevalent situation appears to be the monogenic GSD. In this situation, the chromosome pair that harbors the master SD gene becomes the sex chromosomes,

or gonosomes. Two main sex chromosome configurations exist: the XX/XY system, particularly found in mammals, where males have two types of sex chromosomes (X and Y, male heterogamety), and the ZW/ZZ system, common in birds, where females have two different sex chromosomes (Z and W, female heterogamety) [34, 35]. Many other GSD systems have been reported such as haplodiploidy, where for instance males arise from haploid unfertilized eggs and female from diploid fertilized eggs, like in bees, ants, or some molluscs [36]. In the XX/XY sex determination system in mammals, the *Sry* gene is the male master sex determining gene for almost all species. *Sry* is located on the Y but not on the X chromosome and is therefore present in males but not in females. Non-mammalian species such as the fruit fly *Drosophila melanogaster* or the medaka fish *Oryzias latipes* also have XX/XY sex determination systems but of independent evolutionary origins. The *Sry* gene is absent from these species. In *O. latipes* the Y-linked master gene *dmrt1bY*, which is a Y-specific duplicate of the *dmrt1* gene, drives development toward the male phenotype like *Sry* in mammals [37, 38]. In *D. melanogaster*, the X chromosome carries *Sxl* that has to be in two copies to trigger female differentiation [39]. In this case, the initial choice between the male and female pathways is thus triggered by a dosage effect of the master gene. In birds, a similar process occurs but in a ZW/ZZ system, where ZZ males have two copies of the Z-linked *dmrt1* gene and females only one. This creates a gene dosage difference, leading to male or female differentiation [40]. In the nematode *C. elegans* individuals are either males or hermaphrodites. The presence of two X chromosomes (XX individuals) triggers the differentiation into a hermaphrodite adult that produces both male and female gametes. In contrast, XO individuals differentiate into males as a consequence of the ratio between X chromosomes and autosomes [41, 42].

Once sexual development is initiated, the gonad, which comprises both germ cells and somatic cells, differentiates into either a testis or an ovary. A sex-dependent gene regulatory cascade, initiated in the somatic part of the gonad, controls differentiation [30, 43, 44]. Male and female differentiation cascades are often repressing each other, creating a competition between male and female differentiation genes: the most expressed pathway represses the other one [43]. Finally, once the gonad is differentiated, sex is maintained by the expression of specific genes like those encoding the sexual hormone biosynthesis pathways in mammals. It has been shown in mammals and teleost fish that even in adults, derepressing the opposite pathway can induce sex reversal [45–47]. This demonstrates that expression of at least some of the sexual development network genes is necessary to maintain the differentiated state in sexually mature individuals. Beyond gonads, sex affects many other pathways in the organism, creating a bias in gene expression in several tissues and organs including brain [48–53]. However, gonads remain the most sex-biased organs in terms of gene expression.

Depending on the animal lineage, sexual development and particularly sex determination can show very different evolutionary dynamics. Some SD systems are ancient and at least 100 million years old, such as the mammalian male heterogamety system driven by the Y-linked gene *Sry* [54] or the avian female heterogametic determination controlled by the Z-linked *dmrt1* gene [40]. In other lineages, for instance in teleost fish, sex determination is much more labile, with a frequent switch between and even combination of ESD and GSD, and an important turn-over of sex chromosomes and master sex-determining genes in GSD [55, 56]. For example, the genetic sex determination system is not conserved in the genus *Oryzias*: while *O. latipes*, *O. curvinotus*, *O. luzonensis* and *O. dancena* use a XX/XY system, *O. javanicus* determines sex through ZW/ZZ female heterogamety [57]. Strikingly, Oryzias species with a XX/XY system generally have different sex chromosomes and even different master sex-determining genes: sex is controlled by *dmrt1bY* (aka *dmy*) in *O. latipes* and *O. curvinotus, gsdfY* in *O. luzonensis* and *sox3Y* in *O. dancena* [57]. Hence, the control of sexual development can be considered as a fast-evolving trait in this clade. Beyond the initiation of sex differentiation, the downstream molecular pathways also appear variable among animals: a comparison of genes expressed in medaka fish and mammalian gonads revealed substantial differences [58]. Very interestingly, the control of sexual development sometimes experiences convergent evolution: in both therian mammals (non-egg-laying placental mammals and marsupials) and *Oryzias dancena* for instance, the master sex-determining gene evolved from the *Sox3* gene [59]. This happened independently in the two lineages, 148 to 166 million years ago in a common ancestor of therian mammals, and less than 20 million years ago in *Oryzias dancena*. Another striking example is the *dmrt1* gene in birds and in the tongue sole. This gene was ancestrally located on the vertebrate linkage group A, which became the Z chromosome independently in both lineages [60].

In this review, we reassess the impact of transposable elements on the structure and expression of genes and genomes through the prism of sex by inventorying the known reciprocal interactions between TEs and sexual development and function in animals. The species sample, however, appears heavily biased towards insects and vertebrates, since most of the studies linking TE and sex have been conducted in classical model organisms commonly used in genetics and development. We first focus on the expression of TEs in germ cells and on the

control of their expression. Then, we review how TEs, reciprocally, can impact the expression of sexual development genes. Finally, we document how TEs influence the organization and structural evolution of sexual genes and chromosomes. These diverse and reciprocal influences well illustrate the intricate co-evolution of TEs with their host.

TE expression is tightly controlled in the germline

TEs in the germline: a trade-off between expression and control

Expression and transposition of TEs in the germline are necessary for their vertical transmission to the host progeny, and ultimately for their maintenance within a lineage. The first step of TE transposition consists in the transcription of mRNA to produce enzymes such as a transposase for most DNA transposons, or a reverse transcriptase and an integrase/endonuclease for retroelements. TE mRNAs are expected to be found in cells where TEs are spreading. TE-derived transcripts are indeed found in transcriptomes [61–64], including the germline [65, 66]. In the medaka *Oryzias latipes* for instance, about 1.2 and 3.5% of the transcriptome of ovaries and testes, respectively, can be assigned to TEs (Dechaud et al. unpublished data).

If evolution fosters TEs that are active in gonads, the putative negative effects of TE insertions, at the same time, require repressive mechanisms. The gonadal activity of a TE results in a trade-off, its own survival depending on the survival of the host, which is needed for vertical transmission and maintenance. This follows the "selfish gene" hypothesis according to which, in a gene-centered view of evolution, some genes can enhance their own transmission, sometimes with a negative effect on the organism fitness [16]. Very interestingly, some TEs like the P element in Drosophila produce different transcripts depending on the organ in which they are expressed [67]. In the gonads, the third intron of the P element is excised allowing its transposition, while in the soma, in addition to a transcriptional control, the P element transcript keeps its third intron and is not able to transpose [67]. Such mechanisms allow the element to limit its impacts on the soma while transposing in the germline.

Germline TE expression is controlled by several mechanisms

piRNAs (Fig. 1a)

Piwi-interacting RNAs (piRNAs) are 24–31 nucleotides long small non-coding RNAs expressed in the germline and derived from long RNAs that contain TE sequences [68]. They have been described in eukaryotes only, from humans to protozoans [69, 70] and play a large diversity of roles, such as genome rearrangement in ciliates, sex determination in silkworm, telomere protection in Drosophila, long-term memory in sea slug, or oocyte development in human [70]. piRNAs are produced from specific loci called piRNA clusters that regularly integrate new TE-derived sequences and thus extend their target potentialities. They can further be amplified by the so-called "ping-pong" cycle [71].

piRNAs can regulate TE expression by two different mechanisms. The first mechanism occurs in the nucleus, where piRNAs interact with the Piwi proteins, a subfamily of Argonaute nucleases, to target the TE nascent RNAs to which they present sequence similarities, and adds histone repressive marks in the region by interacting with other proteins [68]. This mechanism inhibits the expression of the targeted TEs. The second mechanism happens in the cytoplasm, where piRNAs form a complex with Aubergine (Aub) proteins, which belong to the Piwi subfamily too. This complex post-transcriptionally silences TE expression by interacting with the TE mRNAs. This also triggers a replication of the piRNA, known as the ping-pong cycle [68]. The ubiquitous presence of this regulatory system in the gonads specifically underlines the importance of controlling TE activity in the germline.

As an example, piRNAs are involved in the P-cytotype regulation in *Drosophila* [72]. In these species, some strains of flies have a DNA transposon, the P element, from which a complementary piRNA is produced. These are called "P strains", for Paternal contributing strains, in opposition to "M strains", for Maternal contributing strains. One model proposes that in P strains, P element-derived piRNAs are transmitted from the mother through the oocyte cytoplasm. The transmitted piRNAs then silence the P element both in the nucleus and the cytoplasm by the mechanisms described above. piRNAs are further amplified in the cytoplasm through the ping-pong cycle, maintaining the silencing of the P element. If no piRNA is transmitted from the mother, the P element is not repressed. Consequently, a P male crossed with an M female will have a dysgenic offspring, with increased mutation rates, frequent sterility and abnormally small gonads [73]. This phenomenon, due to the fact that the offspring have the P element but no silencing through maternal piRNA, is known as "hybrid dysgenesis" [67, 72]. In contrast, the offspring of a P female crossed with an M male is fertile, as the P female brings the P element but also some piRNAs to trigger its repression, as well as the ping-pong amplification cycle.

Repressor proteins (Fig. 1b)

TE expression can also be directly controlled by protein factors. In vertebrates, KRAB-ZNF (for Krüppel-associated box domain zing finger) proteins have been shown to play this role ([74], reviewed in [75]). They constitute

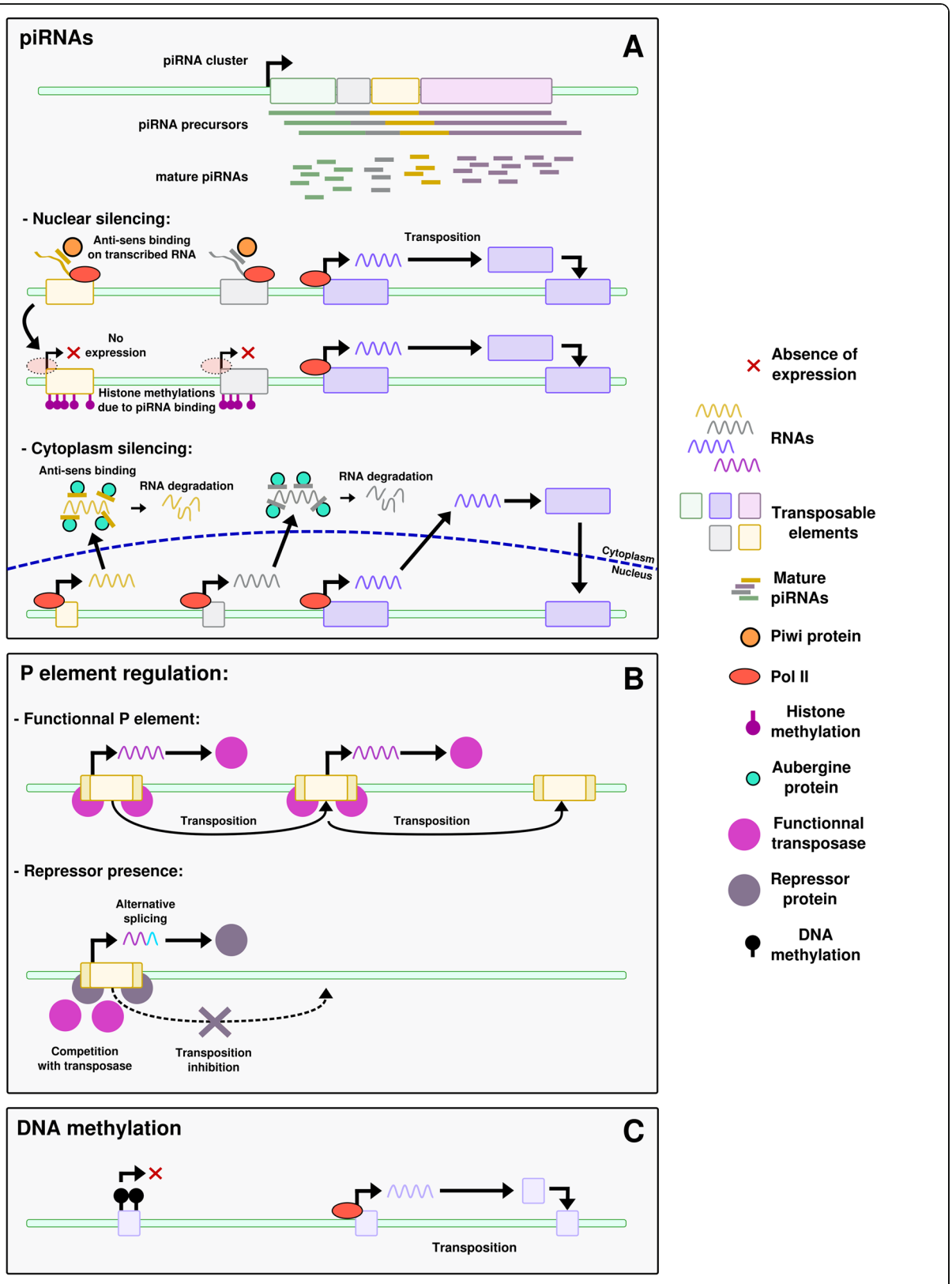

Fig. 1 (See legend on next page.)

(See figure on previous page.)
Fig. 1 Different ways to control TE expression. **a** piRNAs. piRNAs are produced from piRNA clusters, genomic spots where new TEs can integrate. piRNAs can act through two mechanisms. In the nucleus, piRNAs bind to Piwi proteins. They also bind in anti-sense to TE mRNA being transcribed, triggering histone methylation of TEs and thus inhibiting recruitment of Pol II. This leads to the silencing of TE expression. In the cytoplasm, piRNAs bind to other Argonaute proteins, triggering TE mRNA degradation. **b** Repressor proteins. A functional P element produces the transposase that triggers its excision and transposition. When repressor proteins are transmitted from the mother through cytoplasm or when the P element is degenerated, it produces an alternatively spliced mRNA. This mRNA encodes a non-functional transposase that will act as a repressor by competing with the functional transposase, and trigger the production of more alternatively spliced mRNA. This positive repression loop, where the repressor protein activates its own production, prevents the transposition of the TE. **c** DNA methylation. The TE is methylated, preventing its expression

a large family of proteins and are able to bind various DNA sequences via the diversity of their ZNF domains. They recruit KAP1 (for KRAB-associated protein 1) to DNA, which in turn mediates transcriptional silencing through histone modifications. KRAB-ZNF proteins were first discovered in mice where they silence genomic insertions of a murine leukemia virus (MLV) [76], but recent studies demonstrated their action on other retro-elements [77]. Many KRAB-ZNF proteins are expressed during germline development; however the targeted TE families are still to be discovered for most of the KRAB-ZNF members [77–79]. In Drosophila, a second model of P-element control involves repressor proteins. P strains express a repressor protein that prevents the transposition of the P element in the germline. This mechanism is known as the "protein repressor model" [67, 72]. The repressor is produced from degenerated P elements or from alternatively spliced full P element transcripts. If the precise action mechanism of the repressor protein is unknown, the main hypothesis is a competitive inhibition with the P element transcription [72]. This repressor could also further trigger the production of alternatively spliced transcripts, leading to a feedforward repression loop (Fig. . 1); however this action as a splicing modifier has never been demonstrated. It is inherited from the mother through the cytoplasm. Since the discovery of piRNA however, later demonstrated to repress TEs in the germline [80], an alternative model has been proposed for the P-cytotype regulation (see before). Both models are not mutually exclusive and likely coexist within populations or individuals [72].

Epigenetic modifications (Fig. 1c)

TE activity can be controlled by epigenetic regulations such as DNA methylation [9] or histone modifications [80, 81]. These epigenetic controls however are not specific of the germline. The modifications targeting TEs can sometimes also affect neighboring genes, hence participating in shaping their regulation and influencing genome evolution [82]. Indeed, the epigenetic silencing of TEs is known to be released in cases of stress, for example UV exposure or temperature changes [83]. Thus TEs can be reactivated and expand, influencing genome evolution under stress conditions [82].

TE expression can vary between sexes

Epigenetic modifications and gene expression can differ between sexes. One may wonder, because of these epigenetic differences, whether TE activity would also vary between males and females. Some TE families are expressed at unchanged levels in very different contexts, like SINEs in rats [84]. In this study, 11 organs were tested including testis and uterus, each at 4 developmental stages. Contrary to SINEs, LTR appeared to be more likely to be expressed in specific tissues or conditions, and are also found more differentially expressed between sexes [84, 85].

In mammals, the inactivation of the Piwi regulatory system in the germline of males leads to azoospermia (no production of mature gametes) due to a high rate of illegitimate pairing between non-homologous chromosomes at meiosis that trigger apoptosis [86]. Also, piRNA interacting protein expression was found to be impaired in humans with cryptorchidism (absence of both testes, or location outside the scrotum) [87]. In contrast, Piwi system inactivation in female mice does not lead to over-activation of TEs [86], and neither does a knock-out of *dicer*, a protein involved in the siRNA degradation system, which would have suggested the involvement of the RNA interference pathway in TE control. One player of this control corresponds instead to the evolutionarily conserved MAEL protein (encoded by the *maelstrom* gene), found both in mouse and fly [88]. When this factor is mutated, a 2.3-fold excess of L1 mRNA is measured in embryonic day 15.5 mouse oocytes [88]. Although its precise role is still unclear, MAEL intervenes in a silencing step downstream of Piwi [64]. Of note, TEs are hypomethylated in females compared to the male germline. Hence, oocytes seem more resilient to TE transposition than the male germline. It has been suggested that this difference could be linked to the life-long division of spermatogonial cells, in contrast to oocytes, which undergo a long meiotic arrest. Cell division is required for TE transposition, and many more cell divisions occur in the male germline. More cell divisions would allow too many deleterious insertions in the male germline, explaining the need for TE silencing [86].

TEs can regulate the expression of sexual development genes

TEs can have an important impact on gene regulatory networks [89–91]. They can modify the expression of surrounding genes [9, 91] by bringing with them Pol II or III promoters as well as transcription factor binding sites, insulators, splicing sites or epigenetic modifications. TEs could be particularly prone to recruitment into sexual development since they are generally expressed in the gonads.

Regulation in cis (Fig. 2a)

TEs have a strong cis-regulatory potential for host genes through their Pol II or Pol III promoters and binding sites for transcription factors, or other regulatory sequences, which they carry [9]. These regulatory sequences can already exist in the TE sequence, or derive from this sequence by a few point mutations only. Some of the described examples are related to sexual development.

In Drosophila species, MSL Recognition Elements (MREs) are known to trigger dosage compensation for X chromosomal genes. MSL (for Male Specific Lethal) is a male-specific complex that binds to MREs and increases neighboring gene expression in XY males, hence compensating for the absence of one X chromosome compared to XX females. MREs are found at multiple loci interspersed on the X chromosome. Interestingly, they are carried by Helitron DNA transposons that regulate in *cis* genes close to their insertion sites [92, 93]. In *Drosophila miranda* the X chromosome is recent, allowing the detection of the Helitron sequences with alignments methods, while in other *Drosophila* with older X chromosomes, MREs are present but the Helitrons are not detectable anymore. The authors propose that, on these older chromosomes, selection eroded the Helitron TEs outside of the selected MRE motifs [92, 93]. This example illustrates the efficiency of TEs in the rewiring of gene regulatory networks, as they can spread transcription factor binding sites or other types of regulatory sequences that can then co-regulate several genes. This process appears even more efficient than the birth of transcription factor binding sites "from scratch" by a series of point mutations, which would require much more time to target different genes [89]. More recent studies on MSL in *Drosophila* show that other mechanisms such as microsatellites expansion also spread MRE motifs on neo-X chromosomes [94]. In *Drosophila melanogaster*, the promoter of the *Su (Ste)* piRNA – one of the most abundant piRNA in the testes – derives from a *1360* transposon [95, 96]. *Su (Ste)* silences the *Stellate* genes, hindering the accumulation of *Stellate* proteins, which causes formation of crystals and results in male sterility [97].

Other cases of TE-controlled genes have been described in other organisms. In the medaka fish *Oryzias latipes,* the master sex determining gene *dmrt1bY* has been formed through the duplication of the autosomal gene *dmrt1a,* which has a downstream position in the male sex differentiation cascade in vertebrates. *Dmrt1bY* is controlled by different transcription factors including itself, its paralog Dmrt1a and Sox5. Binding sites for these transcription factors are located in the upstream region of *dmrt1bY*, which corresponds to a non-autonomous P element called *Izanagi*, in which a LINE/Rex1 retroelement was inserted later (Fig. 3a) [98]. The binding sites for Dmrt1A and Dmrt1bY are located within *Izanagi,* while the binding site for Sox5 lies within the Rex1-derived sequence [47, 98]. Here, the TEs directly brought the *cis*-regulatory elements that conferred to *dmrt1bY* an expression pattern compatible with a function as a master sex-determining gene. This makes a convincing case for TEs being actors of sex determination evolution (Fig. 3b) [98]. Accordingly, it has also been suggested that recent TE insertions in humans (like *Izanagi* in medaka) usually bring context-specific gene activities, while older TE insertions are more likely to correspond to broad enhancers [99]. In human, enhancers are globally depleted in recent TE insertions. However, enrichment of young TE families is observed in enhancers of genes specifically expressed in testis [99].

Regulation by piRNAs (Fig. 2b)

TEs can affect the regulation of genes in *trans* via piRNAs. If piRNAs are originally devoted to the down-regulation of TEs, there is now accumulating evidence that piRNAs regulate host developmental genes and maternal mRNA decay [100]. As an example, TE-derived piRNAs can target maternally-deposited copies of the Drosophila embryo *nos* mRNA for degradation, which is required for a proper development of the head [101]. The region of the *nos* 3′ untranslated region that is recognized by the piRNAs originates from two different TEs [101]. We can find some evidence of such regulation in gonads. In Drosophila ovarian somatic sheet cells a piRNA knock-down affects the expression of about 100 transcripts [102]. Most of these deregulated transcripts originate from TEs, but a significant part of them still corresponds to host protein-coding genes, with different genes being affected according to cell lineage. Some of these genes presented de novo inserted TEs in their introns or UTRs that induced suppression by the PIWI machinery at the nascent RNA level [102]. In mouse spermatocytes, piRNAs derived from TEs were shown to mediate the degradation of numerous mRNAs and lncRNAs [103]. This regulation involves PIWIL1, a major actor of the piRNA pathway, the

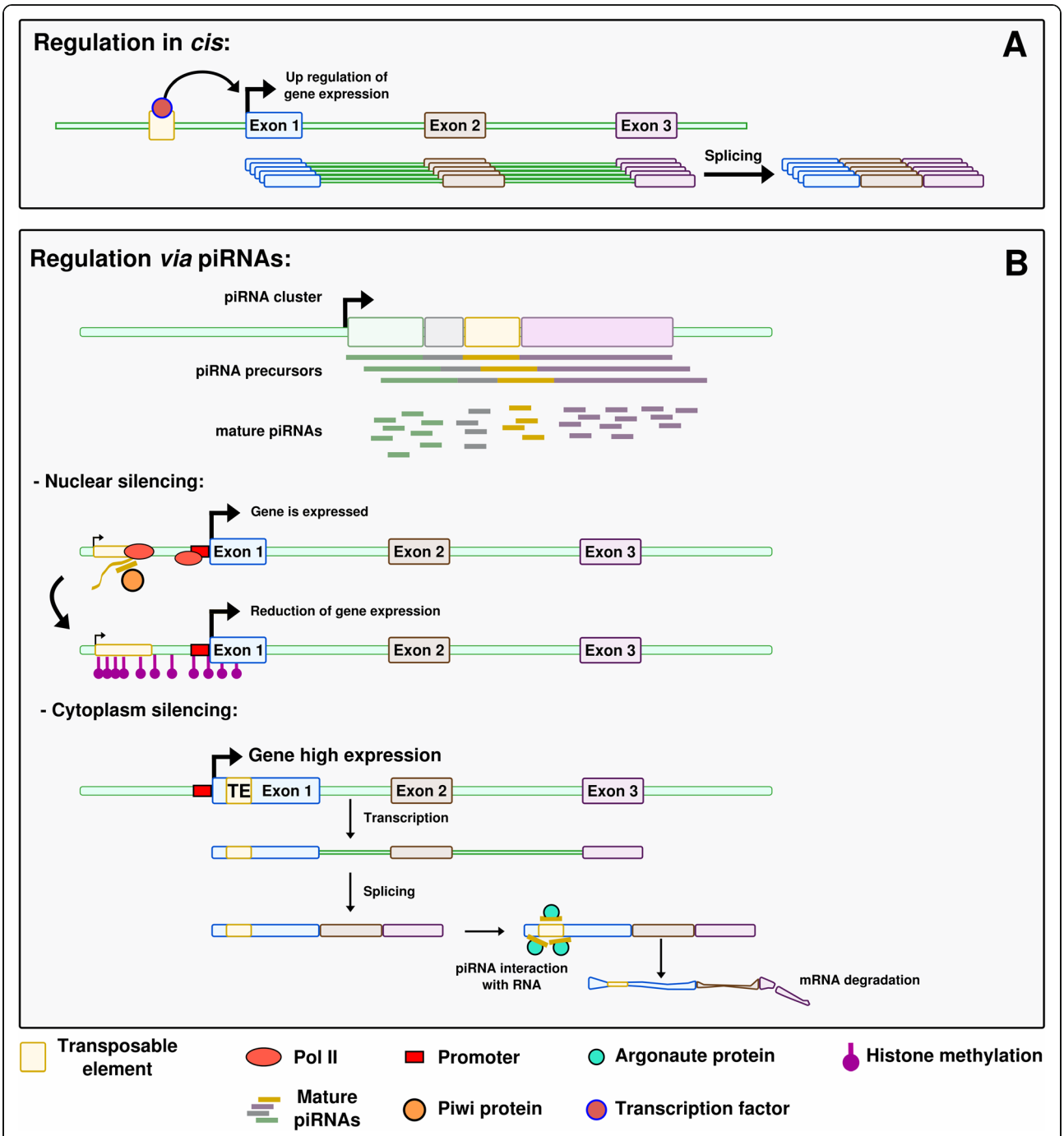

Fig. 2 Different ways how TEs can affect gene expression. **a** Regulation in *cis*. The TE brings a ready-to-use regulatory sequence that carries a transcription factor binding site. The transcription factor can bind on this site and influence expression of the neighboring gene. **b** Regulation via piRNAs. In the nuclear silencing situation, a TE is present close to the gene of interest. The piRNA, via the Piwi protein, triggers histone modifications that silence the TE but also affect the RNA polymerase binding region of the neighboring gene. Because of the epigenetic modification of the TE, the gene expression is reduced. In the cytoplasm silencing situation, a TE-derived sequence is present in the 5'UTR of the gene. piRNAs specific to this TE bind the transcript in the cytoplasm via an Argonaute protein and trigger the degradation of the transcript

knockdown of which leads to the upregulation of 172 genes. piRNAs were shown to target in particular retrotransposon sequences located in the 3′ UTR of mRNAs [103]. TE-derived sequences thus play a role in the control of germline expressed genes through piRNAs.

Some piRNAs have been demonstrated to trigger sex determination. In *Bombyx mori*, a species where the sex

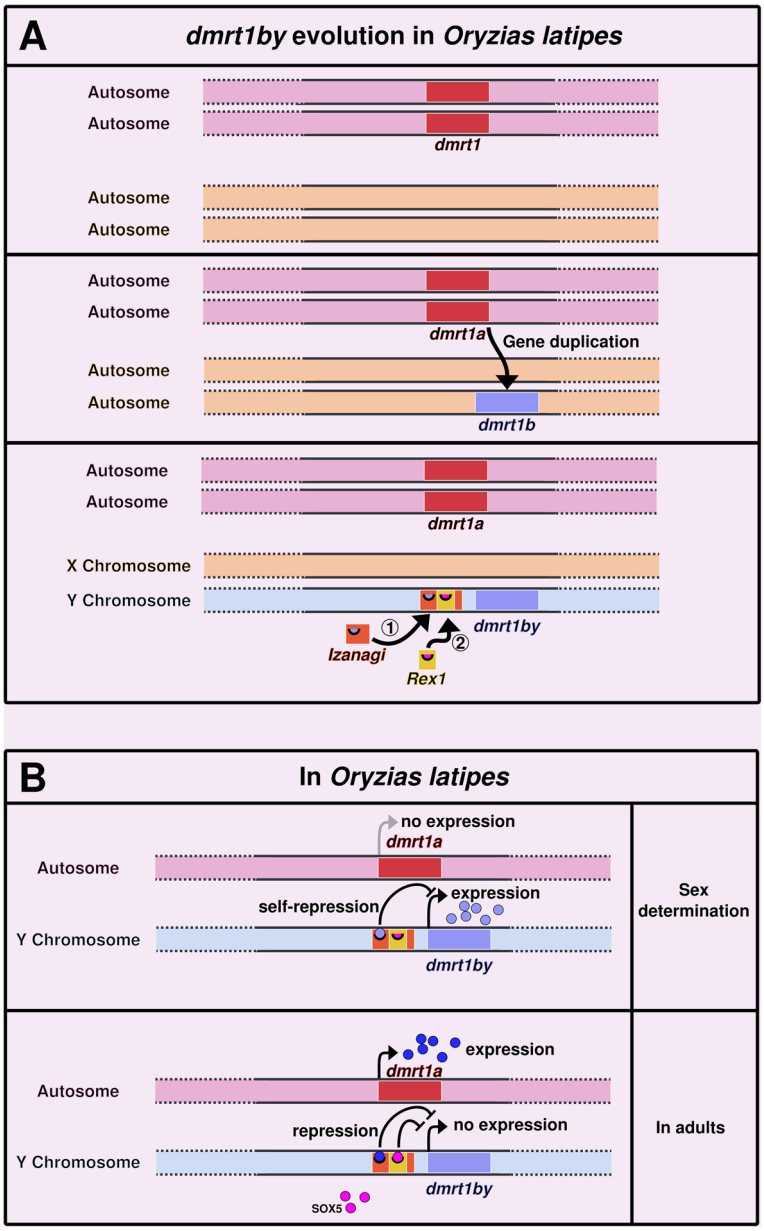

Fig. 3 dmrt1bY evolution and regulation in *Oryzias latipes*. **a** In the ancestor, the *dmrt1* gene existed in a single copy on a pair of autosomes. *dmrt1* was then duplicated into *dmrt1a* and *dmrt1b*. Later, two TEs inserted upstream of *dmrt1b*: *Izanagi,* a DNA/P-element, followed by *Rex1,* a LINE retrotransposon [98]. After the insertion of these TEs, *dmrt1b* became the master sex-determining gene *dmrt1bY* and the chromosome harboring it became the Y chromosome (the gene is absent from the X). **b** *dmrt1bY* is expressed during sex determination in the prospective males. Its product triggers sex determination towards the male phenotype. It also binds on its own binding site in *Izanagi,* down-regulating its own expression. After sex determination and in adults, *dmrt1a,* the ancestral paralog of *dmrt1bY*, is expressed. It binds to *Izanagi*, down-regulating and silencing *dmrt1bY* once sex determination has occurred. This silencing is also ensured by the binding of Sox5 to a motif encompassed in the *Rex1* sequence

determining system is ZW/ZZ, the master sex-determining region is localized on the W chromosome and produces female enriched piRNAs deriving from TEs and repetitive sequences. The *Fem* piRNA encoded in this sex-determining region of the W chromosome derives from a non-TE repetitive region and forms a complex with a silkworm equivalent of the Piwi protein. The complex targets and cleaves a masculinizing protein-coding mRNA transcribed from the Z chromosome, triggering feminization [104, 105]. A similar example has been described in *C. elegans*, where the *21ux-1* piRNA downregulates the *xol-1* gene involved in

X chromosome dosage compensation and sex determination [42]. This piRNA control of *xol-1* appears to be conserved in the related nematode *C. briggsae*, suggesting a robust involvement of piRNA in controlling gene expression [42]. In these two examples however, neither the piRNA nor its target were shown to be derived from TEs. In mammals, as described before, the inactivation of the epigenetic control of TEs in male gonads leads to azoospermia and thus infertility [86]. However, a certain relaxation of epigenetic control is observed in the germline, leading to demethylation of TEs and their reactivation. At a first look, this could be considered as deleterious for the host. The relaxation happening in the germline leads to a low level of TE activity that is actually thought to allow the host to sense the TEs present in the genome [86]. Such sensing would help to better control TE transposition. According to the authors, this sensing could be ensured by piRNAs. Relaxation of the epigenetic control allows TE expression that itself triggers piRNA production. piRNAs could then limit the impact of TEs but also regulate the expression of other genes, and through these possibly participate in sexual development. Taken together, the presence of TEs in genomes could be linked to the fact that they have an indirect effect, via piRNAs, on the control of specific genes, and sometimes on critical event such as sexual development.

TEs are involved in sex chromosome structure and evolution

We have described how sex can influence TEs expression, and reciprocally how TEs can modulate expression of genes involved in sexual development. In addition to effects of TE on host gene expression, genomic differences can exist between males and females in terms of TE and gene position and content. These differences can impact sexual development.

In mammals, the X and Y chromosomes are derived from a same pair of autosomes. Accordingly, even if the Y chromosome has lost many of its genes due to suppression of recombination, most genes carried on the Y chromosome have homologs on the X chromosome. This scenario of gene loss, however, does not appear universal, since in certain cases, like in *Drosophila melanogaster*, sex chromosomes evolved more through gene gain [106]. In the platyfish (*Xiphophorus maculatus)*, an accumulation of *Texim* genes is observed on the Y chromosome [107]. These genes are physically associated to a Helitron transposon, which might have spread the *Texim* sequences on the Y chromosome but not on the X. In salmonids, recent findings on SD showed that the master sex-determining gene, *sdY*, is conserved in many species. However, it does not always locate on the same chromosome, but instead seems to behave like a "jumping gene" [108, 109]. An analysis of the boundaries of the moving region that carries *sdY* revealed the presence of several TE sequences, leading authors to propose a mechanism of TE-associated transduction [108, 109]. This phenomenon could be linked to a rapid turnover of sexual chromosomes in this clade. Other examples of such sex determining "jumping genes" have been described in animals, such as in the house fly [110] or in *Chironomus* species [111]. In these cases the possible involvement of TEs in the translocation of the determining cassette has not been investigated, but we can notice that, in the case of the house fly, about two thirds of the Y-linked scaffolds present sequence similarities with TEs [110].

TEs can also themselves present sex-specific localizations. As described before, in *Drosophila miranda* the recently formed X chromosome, called "neo-X", accumulates Helitron DNA transposons [92]. The success of fixation of this TE on this specific sex chromosome is probably linked to its role in the expression of X-chromosomal genes, bringing an evolutionary advantage (see part 2A) [92]. Sex chromosomes are actually often enriched in TEs [112–115]. This accumulation might be in some cases the consequence of the impossibility for sex chromosomes to recombine and thus eliminate deleterious insertions. In the genome of the African clawed frog *Xenopus laevis*, recombination between W and Z sex chromosomes stopped recently, and a large accumulation of TEs already started in the W specific regions [115]. Such accumulation has also been observed on several young sex chromosomes of teleost fishes [112]. The higher density of TEs on these chromosomes might increase their probability to regulate some key sexual development genes and consequently to impact sexual development. In birds, such as woodpeckers for instance, the female specific chromosome W is enriched in CR1 insertions, which is a retrotransposon [116, 117]. In human, the Y chromosome is a hot spot for specific TE insertions [118]. All TE types show a higher density on the Y compared to autosomes, except for SVA short retrotransposons. In particular, density is 30 times higher than the genome average for LTR elements, and four times higher for *Alu* and L1 elements. The authors assume that this cannot be due to a genome assembly artifact, since the enrichment varies according to TE families. Nevertheless, they do not provide any explanation for the insertion rate differences between TE types on the Y chromosome. This high TE density on the Y chromosome is not explainable by low gene density as human chromosome 13 has a lower gene density and is not enriched for TEs [118]. This accumulation of active elements suggests that the Y chromosome is not shrinking in man, but still expanding through new insertions [119]. Of note, in contrast to what is observed in mammals and birds, the heterogametic sex chromosome (W

or Y), in many fish, reptiles and amphibians, is much larger than the Z or X, and often the largest chromosome of the complement. In these groups, sex chromosomes are usually younger than in mammals and birds, with frequent turnover. In addition to bringing additional DNA material, it has been hypothesized that TE insertions could favor, in a fast and effective manner, structural differences between gonosomes, that in turn help the expansion of the region of suppressed recombination [120]. This could thus lead to an increase in sex chromosome size during the early phase of their differentiation, while size diminishing would occur later in their evolution [120]. The accumulation of TEs and other repetitive sequences on the Y chromosome has been hypothesized to globally impact the chromatin landscape of the genome [121, 122]. Indeed, polymorphic Y chromosomes that differ only by their quantity of repeats are associated to different levels of chromatin repression on autosomes [122]. The high density of TEs and satellite DNA on the Y chromosome could function as a sink for heterochromatin marks, leading to a dilution of these marks in the rest of the genome, and hence to differential expression between males and females [122].

The X chromosome inactivation in mammals, also called Lyonisation, is a dosage compensation process in which one of the two X chromosomes is inactivated in XX females, preventing gene overexpression compared to males, which have a single X [123, 124]. The enrichment of LINE retrotransposons on the X chromosomes of human and mice led to the hypothesis of an involvement of LINEs in this process [114, 124]. This hypothesis has been tested in the spiny rat *Tokudaia osimensis*, where males and females are XO [125]. No dosage compensation by X inactivation is required here, suggesting that LINEs would not be required on this X chromosome. Interestingly, the authors describe a similar high concentration of LINEs on this X chromosome compared to humans or mice. They conclude that the accumulation of TEs on X chromosomes might be only a by-product of reduced recombination [125]. This idea was also reviewed later by Lyon, leading to the same conclusion [126]. Further investigations on the role of LINEs in X chromosome inactivation have been conducted in mammals. On the human X chromosome, regions poor in L1 elements contain genes escaping X inactivation [127]. In placental mammals, the inactivated X chromosome is coated with Xist (X-inactive specific transcript) RNAs, which have a silencing effect. These regions are composed of silent LINEs that are closed in chromatin 3D structure, and are formed prior to gene inactivation [128, 129]. As genes "move" in the Xist silenced region via a modification of the 3D conformation of the chromosome, they become inactivated. Conversely, LINE poor regions are physically distant from Xist silenced regions [123, 129]. In these studies, the authors show that LINEs play a role in the spread of X chromosome silencing by recruiting Xist RNAs, suggesting a general role in the regulation of X-chromosomal gene expression. This phenomenon also exemplifies that for understanding chromosomal organization the intricate structure and function relationships have to be considered.

Conclusions

Sex is an important parameter to take into account when performing experiments, in particular when analyzing gene expression [130]. Many studies, including genome sequencing, are conducted in individuals of only one sex, and results observed might not be generalizable to the other [131]. We presented in this review the many facets linking sex with TEs, both influencing each other in a co-evolutionary process. TE expression in germlines is essential for them to get fixed in the genome and be transmitted vertically. Conversely, TEs have an influence on sex differentiation mechanisms, for example through the intermediary of piRNAs. They could also influence sex evolution by the regulatory novelties they create. TEs are indeed great tools for evolution as they can rapidly propagate regulatory elements and thus provide the necessary rewiring of the genetic network. The high density of TEs on sex chromosomes, linked to the absence of recombination of these chromosomes, could increase the probability for TEs to locate in the vicinity of sexual development genes and interact with them. They can influence and be influenced by sex depending on the process studied.

Another way TEs can influence gene expression is by triggering alternative splicing, via the new splicing sites they sometimes bring with them [9]. In the case of sexual development gene regulation, however, such involvement of TEs has yet to be demonstrated. In *Drosophila melanogaster*, some intron retention events are known to be linked to sex [132]. Although the exact trigger of the alternative splicing is not clearly elucidated for now, a hypothesis proposed that the high coverage of repetitive sequences on the Y chromosome could be involved in the process, as presented earlier in this review: the Y chromosome would attract on its repeats high quantities of chromatin-modifying proteins, which would in turn lead to a global modification of the chromatin state on other chromosomes, and in the end would affect the accessibility of splicing factors to the nascent transcripts. Here, the impact of TEs on the splicing machinery would thus be indirect and not specific to particular genes.

Finally, genes involved in sexual development and sexual functions seem to evolve faster than other genes

[133, 134]. These observations of positive selection and rapid evolution are not really consistent with earlier observations of the sex determination and differentiation cascade. Indeed, a popular model, formulated by Graham in 2003, states that "masters change, slaves remain" [135], where "masters" refer to genes at the top of the sex determination cascade, and "slaves" to genes acting at the end of the cascade. A renewal of this initial proposition has been proposed by Herpin et al.: "When masters change, some slaves remain, others are dismissed or acquire new tasks, and new ones can be hired" [34, 55]. Knowing that TEs are a source of genomic diversification, studying the evolution of sexual development genes in the perspective of TEs, just as the evolution of their regulation, could reveal interesting trends. A perspective could be to investigate RNA-seq dataset for species-specific sex-biased genes associated to TE location variation between closely related species to reveal candidate genes recently controlled by TEs. Global approaches by sequencing piRNAs and mapping them to sex-biased genes could also give more clues about the regulation and evolution of genes involved in sexual development and function.

Abbreviations

ESD: Environmental Sex Determination; GSD: Genetic Sex Determination; KAP1: KRAB-associated protein 1; KRAB-ZNF: Krüppel-associated box domain zinc finger; MRE: MSL Recognition Element; MSL: Male Specific Lethal; piRNA: Piwi-Interacting RNA; SD: Sex Determination; TE: Transposable Element

Acknowledgments

The authors sincerely thank Joanne Burden for her diligent proofreading of the article.

Authors' contributions

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

Funding

This work was supported by a PRCI (International Collaborative Research Project) grant overseen by the French National Research Agency (ANR) together with the German Research Agency (Deutsche Forschungsgemeinschaft) (ANR-16-CE92–0019 – EVOBOOSTER). This publication was funded by the German Research Foundation (DFG) and the University of Würzburg in the funding program Open Access Publishing.

Availability of data and materials

Data sharing not applicable to this article as no datasets were generated or analysed during the current study.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

[1]Institut de Genomique Fonctionnelle de Lyon, Univ Lyon, CNRS UMR 5242, Ecole Normale Superieure de Lyon, Universite Claude Bernard Lyon 1, 46 allee d'Italie, F-69364 Lyon, France. [2]Entwicklungsbiochemie, Biozentrum, Universität Würzburg, Würzburg, Germany. [3]The Xiphophorus Genetic Stock Center, Department of Chemistry and Biochemistry, Texas State University, San Marcos, TX, USA.

Received: 27 August 2019 Accepted: 21 October 2019
Published online: 03 November 2019

References

1. Biémont C, Vieira C. Genetics: junk DNA as an evolutionary force. Nature. 2006;443(7111):521–4.
2. McClintock B. The origin and behavior of mutable loci in maize. Proc Natl Acad Sci. 1950;36(6):344–55.
3. 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.
4. Kapitonov VV, Jurka J. A universal classification of eukaryotic transposable elements implemented in Repbase. Nat Rev Genet. 2008;9(5):411–2.
5. Seberg O, Petersen G. A unified classification system for eukaryotic transposable elements should reflect their phylogeny. Nat Rev Genet. 2009; 10(4):276.
6. Rishishwar L, Tellez Villa CE, Jordan IK. Transposable element polymorphisms recapitulate human evolution. Mob DNA. 2015;6:21.
7. Bourgeois Y, Boissinot S. On the Population Dynamics of Junk: A Review on the Population Genomics of Transposable Elements. Genes. 2019;10(6):419.
8. Petrov DA, Fiston-Lavier A-S, Lipatov M, Lenkov K, Gonzalez J. Population genomics of transposable elements in Drosophila melanogaster. Mol Biol Evol. 2011;28(5):1633–44.
9. Rebollo R, Romanish MT, Mager DL. Transposable elements: an abundant and natural source of regulatory sequences for host genes. Annu Rev Genet. 2012;46(1):21–42.
10. Dupressoir A, Marceau G, Vernochet C, Bénit L, Kanellopoulos C, Sapin V, et al. Syncytin-a and syncytin-B, two fusogenic placenta-specific murine envelope genes of retroviral origin conserved in Muridae. Proc Natl Acad Sci U S A. 2005;102(3):725–30.
11. Volff J-N. Turning junk into gold: domestication of transposable elements and the creation of new genes in eukaryotes. BioEssays News Rev Mol Cell Dev Biol. 2006;28(9):913–22.
12. Gilbert C, Feschotte C. Horizontal acquisition of transposable elements and viral sequences: patterns and consequences. Curr Opin Genet Dev. 2018;49:15–24.
13. Kapitonov VV, Jurka J. RAG1 core and V(D) J recombination signal sequences were derived from Transib transposons. PLoS Biol. 2005;3(6):e181.
14. Naville M, Warren IA, Haftek-Terreau Z, Chalopin D, Brunet F, Levin P, et al. Not so bad after all: retroviruses and long terminal repeat retrotransposons as a source of new genes in vertebrates. Clin Microbiol Infect. 2016;22(4): 312–23.
15. Dolgin ES, Charlesworth B. The fate of transposable elements in asexual populations. Genetics. 2006;174(2):817–27.
16. Hickey DA. Selfish DNA: a sexually-transmitted nuclear parasite. Genetics. 1982;101(3–4):519–31.
17. Wright S, Finnegan D. Genome evolution: sex and the transposable element. Curr Biol. 2001;11(8):R296–9.
18. Schaack S, Gilbert C, Feschotte C. Promiscuous DNA: horizontal transfer of transposable elements and why it matters for eukaryotic evolution. Trends Ecol Evol. 2010;25(9):537–46.
19. Zeyl C, Bell G, Green DM. Sex and the spread of retrotransposon Ty3 in experimental populations of *Saccharomyces cerevisiae*. Genetics. 1996;143(4): 1567–77.
20. Felsenstein J. The evolutionary advantage of recombination. Genetics. 1974; 78(2):737–56.
21. Hill WG, Robertson A. The effect of linkage on limits to artificial selection. Genet Res. 1966;8(3):269–94.
22. Arkhipova IR. Mobile genetic elements and sexual reproduction. Cytogenet Genome Res. 2005;110(1–4):372–82.
23. Barsoum E, Martinez P, Astrom SU. Alpha 3, a transposable element that promotes host sexual reproduction. Genes Dev. 2010;24(1):33–44.
24. Bestor TH. Sex brings transposons and genomes into conflict. Genetica. 1999;107(1–3):289–95.
25. Kraaijeveld K, Zwanenburg B, Hubert B, Vieira C, De Pater S, Van Alphen JJM, et al. Transposon proliferation in an asexual parasitoid. Mol Ecol. 2012;21(16): 3898–906.

26. Warren WC, García-Pérez R, Xu S, Lampert KP, Chalopin D, Stöck M, et al. Clonal polymorphism and high heterozygosity in the celibate genome of the Amazon molly. Nat Ecol Evol. 2018;2(4):669–79.
27. Arkhipova I, Meselson M. Transposable elements in sexual and ancient asexual taxa. Proc Natl Acad Sci U S A. 2000;97(26):14473–7.
28. Flot J-F, Hespeels B, Li X, Noel B, Arkhipova I, Danchin EGJ, et al. Genomic evidence for ameiotic evolution in the bdelloid rotifer Adineta vaga. Nature. 2013;500(7463):453–7.
29. Capel B. Vertebrate sex determination: evolutionary plasticity of a fundamental switch. Nat Rev Genet. 2017;18(11):675–89.
30. Pan Q, Anderson J, Bertho S, Herpin A, Wilson C, Postlethwait JH, et al. Vertebrate sex-determining genes play musical chairs. C R Biol. 2016;339(7–8):258–62.
31. Bull JJ, Vogt RC. Temperature-dependent sex determination in turtles. Science. 1979;206(4423):1186–8.
32. Lang JW, Andrews HV. Temperature-dependent sex determination in crocodilians. J Exp Zool. 1994;270(1):28–44.
33. Honeycutt JL, Deck CA, Miller SC, Severance ME, Atkins EB, Luckenbach JA, et al. Warmer waters masculinize wild populations of a fish with temperature-dependent sex determination. Sci Rep. 2019;9(1):6527.
34. Bachtrog D, Mank JE, Peichel CL, Kirkpatrick M, Otto SP, Ashman T-L, et al. Sex determination: why so many ways of doing it? PLoS Biol. 2014;12(7):e1001899.
35. Schartl M. Sex chromosome evolution in non-mammalian vertebrates. Curr Opin Genet Dev. 2004;14(6):634–41.
36. Koene JM. Sex determination and gender expression: reproductive investment in snails. Mol Reprod Dev. 2017;84(2):132–43.
37. Matsuda M. Sex determination in the teleost medaka, *Oryzias latipes*. Annu Rev Genet. 2005;39:293–307.
38. Nanda I, Kondo M, Hornung U, Asakawa S, Winkler C, Shimizu A, et al. A duplicated copy of DMRT1 in the sex-determining region of the Y chromosome of the medaka, *Oryzias latipes*. Proc Natl Acad Sci. 2002;99(18):11778–83.
39. Gilbert SF. Chromosomal sex determination in drosophila. Dev Biol 6th Ed. 2000.
40. Sanchez L, Chaouiya C. Logical modelling uncovers developmental constraints for primary sex determination of chicken gonads. J R Soc Interface. 2018;15(142):20180165.
41. Stothard P, Pilgrim D. Sex-determination gene and pathway evolution in nematodes. BioEssays News Rev Mol Cell Dev Biol. 2003;25(3):221–31.
42. Tang W, Seth M, Tu S, Shen E-Z, Li Q, Shirayama M, et al. A Sex Chromosome piRNA Promotes Robust Dosage Compensation and Sex Determination in *C. elegans*. Dev Cell. 2018;44(6):762–770.e3.
43. Eggers S, Ohnesorg T, Sinclair A. Genetic regulation of mammalian gonad development. Nat Rev Endocrinol. 2014;10(11):673–83.
44. Hsu C, Pan Y-J, Wang Y-W, Tong S-K, Chung B. Changes in the morphology and gene expression of developing zebrafish gonads. Gen Comp Endocrinol. 2018;265:154–9.
45. Croft B, Ohnesorg T, Hewitt J, Bowles J, Quinn A, Tan J, et al. Human sex reversal is caused by duplication or deletion of core enhancers upstream of SOX9. Nat Commun. 2018;9(1):5319.
46. Gonen N, Futtner CR, Wood S, Garcia-Moreno SA, Salamone IM, Samson SC, et al. Sex reversal following deletion of a single distal enhancer of Sox9. Science. 2018;360(6396):1469–73.
47. Schartl M, Schories S, Wakamatsu Y, Nagao Y, Hashimoto H, Bertin C, et al. Sox5 is involved in germ-cell regulation and sex determination in medaka following co-option of nested transposable elements. BMC Biol. 2018;16:16.
48. Assis R, Zhou Q, Bachtrog D. Sex-biased transcriptome evolution in drosophila. Genome Biol Evol. 2012;4(11):1189–200.
49. Böhne A, Sengstag T, Salzburger W. Comparative transcriptomics in east african cichlids reveals sex- and species-specific expression and new candidates for sex differentiation in fishes. Genome Biol Evol. 2014;6(9):2567–85.
50. Catalan A, Hutter S, Parsch J. Population and sex differences in *Drosophila melanogaster* brain gene expression. BMC Genomics. 2012;13:1–12.
51. Ledón-Rettig CC, Zattara EE, Moczek AP. Asymmetric interactions between doublesex and tissue- and sex-specific target genes mediate sexual dimorphism in beetles. Nat Commun. 2017;8:14593.
52. Liu H, Lamm MS, Rutherford K, Black MA, Godwin JR, Gemmell NJ. Large-scale transcriptome sequencing reveals novel expression patterns for key sex-related genes in a sex-changing fish. Biol Sex Differ. 2015;6:26.
53. Shi L, Zhang Z, Su B. Sex biased gene expression profiling of human brains at major developmental stages. Sci Rep. 2016;6:21181.
54. Waters PD, Wallis MC, Graves JAM. Mammalian sex—origin and evolution of the Y chromosome and SRY. Semin Cell Dev Biol. 2007;18(3):389–400.
55. Herpin A, Schartl M. Plasticity of gene-regulatory networks controlling sex determination: of masters, slaves, usual suspects, newcomers, and usurpators. EMBO Rep. 2015;16(10):1260–74.
56. Schartl M. A comparative view on sex determination in medaka. Mech Dev. 2004;121(7–8):639–45.
57. Matsuda M, Sakaizumi M. Evolution of the sex-determining gene in the teleostean genus Oryzias. Gen Comp Endocrinol. 2016;239:80–8.
58. Herpin A, Adolfi MC, Nicol B, Hinzmann M, Schmidt C, Klughammer J, et al. Divergent expression regulation of gonad development genes in medaka shows incomplete conservation of the downstream regulatory network of vertebrate sex determination. Mol Biol Evol. 2013;30(10):2328–46.
59. Takehana Y, Matsuda M, Myosho T, Suster ML, Kawakami K, Shin T, et al. Co-option of Sox3 as the male-determining factor on the Y chromosome in the fish *Oryzias dancena*. Nat Commun. 2014;5:4157.
60. Chen S, Zhang G, Shao C, Huang Q, Liu G, Zhang P, et al. Whole-genome sequence of a flatfish provides insights into ZW sex chromosome evolution and adaptation to a benthic lifestyle. Nat Genet. 2014;46(3):253–60.
61. Deloger M, Cavalli FMG, Lerat E, Biémont C, Sagot M-F, Vieira C. Identification of expressed transposable element insertions in the sequenced genome of *Drosophila melanogaster*. Gene. 2009;439(1–2):55–62.
62. Graveley BR, Brooks AN, Carlson JW, Duff MO, Landolin JM, Yang L, et al. The developmental transcriptome of *Drosophila melanogaster*. Nature. 2011;471(7339):473–9.
63. Lipatov M, Lenkov K, Petrov DA, Bergman CM. Paucity of chimeric gene-transposable element transcripts in the *Drosophila melanogaster* genome. BMC Biol. 2005;3(1):24.
64. Sienski G, Dönertas D, Brennecke J. Transcriptional silencing of transposons by Piwi and maelstrom and its impact on chromatin state and gene expression. Cell. 2012;151(5):964–80.
65. Brunet F, Roche A, Chalopin D, Naville M, Klopp C, Vizziano-Cantonnet D, et al. Analysis of transposable elements expressed in the gonads of the siberian sturgeon. In: Williot P, Nonnotte G, Vizziano-Cantonnet D, Chebanov M, editors. The Siberian Sturgeon (*Acipenser baerii*, Brandt, 1869) Volume 1 - Biology. Cham: Springer International Publishing; 2018. p. 115–30.
66. Esteve-Codina A, Kofler R, Palmieri N, Bussotti G, Notredame C, Perez-Enciso M. Exploring the gonad transcriptome of two extreme male pigs with RNA-seq. BMC Genomics. 2011;12:552.
67. Laski FA, Rio DC, Rubin GM. Tissue specificity of Drosophila P element transposition is regulated at the level of mRNA splicing. Cell. 1986;44(1):7–19.
68. Iwasaki YW, Siomi MC, Siomi H. PIWI-Interacting RNA: Its biogenesis and functions. Annu Rev Biochem. 2015;84(1):405–33.
69. 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.
70. 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.
71. Kim VN, Han J, Siomi MC. Biogenesis of small RNAs in animals. Nat Rev Mol Cell Biol. 2009;10(2):126–39.
72. Kelleher ES. Reexamining the P-element invasion of *Drosophila melanogaster* through the lens of piRNA silencing. Genetics. 2016;203(4):1513–31.
73. Hill T, Schlötterer C, Betancourt AJ. Hybrid dysgenesis in Drosophila simulans associated with a rapid invasion of the P-element. PLoS Genet. 2016;12(3):e1005920.
74. Ecco G, Imbeault M, Trono D. KRAB zinc finger proteins. Dev. 2017;144(15):2719–29.
75. Molaro A, Malik HS. Hide and seek: how chromatin-based pathways silence retroelements in the mammalian germline. Curr Opin Genet Dev. 2016;37:51–8.
76. Wolf D, Goff SP. Embryonic stem cells use ZFP809 to silence retroviral DNAs. Nature. 2009;458(7242):1201–4.
77. Jacobs FMJ, Greenberg D, Nguyen N, Haeussler M, Ewing AD, Katzman S, et al. An evolutionary arms race between KRAB zinc-finger genes ZNF91/93 and SVA/L1 retrotransposons. Nature. 2014;516(7530):242–5.
78. Tan X, Xu X, Elkenani M, Smorag L, Zechner U, Nolte J, et al. Zfp819, a novel KRAB-zinc finger protein, interacts with KAP1 and functions in genomic integrity maintenance of mouse embryonic stem cells. Stem Cell Res. 2013;11(3):1045–59.

79. Helleboid P-Y, Heusel M, Duc J, Piot C, Thorball CW, Coluccio A, et al. The interactome of KRAB zinc finger proteins reveals the evolutionary history of their functional diversification. EMBO J. 2019;38(18):e101220.
80. Brennecke J, Malone CD, Aravin AA, Sachidanandam R, Stark A, Hannon GJ. An epigenetic role for maternally inherited piRNAs in transposon silencing. Science. 2008;322(5906):1387–92.
81. Zempleni J, Chew YC, Bao B, Pestinger V, Wijeratne SSK. Repression of transposable elements by histone biotinylation. J Nutr. 2009;139(12):2389–92.
82. Slotkin RK, Martienssen R. Transposable elements and the epigenetic regulation of the genome. Nat Rev Genet. 2007;8(4):272–85.
83. Capy P, Gasperi G, Biémont C, Bazin C. Stress and transposable elements: co-evolution or useful parasites? Heredity. 2000;85(2):101.
84. Dong Y, Huang Z, Kuang Q, Wen Z, Liu Z, Li Y, et al. Expression dynamics and relations with nearby genes of rat transposable elements across 11 organs, 4 developmental stages and both sexes. BMC Genomics. 2017;18:666.
85. 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.
86. Zamudio N, Bourc'his D. Transposable elements in the mammalian germline: a comfortable niche or a deadly trap? Heredity. 2010;105(1):92–104.
87. Hadziselimovic F, Hadziselimovic NO, Demougin P, Krey G, Oakeley E. Piwi-pathway alteration induces LINE-1 transposon derepression and infertility development in cryptorchidism. Sex Dev. 2015;9(2):98–104.
88. Malki S, van der Heijden GW, O'Donnell KA, Martin SL, Bortvin A. A role for Retrotransposon LINE-1 in fetal oocyte attrition in mice. Dev Cell. 2014;29(5):521–33.
89. Feschotte C. Transposable elements and the evolution of regulatory networks. Nat Rev Genet. 2008;9(5):397–405.
90. Sundaram V, Wang T. Transposable element mediated innovation in gene regulatory landscapes of cells: re-visiting the "gene-battery" model. BioEssays. 2017;40(1):1700155.
91. Sundaram V, Choudhary MNK, Pehrsson E, Xing X, Fiore C, Pandey M, et al. Functional cis-regulatory modules encoded by mouse-specific endogenousretrovirus. Nat Commun. 2017;8:14550.
92. Ellison CE, Bachtrog D. Dosage compensation via transposable element mediated rewiring of a regulatory network. Science. 2013;342(6160):846–50.
93. Ellison CE, Bachtrog D. Non-allelic gene conversion enables rapid evolutionary change at multiple regulatory sites encoded by transposable elements. Elife. 2015;4:e05899.
94. Ellison C, Bachtrog D. Contingency in the convergent evolution of a regulatory network: dosage compensation in Drosophila. PLoS Biol. 2019;17(2):e3000094.
95. Nagao A, Mituyama T, Huang H, Chen D, Siomi MC, Siomi H. Biogenesis pathways of piRNAs loaded onto AGO3 in the Drosophila testis. RNA. 2010;16(12):2503–15.
96. Malone CD, Lehmann R, Teixeira FK. The cellular basis of hybrid dysgenesis and stellate regulation in Drosophila. Curr Opin Genet Dev. 2015;34:88–94.
97. Kotelnikov RN, Klenov MS, Rozovsky YM, Olenina LV, Kibanov MV, Gvozdev VA. Peculiarities of piRNA-mediated post-transcriptional silencing of stellate repeats in testes of *Drosophila melanogaster*. Nucleic Acids Res. 2009;37(10):3254–63.
98. 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. Petrov DA, éditeur. PLoS Genet. 2010;6(2):e1000844.
99. 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.
100. Rojas-Ríos P, Simonelig M. piRNAs and PIWI proteins: regulators of gene expression in development and stem cells. Dev. 2018;145(17):dev161786.
101. Rouget C, Papin C, Boureux A, Meunier A-C, Franco B, Robine N, et al. Maternal mRNA deadenylation and decay by the piRNA pathway in the early Drosophila embryo. Nature. 2010;467(7319):1128–32.
102. Sytnikova YA, Rahman R, Chirn G, Clark JP, Lau NC. Transposable element dynamics and PIWI regulation impacts lncRNA and gene expression diversity in Drosophila ovarian cell cultures. Genome Res. 2014;24(12):1977–90.
103. 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.
104. Katsuma S, Kawamoto M, Kiuchi T. Guardian small RNAs and sex determination. RNA Biol. 2014;11(10):1238–42.
105. Kiuchi T, Koga H, Kawamoto M, Shoji K, Sakai H, Arai Y, et al. A single female-specific piRNA is the primary determiner of sex in the silkworm. Nature. 2014;509(7502):633–6.
106. Carvalho AB, Vicoso B, Russo CAM, Swenor B, Clark AG. Birth of a new gene on the Y chromosome of *Drosophila melanogaster*. Proc Natl Acad Sci U S A. 2015;112(40):12450–5.
107. Tomaszkiewicz M, Chalopin D, Schartl M, Galiana D, Volff J-N. A multicopy Y-chromosomal SGNH hydrolase gene expressed in the testis of the platyfish has been captured and mobilized by a Helitron transposon. BMC Genet. 2014;15:44.
108. Faber-Hammond JJ, Phillips RB, Brown KH. Comparative analysis of the shared sex-determination region (SDR) among salmonid fishes. Genome Biol Evol. 2015;7(7):1972–87.
109. Lubieniecki KP, Lin S, Cabana EI, Li J, Lai YYY, Davidson WS. Genomic Instability of the Sex-Determining Locus in Atlantic Salmon (*Salmo salar*). G3. 2015;5(11):2513–22.
110. Meisel RP, Gonzales CA, Luu H. The house fly Y chromosome is young and minimally differentiated from its ancient X chromosome partner. Genome Res. 2017;27(8):1417–26.
111. Martin J, Kuvangkadilok C, Peart DH, Lee BTO. Multiple sex determining regions in a group of related Chironomus species (Diptera:Chironomidae). Heredity. 1980;44(3):367–82.
112. Chalopin D, Naville M, Plard F, Galiana D, Volff J-N. Comparative analysis of transposable elements highlights mobilome diversity and evolution in vertebrates. Genome Biol Evol. 2015;7(2):567–80.
113. Erlandsson R, Wilson JF, Pääbo S. Sex chromosomal transposable element accumulation and male-driven substitutional evolution in humans. Mol Biol Evol. 2000;17(5):804–12.
114. Lyon MF. The Lyon and the LINE hypothesis. Semin Cell Dev Biol. 2003;14(6):313–8.
115. Mawaribuchi S, Takahashi S, Wada M, Uno Y, Matsuda Y, Kondo M, et al. Sex chromosome differentiation and the W- and Z-specific loci in *Xenopus laevis*. Dev Biol. 2017;426(2):393–400.
116. Bertocchi NA, de Oliveira TD, Del Valle GA, Coan RLB, Gunski RJ, Martins C, et al. Distribution of CR1-like transposable element in woodpeckers (*Aves Piciformes*): Z sex chromosomes can act as a refuge for transposable elements. Chromosom Res. 2018;26(4):333–43.
117. Śliwińska EB, Martyka R, Tryjanowski P. Evolutionary interaction between W/Y chromosome and transposable elements. Genetica. 2016;144(3):267–78.
118. Tang W, Mun S, Joshi A, Han K, Liang P. Mobile elements contribute to the uniqueness of human genome with 15,000 human-specific insertions and 14 Mbp sequence increase. DNA Res. 2018;25(5):521–33.
119. Griffin DK. Is the Y chromosome disappearing?—both sides of the argument. Chromosom Res. 2012;20(1):35–45.
120. Schartl M, Schmid M, Nanda I. Dynamics of vertebrate sex chromosome evolution: from equal size to giants and dwarfs. Chromosoma. 2016;125(3):553–71.
121. Brown EJ, Bachtrog D. The chromatin landscape of Drosophila: comparisons between species, sexes, and chromosomes. Genome Res. 2014;24(7):1125–37.
122. Lemos B, Branco AT, Hartl DL. Epigenetic effects of polymorphic Y chromosomes modulate chromatin components, immune response, and sexual conflict. Proc Natl Acad Sci. 2010;107(36):15826–31.
123. Chow J, Heard E. X inactivation and the complexities of silencing a sex chromosome. Curr Opin Cell Biol. 2009;21(3):359–66.
124. Lyon MF. X-chromosome inactivation: a repeat hypothesis. Cytogenet Cell Genet. 1998;80(1–4):133–7.
125. Scott LA, Kuroiwa A, Matsuda Y, Wichman HA. X accumulation of LINE-1 retrotransposons in *Tokudaia osimensis*, a spiny rat with the karyotype XO. Cytogenet Genome Res. 2006;112(3–4):261–9.
126. Lyon MF. Do LINEs have a role in X-chromosome inactivation? J Biomed Biotechnol. 2006;2006(1):59746.
127. Bailey JA, Carrel L, Chakravarti A, Eichler EE. Molecular evidence for a relationship between LINE-1 elements and X chromosome inactivation: the Lyon repeat hypothesis. Proc Natl Acad Sci U S A. 2000;97(12):6634–9.
128. Chaumeil J, Le Baccon P, Wutz A, Heard E. A novel role for Xist RNA in the formation of a repressive nuclear compartment into which genes are recruited when silenced. Genes Dev. 2006;20(16):2223–37.
129. 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.
130. Jazin E, Cahill L. Sex differences in molecular neuroscience: from fruit flies to humans. Nat Rev Neurosci. 2010;11(1):9–17.
131. Shansky RM. Are hormones a "female problem" for animal research? Science. 2019;364(6443):825–6.

132. Wang M, Branco AT, Lemos B. The Y chromosome modulates splicing and sex-biased intron retention rates in Drosophila. Genetics. 2018; 208(3):1057–67.
133. Ellegren H, Parsch J. The evolution of sex-biased genes and sex-biased gene expression. Nat Rev Genet. 2007;8(9):689–98.
134. Yang L, Zhang Z, He S. Both male-biased and female-biased genes evolve faster in fish genomes. Genome Biol Evol. 2016;8(11):3433–45.
135. Graham P, Penn JKM, Schedl P. Masters change, slaves remain. BioEssays. 2003;25(1):1–4.

Publisher's Note

1.4 L'outil informatique au service de l'analyse des éléments transposables

L'avènement des techniques de séquençage de nouvelle génération (NGS) permet d'avoir accès au génome de plus en plus d'espèces de plus en plus rapidement. La présence d'ET dans les génomes des eucaryotes étudiés à ce jour (C. ELEGANS SEQUENCING CONSORTIUM, 1998; CHINWALLA *et al.*, 2002; LANDER *et al.*, 2001; THE ARABIDOPSIS GENOME INITIATIVE, 2000), couplée au séquençage de nouveaux génomes, est une source de données majeure pour l'étude de l'évolution des ET et des génomes. L'outil informatique est central à l'analyse des données issues des techniques de séquençage récentes, aussi bien pour l'assemblage des génomes que pour leur annotation. De nombreux programmes ont en particulier été développés dans le but d'étudier la dynamique des ET dans les génomes. Certains permettent la construction de banques de consensus décrivant les différentes familles, d'autres les annotent directement dans les génomes. Enfin certains outils sont spécialisés dans l'analyse de l'expression des ET à partir de données transcriptomiques. Ces outils ont en commun leur capacité à gérer de fortes similarités de séquences entre copies multiples d'ET, ce qui est moins problématique lors de l'étude des gènes.

1.4.1 Le séquençage nouvelle génération

1.4.1.1 Historique et différentes techniques

Les séquençages de type Sanger et de type Maxam-Gilbert, élaborés dans les années 70, ont permis pour la première fois d'accéder à la séquence en nucléotides d'une molécule d'ADN. Le séquençage de type Sanger, plus facilement automatisable, permet de séquencer l'ADN sur des longueurs de l'ordre de 1000 nucléotides avec une grande fiabilité. Il a notamment permis d'aboutir aux premières séquences complètes du génome humain en 2001 après environ 10 ans de travail (LANDER *et al.*, 2001; VENTER *et al.*, 2001). A la fin des années 2000, l'avènement du séquençage de nouvelle génération (NGS), aussi appelé séquençage de seconde génération, a apporté de nouvelles perspectives. Il est aujourd'hui possible de reséquencer en une fois un génome humain complet pour moins de 1000€ (**Fig. 1.14.**). La technologie Illumina est la plus répandue actuellement (ILLUMINA, 2010). Elle diffère du séquençage Sanger en plusieurs points. Plutôt que d'obtenir des séquences relativement longues, le séquençage Illumina génère et séquence de courts fragments d'ADN, de l'ordre de 100pb. Le principe est de fragmenter l'ADN à séquencer et de lier des adaptateurs à chaque fragment généré. Ces adaptateurs permettent de fixer les fragments d'ADN obtenus à une lame de verre ("flow cell") avant de procéder à une étape d'amplification par PCR (Réaction de Polymérisation en Chaine), nécessaire à la détection du signal lumineux permettant de suivre le séquençage. Sur la lame de verre, des groupes d'environ 1000 copies de chaque fragment sont formés. Ensuite, l'un des deux brins d'ADN est éliminé de chaque fragment pour pouvoir procéder au séquençage par synthèse. L'incorporation des nucléotides émet un signal fluorescent de couleur différente selon la base, signal qui est enregistré (ILLUMINA, 2010). Les différents fragments initialement formés sont ainsi séquencés en parallèle. Les NGS permettent de séquencer des génomes (DNA-seq), mais aussi des transcriptomes (RNA-seq), qui représentent tous les ARN dans le tissu, l'organe, ou l'individu étudié. Généralement, les transcrits d'intérêt

sont ceux exprimés par des gènes codants. Ces ARN messagers représentent une petite fraction des ARN totaux présents dans les cellules, alors qu'une grande part d'entre eux correspond à des ARN ribosomiques (WARNER, 1999). Deux techniques existent pour sélectionner les ARNm avant séquençage : utiliser des oligo-dT qui s'hybrident à la queue poly-A des ARNm et permettent de les purifier, ou appliquer des protocoles de déplétion des ARN ribosomiques, ce qui peut être utile si l'on s'intéresse aussi aux ARN non-codants (ZHAO *et al.*, 2018).

Diverses variantes aux protocoles de séquençage NGS classiques existent et permettent d'accéder à différentes données. Le ChIP-seq (Chromatin Immuno-Precipitation), utilise des anticorps dirigés contre un facteur de transcription ou une histone d'intérêt pour selectionner les régions du génome où la protéine se lie (PARK, 2009). Ces régions sont ensuite séquencées, ce qui permet de connaître les sites de fixation du facteur de transcription, ou la localisation de l'histone d'intérêt. Le Mnase-seq permet d'identifier les régions du génome occupées par des nucléosomes en digérant l'ADN libre (SCHONES *et al.*, 2008; VALOUEV *et al.*, 2011). Dans les approches Hi-C (Chromosome Conformation Capture), les régions du génome proches dans l'espace sont liées entre elles avant d'être séquencées, ce qui permet de déduire les interactions physiques entre différentes parties du génome (BELTON *et al.*, 2012).

Il existe aujourd'hui des techniques de séquençage dites de troisième génération (proposées notamment par Pacific Biosciences et Oxford Nanopore Technology) permettant d'obtenir un grand nombre de séquences plus longues (de l'ordre de 10 000 nucléotides), mais pour l'instant avec un taux d'erreur plus élevé qu'un NGS classique.

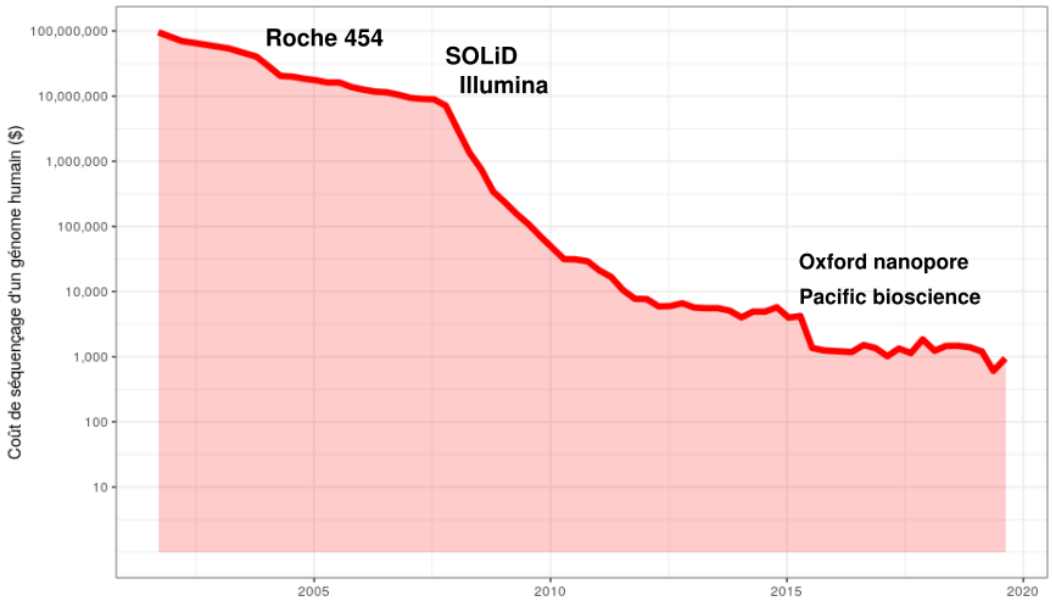

FIGURE 1.14 – **Évolution du coût de séquençage d'un génome humain.** Données : NIH. https ://www.genome.gov/about-genomics/fact-sheets/DNA-Sequencing-Costs-Data

1.4.1.2 Traitement des données de séquençage

Pour pouvoir analyser un génome séquencé en DNA-seq, des algorithmes dits d'« assemblage » comme SPAdes (BANKEVICH *et al.*, 2012) ont été développés. En effet, il faut être capable d'assembler

les différents fragments séquencés appelés « lectures » pour reconstituer la séquence initiale du génome. La problématique est similaire en RNA-seq, avec l'objectif de reconstituer les ARNm entiers et non des chromosomes ; dans ce contexte c'est le programme Trinity qui est généralement utilisé (HAAS *et al.*, 2013). Dans le cas où l'on travaille sur une espèce avec un génome déjà connu et assemblé, il peut être intéressant d'aligner les lectures de DNA-seq directement sur ce génome de référence pour y trouver des différences ;cela est généralement fait avec Bowtie2 (LANGMEAD et SALZBERG, 2012). Dans le cas du RNA-seq, un alignement sur le génome de référence avec HiSat2 (KIM *et al.*, 2019) permet de localiser les transcrits sur le génome en tenant compte de la présence d'introns. La nature répétée de la majorité des ET constitue une vraie difficulté pour ces algorithmes. L'assemblage est rendu complexe par les répétitions et crée des régions mal résolues (**Fig. 1.15.**). Dans le cas de l'alignement sur un génome de référence, les lectures issues d'ET peuvent ainsi s'aligner à plusieurs endroits, il est donc difficile de connaître leur origine exacte. Les algorithmes permettant de travailler sur les ET comme TETools (LERAT *et al.*, 2016), TEtranscripts (JIN *et al.*, 2015) ou SQuIRE (YANG *et al.*, 2019) doivent être en mesure de prendre en compte ces biais ; cela sera détaillé dans la partie suivante.

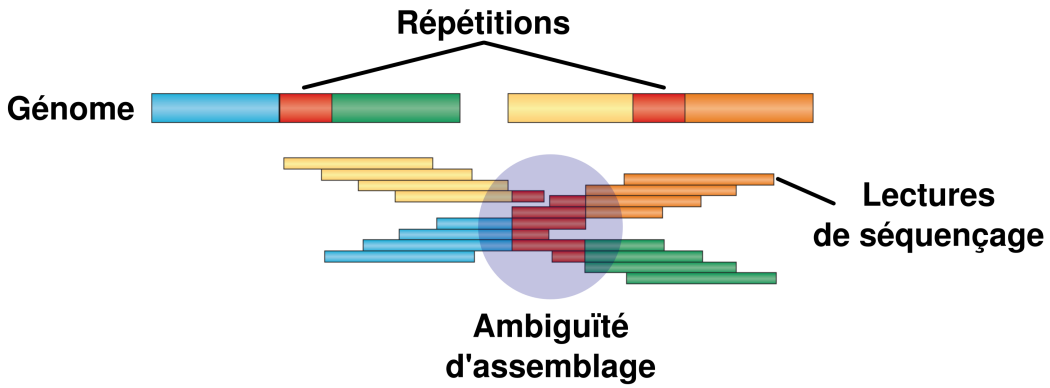

FIGURE 1.15 – **Difficulté d'assemblage des lectures de séquençage due à la présence d'une répétition.** Le génome contient une répétition (en rouge). Après séquençage et assemblage des lectures, il n'est pas possible de savoir si la région bleue est adjacente à la région verte ou à la région orange. Pour résoudre cette ambiguïté, il faudrait des lectures de séquençage plus grandes que la répétition. Adapté de Chaisson *et al.* (CHAISSON *et al.*, 2015).

1.4.2 Analyse des éléments transposables dans les génomes

1.4.2.1 Après séquençage du génome d'une nouvelle espèce

Lors d'un premier séquençage NGS, généralement sur des espèces non-modèles dont le génome n'est pas séquencé, certains algorithmes permettent d'avoir une idée rapide du contenu en ET d'un génome. C'est le cas de DNApipeTE (GOUBERT *et al.*, 2015) qui utilise un échantillon des lectures de séquençage pour réaliser un assemblage. Cet assemblage ne pourra reconstruire que les parties du génome qui sont répétées, c'est-à-dire les ET en grand nombre de copies et certains gènes en multiples copies comme les gènes d'ARNt. Le reste du génome, souvent en copie unique ou en faible nombre, demande une couverture de séquençage plus élevée pour être assemblé par les méthodes classiques. Par la suite, les ET ainsi reconstruits sont classifiés comme décrit au chapitre 2, et il est possible d'avoir une bonne approximation de la part du génome couverte par des ET

répétés. En revanche, cette méthode ne permet pas de connaître la localisation précise de chaque copie puisque le génome total n'est pas encore assemblé.

1.4.2.2 Analyse d'un génome déjà assemblé

Génération de la banque d'éléments transposables

Pour annoter les ET d'un génome, c'est-à-dire identifier la localisation génomique de chaque copie ainsi que sa famille, les différents algorithmes travaillent généralement en deux temps. La première étape consiste à former une banque contenant les séquences consensus des différentes familles d'ET présentes dans le génome analysé. Cette banque est ensuite utilisée pour détecter les différentes copies d'ET dans le génome par similarité. Il est fortement conseillé de constituer une banque spécifique de l'espèce étudiée, plutôt que d'utiliser une banque rassemblant différents éléments identifiés dans d'autres espèces. Il peut en effet y avoir des biais à utiliser la banque d'une autre espèce, proche ou non, comme le risque de ne pas détecter des ET spécifiques de l'espèce considérée, ou celui de détecter des séquences qui ne sont pas des ET chez l'espèce considérée (PLATT *et al.*, 2016). Différents algorithmes existent aujourd'hui pour constituer cette banque à partir d'un génome. Les deux programmes les plus utilisés par la communauté sont RepeatModeler2 et REPET/Tedenovo. Ces deux approches regroupent plusieurs algorithmes permettant de détecter les ET de différentes façons.

RepeatModeler2 (FLYNN *et al.*, 2020) utilise en combinaison RepeatScout (PRICE *et al.*, 2005), pour identifier les familles d'ET récentes qui sont les plus abondantes, et RECON (BAO et EDDY, 2002), qui est meilleur pour la détection de familles d'ET plus anciennes. Ces deux programmes se basent sur la nature répétée des ET pour les détecter. La détection des ET à LTR est quant à elle basée sur leur structure, et réalisée grâce aux programmes LTR_finder (XU et WANG, 2007) et LTR_retriever (OU et JIANG, 2018). Enfin, RepeatModeler2 élimine la redondance entre les prédictions obtenues par les différentes approches et classifie les consensus trouvés en utilisant des banques d'ET déjà annotés. Les nouveaux éléments ou ceux qui ne sont pas reconnus ne sont pas classifiés et sont annotés comme inconnus.

REPET3.0 (FLUTRE *et al.*, 2011; QUESNEVILLE *et al.*, 2005) a été développé pour générer une banque d'ET et s'en servir pour annoter chaque copie d'ET dans un génome. La première partie de REPET, nommée TEdenovo, permet de générer la banque. TEdenovo commence par aligner le génome avec lui même en utilisant BLASTER. Les différentes homologies trouvées sont ensuite regroupées en utilisant GROUPER (QUESNEVILLE *et al.*, 2003), PILER (EDGAR et MYERS, 2005) et RECON (BAO et EDDY, 2002). Enfin, comme pour ReapeatModeler2, une étape permet d'éliminer la redondance et de classifier les consensus détectés. Selon les auteurs de RepeatModeler2, REPET trouve plus de consensus d'ET, mais RepeatModeler2 produit plus de consensus exacts. Ils suggèrent que les deux approches peuvent être complémentaires selon le but visé par l'analyse.

Annotation du génome à partir d'une banque d'éléments transposables

L'approche la plus classiquement utilisée pour annoter les ET dans un génome est le programme RepeatMasker. A partir d'une banque de consensus de familles d'ET, le programme annote chaque copie du génome par similarité de séquences. La suite REPET dispose de son propre programme d'annotation appelé TEannot. Avec des contrôles supplémentaires, TEannot réduit le nombre de

faux positifs et regroupe les fragments d'une même copie qui peuvent être séparés dans le génome à cause d'insertions ultérieures dans l'élément.

Détection de polymorphismes d'insertion

Entre différentes populations ou individus d'une même espèce, certains ET peuvent présenter des insertions spécifiques. Il est parfois intéressant, particulièrement en génomique des populations, de pouvoir associer certaines insertions à différents phénotypes. Pour identifier les insertions spécifiques d'une population ou d'un individu, il faut séquencer de nouvelles populations ou de nouveaux individus et identifier les différences avec le génome de référence. Il existe différents programmes capables de détecter les insertions d'ET polymorphes sans assembler le génome des nouveaux individus séquencés (PopoolationTE2 (KOFLER *et al.*, 2016), Tlex3 (BOGAERTS-MÁRQUEZ *et al.*, 2019)). L'idée est par exemple de rechercher des lectures qui s'alignent en deux parties, de part et d'autre d'insertions présentes dans le génome de référence. Ces lectures attestent de l'absence de l'insertion dans le nouvel individu séquencé.

1.4.3 Étudier les éléments transposables dans les transcriptomes

1.4.3.1 Les données transcriptomiques...

Le séquençage ARN de nouvelle génération (RNA-seq) possède un double avantage. Il permet d'obtenir des informations en terme de niveau d'expression des gènes, mais aussi en terme de séquences nucléotidiques. Pour être séquencés, les ARN poly-A sont d'abord sélectionnés comme vu précédemment. Ensuite, les ARN sont fractionnés et rétro-transcrits en ADN complémentaire. Des amorces aléatoires de quelques nucléotides sont utilisées pour rétro-transcrire toutes les séquences d'ARNm fractionnées. Les ADNc ainsi obtenus sont ligués à des amorces qui seront utilisées pour initialiser la réaction de séquençage. Cette méthode permet de séquencer tous les ARNm d'un tissu, et donne des informations en terme de séquence, puisque chaque ARNm est séquencé. Elle donne aussi des informations en terme de quantité, puisqu'un ARNm fortement exprimé sera séquencé un plus grand nombre de fois. Il faut garder en tête que les ARNm ne sont pas séquencés sur la totalité de leur longueur, mais par fragments, ce qui demande une reconstruction bioinformatique des transcrits séquencés.

1.4.3.2 ... pour l'analyse d'expression des gènes...

Chez une espèce modèle, dont le génome est séquencé et annoté, chaque lecture de séquençage (les fragments rétro-transcrits et séquencés) est alignée sur le génome de référence ou sur une banque contenant l'ensemble des transcrits exprimés dans l'espèce. Cet alignement doit être fait en utilisant des outils bioinformatiques adaptés, comme Hisat2 (KIM *et al.*, 2019) qui permet d'aligner des lectures de RNAseq sur un génome en autorisant les « trous » liés à la présence d'introns. Pour aligner des lectures de RNAseq sur une banque de transcrits, bowtie2 (LANGMEAD et SALZBERG, 2012) est recommandé, car la gestion des « trous » n'est pas nécessaire. Des outils permettent ensuite de compter le nombre de lectures qui s'alignent sur les différents gènes. Il reste à comparer, pour un gène donné, le nombre de lectures qui s' alignent dessus entre deux conditions, afin de voir s'il est plus exprimé dans une condition que dans une autre. Des outils comme DESeq2 (LOVE

et al., 2014) permettent de faire cette comparaison en appliquant des normalisations et des tests statistiques pour identifier les gènes différentiellement exprimés.

1.4.3.3 ... ou l'expression des éléments transposables.

Les ET, aussi bien ceux de classe I que ceux de classe II qui possèdent des signaux de polyadénylation, expriment des ARNm poly-adénylés qui sont donc bien séquencés avec les protocoles classiques de RNA-seq (DEININGER et BATZER, 2002; LANCIANO et CRISTOFARI, 2020; LERAT *et al.*, 2016). Certains types d'ET cependant ne possèdent pas ces signaux, comme les *Alu* ou les MITE (LANCIANO et CRISTOFARI, 2020), et ne sont donc pas séquencés par les protocoles classiques de RNA-seq. En outre, il n'est pas possible d'utiliser les mêmes outils pour analyser l'expression des gènes et celle des ET. Des problématiques supplémentaires viennent s'ajouter, dues aux caractéristiques des ET (**Fig. 1.16.**).

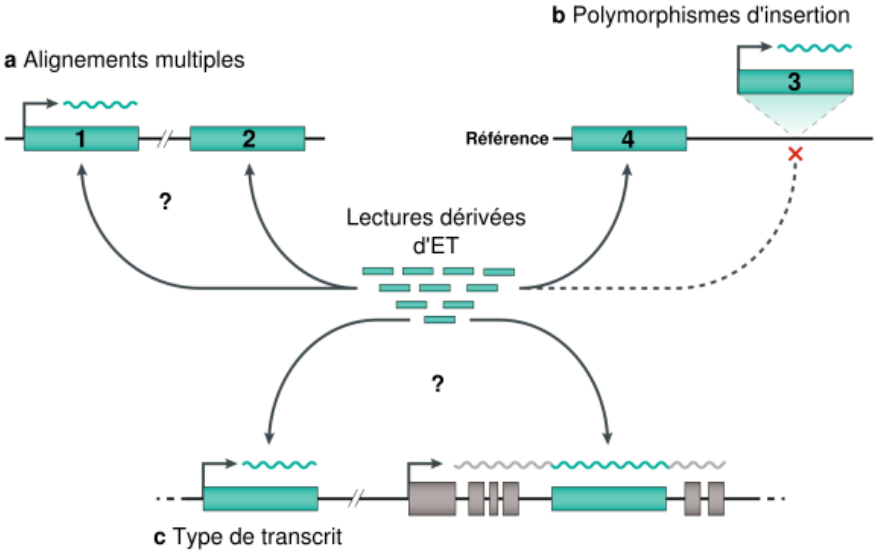

FIGURE 1.16 – **Les ET posent des problématiques supplémentaires lorsqu'on étudie leur expression.** a. Les copies récentes d'ET sont très similaires. Si seule la copie 1 est exprimée, les lectures de séquençage pourront s'aligner aussi bien sur les copies 1 et 2. Ce problème est particulièrement marqué pour les copies récentes qui n'ont pas encore divergé. b. En cas de polymorphisme d'insertion entre le génome de référence utilisé pour travailler et l'individu séquencé, une copie exprimée peut être absente de la référence (copie 3). Dans ce cas une copie similaire (copie 4) bien présente dans la référence va être artificiellement considérée comme exprimée. c. Il est difficile de faire la différence entre un ET autonome exprimé par son promoteur interne, et un ARNm d'un gène qui contient un ET. Dans le cas où un transcrit chimérique d'un gène contient un ET, considère-t-on l'ET comme exprimé? Adapté de Lanciano et Cristofari (LANCIANO et CRISTOFARI, 2020).

La nature répétée des ET entraîne d'abord un alignement multiple des lectures issues d'ARN exprimés issus d'ET, souvent sur différentes copies d'ET d'une même famille (**Fig. 1.16.a.**). Cette problématique peut aussi exister pour des gènes issus de familles multigéniques; dans ce cas la solution généralement utilisée est de simplement éliminer les lectures qui ont cette propriété. Une faible quantité d'information est perdue sachant que peu de gènes sont suffisamment similaires pour entraîner des alignements multiples. En ce qui concerne les ET, cette solution n'est pas envisageable car la majorité des lectures serait éliminée. De plus, les lectures qui s'alignent de manière unique s'aligneraient favorablement sur les copies d'ET anciennes qui ont accumulé des mutations et divergé; entraînant une surestimation de l'expression de celles-ci. Une autre

problématique est l'existence de polymorphismes d'insertion d'ET, entre différentes populations ou différents individus appartenant à une même espèce (**Fig. 1.16.b.**). Des copies exprimées et récentes présentes chez l'individu séquencé peuvent être absentes du génome de référence de l'espèce. Les lectures de séquençage seront donc attribuées à tort à une autre copie présente dans le génome de référence. Enfin, il est difficile de dissocier un transcrit produit par un ET autonome possédant son propre promoteur, d'un transcrit issu d'un gène contenant un ET (**Fig. 1.16.c.**).

Différents outils ont été développés avec des approches différentes pour répondre à ces problématiques. La première façon de procéder est de quantifier le nombre de lectures de séquençage par famille d'ET. La plupart des alignements multiples se faisant sur des copies d'une même famille, il est possible de s'en affranchir. C'est le cas des outils tels que TETools (LERAT *et al.*, 2016) ou TEtranscripts (JIN et HAMMELL, 2018; JIN *et al.*, 2015) (**Fig. 1.17.a.**). TETools a l'avantage de ne pas nécessiter la disponibilité du génome de référence d'une espèce pour être utilisé, mais seulement d'un pseudo-génome composé des copies d'ET sur lesquelles vont être alignées les données de séquençage. L'inconvénient est que des séquences exprimées qui ne sont pas dérivées d'ET pourront subir un alignement forcé sur des séquences non-homologues.

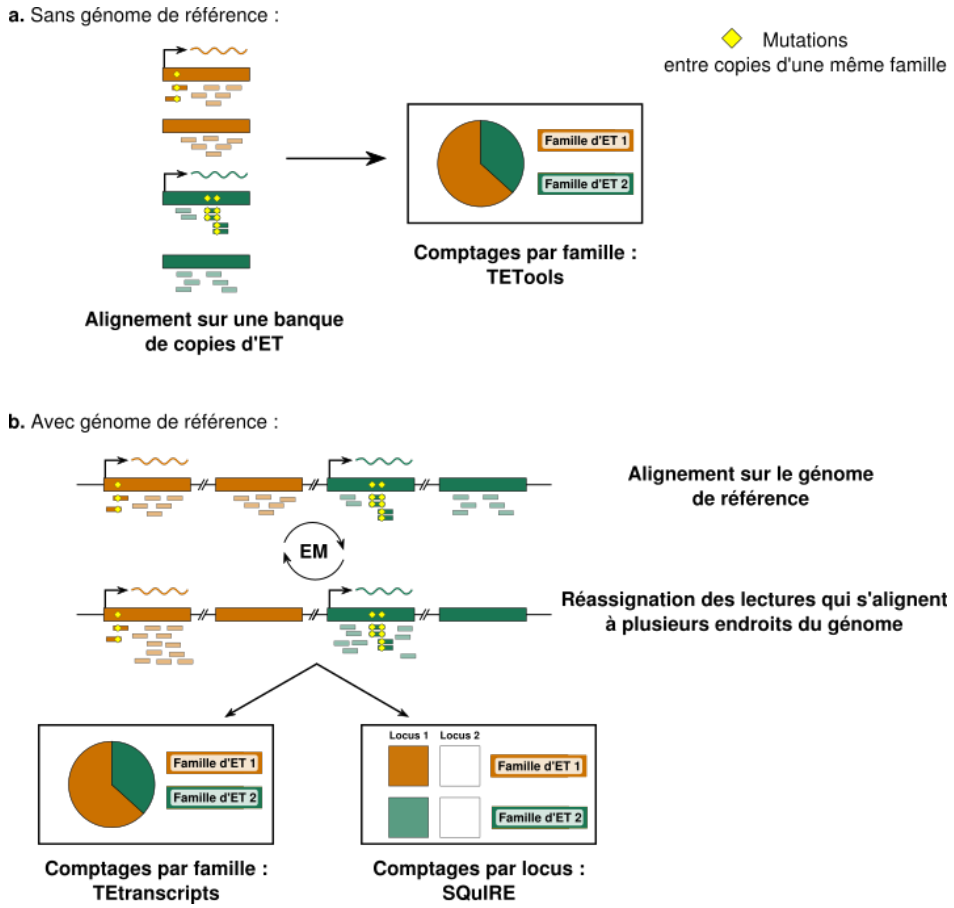

FIGURE 1.17 – **Principe des méthodes permettant d'estimer l'expression des ET dans les données de RNA-seq.** a. TETools n'utilise pas de génome de référence ; il compte les lectures qui s'alignent sur les copies pour faire le total par famille. b. TEtranscripts et SQuIRE alignent les lectures sur un génome de référence et utilisent un algorithme EM pour positionner les lectures qui s'alignent de manière multiple dans le génome. SQuIRE estime ensuite l'expression par copie, alors que TEtranscripts compte par famille. Adapté de Lanciano et Cristofari (LANCIANO et CRISTOFARI, 2020).

A l'inverse, TEtranscripts (**Fig. 1.17.b.**) nécessite un génome de référence, mais ne présente

pas le biais cité précédemment. Ces outils permettent de détecter des familles d'ET plus ou moins actives dans une condition, mais ne permettent pas en revanche de mesurer le niveau d'expression de chaque copie d'ET prise indépendamment. Même si seule une copie d'une famille est exprimée, ces outils vont moyenner le niveau d'expression sur l'ensemble de la famille. Certaines approches comme TEcandidates (VALDEBENITO-MATURANA et RIADI, 2018) partent d'ailleurs de l'a priori fort selon lequel seul un nombre très réduit de copies sont exprimées au sein de chaque famille. TEcandidates élimine ainsi la majorité des copies d'une famille afin de limiter les alignements multiples possibles et de pouvoir traiter les ET comme des gènes. Enfin, certains outils permettent d'estimer l'expression de chacune des copies d'ET. C'est notamment le cas de SQuIRE (YANG *et al.*, 2019) (**Fig. 1.17.b.**), qui utilise plusieurs itérations d'algorithmes EM (Expectation Maximisation) pour réattribuer les lectures ambiguës (celles qui peuvent s'aligner sur plusieurs insertions) à leur copie d'origine en se basant sur les lectures alignées de manière unique sur le génome. Quel que soit l'outil utilisé, il est primordial d'analyser l'expression des ET indépendamment de l'expression des gènes, car les biais gérés par les différents outils sont très différents. Quand des conclusions sont tirées en utilisant l'une de ces approches, il est important de garder en tête et de discuter les biais qui lui sont inhérents et qui ne peuvent pour l'instant pas être évités. Dans le futur, l'émergence des technologies de séquençage avec lectures longues améliorera l'estimation de l'expression des ET en séquençant les transcrits sur toute leur longueur, augmentant la probabilité pour chaque lecture de contenir un polymorphisme diagnostique, ce qui limitera le problème des alignements multiples.

1.5 Objectifs de la thèse

Le projet de ma thèse était d'étudier l'impact éventuel des ET sur les réseaux de régulation de gènes à évolution rapide. Les poissons téléostéens sont un excellent modèle pour essayer d'apporter des réponses à cette question. Comme détaillé précédemment, le développement sexuel est régulé par un ensemble de gènes. Chez les poissons téléostéens, ces gènes ont la particularité d'évoluer rapidement, aussi bien pour les gènes maîtres du déterminisme que pour certains gènes de la cascade d'activation. De fait, le sexe chez les poissons fournit un excellent exemple de réseau de régulation de gènes à évolution rapide. De plus, comme détaillé dans le chapitre 2, les poissons téléostéens présentent une grande diversité de familles d'ET par rapport aux autres vertébrés comme les mammifères ou les oiseaux. Des études de plus en plus nombreuses montrent que les ET sont impliqués dans le contrôle de l'expression des gènes et le « recâblage » des réseaux de régulation de gènes, par différents mécanismes détaillés dans le chapitre 3. C'est pourquoi le modèle des poissons téléostéens est idéal pour tester l'hypothèse selon laquelle les ET joueraient un rôle important sur les réseaux de régulation de gènes à évolution rapide.

Le médaka *O. latipes* est une espèce modèle de poisson téléostéen, même s'il est un peu moins étudié que le poisson-zèbre. Il est particulièrement utilisé dans des études d'écotoxicologie sur les espèces aquatiques, ce qui a conduit à une bonne connaissance de sa physiologie, de son génome, et de son développement sexuel. La variabilité en terme de déterminisme sexuel entre espèce proches d'*Oryzias*, couplée au déterminisme du sexe lié à un ET chez *O. latipes*, en fait un modèle idéal pour répondre à la problématique proposée.

Les génomes de cette espèce et de quatre espèces proches (*O. luzonensis, O. curvinotus, O. javanicus* et *X. maculatus*) sont disponibles. L'équipe ayant établi précédemment une annotation des ET d'*O. latipes*, le transcriptome des gonades mâles et femelles d'*O. latipes* a été séquencé ainsi que celui des 4 autres espèces. L'ensemble de ces données m'a permis d'estimer l'expression des gènes et des ET dans ces deux tissus.

Afin d'apporter des éléments de réponses à la problématique, j'ai d'abord recherché des régions du génome enrichies en gènes et ET sexe-biaisés à la fois en utilisant des méthodes existantes et en développant une nouvelle approche. Je me suis ainsi intéressé à la colocalisation dans le génome des gènes sexe-biaisés et des ET. Les résultats obtenus ont abouti à l'écriture d'un article présenté dans la première partie. Dans un second temps, je décris une approche permettant de détecter des familles précises d'ET candidates pour réguler l'expression des gènes. Certaines familles qui ont retenu notre attention sont décrites plus en détail.

Ces deux parties abordent le problème à différentes échelles. La première partie adopte un point de vue global à l'échelle du génome ; alors que la seconde se concentre sur des exemples précis de familles d'ET candidates qui pourraient réguler l'expression des gènes impliqués dans le développement sexuel.

2 Clustering of sex-biased genes and transposable elements in the genome of the medaka fish *O. latipes*

Corentin Dechaud[1], Sho Miyake[1], Anabel Martinez-Bengochea[2], Manfred Schartl[2,3], Jean-Nicolas Volff[1], Magali Naville[1].

Corresponding author : Magali Naville - magali.naville@ens-lyon.fr

Affiliations :

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

2 : Entwicklungsbiochemie, Biozentrum, Universität Würzburg, Würzburg, Germany.

3 : The Xiphophorus Genetic Stock Center, Department of Chemistry and Biochemistry, Texas State University, San Marcos, TX, USA.

Sommaire

2.1 Avant-propos

L'analyse de données transcriptomiques de gonades de médaka *O. latipes* adultes ont permis d'estimer l'expression des gène et des éléments transposables à l'échelle génomique. Les résultats présentés ici sous la forme d'un article ont été soumis dans le journal Molecular Biology and Evolution https://academic.oup.com/mbe/. Ils montrent une organisation des gènes et des éléments transposables en fonction de leur expression dans les gonades mâles et femelles. Cette version non-définitive est présentée après une première relecture par les pairs. Ce travail a été réalisé en partie avec la collaboration de Sho Miyake pendant son stage de master 1.

2.2 Abstract

While genes with similar expression patterns are sometimes found in the same genomic regions, almost nothing is known on the relative organization in genomes of genes and transposable elements, which might influence each other at the regulatory level. In this study, we used transcriptomic data from male and female gonads of the Japanese medaka *Oryzias latipes* to define sexually biased genes and transposable elements and analyze their relative genomic localization. We identified 20,588 genes expressed in the adult gonads of *O. latipes*. Around 39% of these genes are differentially expressed between male and female gonads. We further analyzed the expression of transposable elements using the program SQuIRE and showed that more TE copies are overexpressed in testis than in ovaries (36% vs. 10%, respectively). We then developed a method to detect genomic regions enriched in testis or ovary-biased genes. This revealed that sex-biased genes and TEs are not randomly distributed in the genome and a part of them form clusters with the same expression bias. We also found a correlation of expression between TE copies and their closest genes, which increases with decreasing intervening distance. Such a genomic organization suggests either that TEs hijack the regulatory sequences of neighboring sexual genes, allowing their expression in germ line cells and consequently new insertions to be transmitted to the next generation, or that TEs are involved in the regulation of sexual genes, and might therefore through their mobility participate in the rewiring of sex regulatory networks.

2.3 Introduction

With 26,000 species (NELSON *et al.*, 2006), teleost fish form the largest group of extant vertebrates. They present a high diversity of morphology, physiology and behavior, this diversity also affecting their sexual development and function (KOBAYASHI *et al.*, 2013; VOLFF *et al.*, 2007). Many sexual modes exist in fish, from hermaphroditism, where one individual can have both sexes, either simultaneously or sequentially, to gonochorism (individuals are either male or female). Sex determination, corresponding to the process by which the future sex is decided in gonochoristic species, is also diverse in teleosts. Several systems exist ranging from environmental sex determination (ESD, where sex can be determined by water temperature for example) to genetic sex determination (GSD, where sex is controlled by a particular gene or a particular set of genes) with or without sex chromosomes (BACHTROG *et al.*, 2014). In some fish species, both environmental and genetic sex determination occur and interact. In the medaka fish *Oryzias latipes* for instance, where sex is controlled by an XY chromosome system, high temperatures trigger female-to-male sex reversal (ADOLFI *et al.*, 2019). In mammals, the mechanism of GSD is highly conserved and ancient (180 – 210 Mya) (WATERS *et al.*, 2007), with almost all mammals using *Sry* as the male master sex-determining gene present on the Y chromosome. In contrast, in fish with GSD, different master sex determination genes can exist in different species, and in most species the master gene is still unknown (KIKUCHI et HAMAGUCHI, 2013). In *O. latipes* and in its related species *O. curvinotus*, sex is controlled by the dmrt1by gene located on the Y chromosome. The *dmrt1by* gene appeared 10 Mya in the common ancestor of *O. curvinotus* and *O. latipes* and was subsequently lost in *O. luzonensis*, the sister species of *O. curvinotus* (KIKUCHI et HAMAGUCHI, 2013). In O. luzonensis, the master sex-determining gene is *Gsdf* coding for the gonadal soma-derived factor. This gene is located on both X and Y chromosomes but in two different allelic forms, with the *GsdfY* allele triggering male differentiation. In *Oryzias dancena*, the master sex-determining gene is *Sox3Y* that evolved from *Sox3* as *Sry* did in mammals (HERPIN et SCHARTL, 2015; TAKEHANA *et al.*, 2014). The *Oryzias* group thus illustrates the high variability of master sex-determining genes that can control GSD in fish.

Sexual genes can be involved in sex determination *(i.e.* when sex is defined), but also in sexual differentiation (*i.e.* when the undifferentiated gonads become testes or ovaries) or sexual function. A way to detect such genes is to analyze gene expression between males and females (usually in the gonads, but not only; (GRATH et PARSCH, 2016)) and to retrieve differentially expressed genes, called sex-biased genes, *i.e.* genes more expressed in males than in females or vice-versa. An evolutionary conserved feature of sex-biased genes is their fast evolution, due to stronger positive selection, as observed in Drosophila (ASSIS *et al.*, 2012), *C. elegans* (CUTTER et WARD, 2005), fish (YANG *et al.*, 2016), and primates (KHAITOVICH *et al.*, 2005), showing that this trend is conserved throughout evolution. Moreover, sex-biased genes often appear not randomly distributed on chromosomes. In mice and flies, female-biased genes are preferentially located on the X chromosome (MEISEL *et al.*, 2020). In Drosophila and mouse, testis-biased genes tend to co-localize in the genome and form clusters (BOUTANAEV *et al.*, 2002; DORUS *et al.*, 2006; LI *et al.*, 2005).

Teleost fish genomes also harbor a large diversity of transposable element (TE) families compared to other vertebrates, particularly birds and mammals (CHALOPIN *et al.*, 2015). TEs are

sequences able to insert in the genome. They are often repeated and found in the genome of all eukaryotes analyzed to date. The high TE diversity observed in fish constitutes an important source of potential regulatory motifs for host genes. Indeed, if the majority of TE insertions are neutral or deleterious for the host, some can also be selected for adaptive functions. Several examples have been described of TEs with major roles in the rewiring of gene regulatory networks (CHUONG *et al.*, 2016; FESCHOTTE, 2008; LYNCH *et al.*, 2011; REBOLLO *et al.*, 2012b; SORRELLS et JOHNSON, 2015) some of which being related to sex (DECHAUD *et al.*, 2019; ELLISON et BACHTROG, 2013, 2015; HERPIN *et al.*, 2010). Interestingly, the expression of *dmrt1by*, the master sex-determining gene of *O. latipes*, is partly controlled by a regulatory sequence carried by a TE called *Izanagi* (HERPIN *et al.*, 2010). This TE-derived enhancer allows the tightly regulated expression of *dmrt1by* limited to a short period of time before hatching, when sex determination occurs.

TEs are not randomly distributed in the genome. Patterns of TE insertions result from insertion preferences, selection and genetic drift (BOURQUE *et al.*, 2018). Since only TEs that insert in germline cells can be fixed in the genome to be transmitted to the next generation, these patterns could be particularly influenced by the structure of the chromatin in these cells, and thus related to gene expression. Conversely, TEs could bring regulatory elements with them and modify the expression of neighboring genes, participating in the evolution of regulatory networks in germ cells but also in gonads in general. To disentangle these potential functional regulatory links between sexual genes and TEs, we investigate here the localization of sex-biased TEs with respect to sex-biased genes. As sexual development is particularly diverse in *Oryzias* and TEs are highly diverse in teleost fish, we decided to use *O. latipes* as a model of study. We generated RNA-seq data from male and female adult gonads of *O. latipes*, and identified genes and TEs with sex-biased expression. While gene expression in genome is equivalently biased between males and females, TEs globally present a clear male-biased expression in gonads. We show that the closer genes and TEs are, the more similar is their expression bias. Additionally, TEs located in sex-biased gene clusters tend to follow the cluster expression bias. Finally, we find that some male-biased TE families are enriched in male-biased gene clusters. These families constitute good candidates for TEs potentially involved in sexual gene regulation and its variability. Altogether, our study constitutes a first step towards a better understanding of the mutual regulatory influence between genes and TEs in the gonads.

2.4 Results

2.4.1 Identification of gonadal sex-biased genes

We first identified sex-biased genes by sequencing the transcriptome of three testis and three ovary replicates of *O. latipes*. These gonadal tissues are composed of both germline and somatic cells, as the two populations cannot be simply separated by dissection. However, as we are interested here in the sexual function in general, and not only in germ cells, this is not limiting for our study. In addition, this did not prevent to identify sex-biased germ cell genes, for instance genes expressed in spermatogenesis (see below). Since gonads are specialized tissues that were not used to construct the reference genome annotation, this annotation (25,167 protein coding and non-coding genes) could lack some transcripts expressed in our data. To take this into account, we decided to apply the "*new tuxedo*" approach (PERTEA *et al.*, 2016), which is based on the reconstruction of a transcript

annotation from the coordinates of the read alignments on the genome and the comparison of this new annotation with the reference from refseq (NCBI reference ASM223467v1; see methods). We detected 45,444 expressed transcripts corresponding to 27,096 genes according to the pipeline. These gene models contain protein coding genes, non-coding genes, miss-assembled transcripts due to bioinformatic predictions, transposable elements, or any other type of expressed polyA RNA. We filtered these gene models to generate a set of coding-genes and a set of non-coding genes. Among the 17,254 coding genes detected, 16,586 were already present in the refseq annotation (96.1%), and among the 3,334 non-coding genes detected, 1,845 were already present in the refseq annotation (55.3%).

In the whole, 40% of coding genes were found to be sex-biased, including 3,600 (20.9%) genes over-expressed in testis compared to ovary, and 3,293 (19.1%) genes over-expressed in ovary compared to testis. The remaining 10,361 genes were not differentially expressed (**supp. Fig. 2.6.**). The coding transcriptome is thus equivalently biased between testis and ovary. For what follows, we define genes differentially expressed between male and female gonads as "sex-biased genes". We assume that they could be involved in sexual differentiation, maintenance or function.

With respect to the expression of particular genes, we compared the patterns we obtained with previous studies drawn in medaka only. Indeed, recent observations suggest that sexual gene expression might greatly vary between species, which would reflect the rapid evolution of this pathway (HERPIN et SCHARTL, 2015). Some studies analyzed the expression of few genes in the gonads of *O. latipes* by RT-qPCR (HERPIN *et al.*, 2013; HORIE *et al.*, 2016; KOBAYASHI *et al.*, 2017; NAKAMOTO *et al.*, 2006). Three main genes have been studied in details: dmrt1, the ancestral paralog of *dmrt1by*, *gsdf*, the gonadal soma-derived factor, and *foxl2*, a transcription factor involved in ovarian development. Both *dmrt1* and *gsdf* are involved in testis development. *dmrt1*, along with gsdf, was always found to be overexpressed in testis compared to ovary, which corresponds to our data (**supp. Fig. 2.6.**). *foxl2* was found to be highly expressed in ovary in previous studies, and is coherently detected ovary-biased using our RNAseq data (**supp. Fig. 2.6.**). Additionally, RT-PCR experiments detected an expression of *amh* and *sox9b* in both gonads, and of aromatase in ovary only (KUROKAWA *et al.*, 2007). We observe similar patterns of expression for these genes using RNAseq data.

In the case of non-coding genes, 33.2% were found to be sex-biased, including 695 (20.8%) genes over-expressed in testis, and 412 (12.4%) genes over-expressed in ovary compared to testis. The non-coding transcriptome is thus more male-biased than the coding transcriptome, with a lower contribution of ovary-biased genes.

2.4.2 Identification of TE copies and families with sex-biased gonadal expression

To analyze TE expression relative to gene expression, we then characterized gonadal expression of TEs at the single copy level. Our annotation of TEs (**supp. data 1 and 2**) covers 34% of the medaka genome, which corresponds to a previously obtained coverage (CHALOPIN *et al.*, 2015). SQuIRE (YANG *et al.*, 2019) allows to retrieve expression of each TE locus from RNAseq data. We identified 37,108 loci as expressed TE copies (corresponding to 3.7% of all TE loci). Among them, 13,325 (35.9%) were found to be testis-biased (adjusted $p-value < 0.05$, $log_2\text{FC} < -1$), while 3,842

(10.4%) were ovary-biased (adjusted $p-value < 0.05$, $log_2\text{FC} > 1$). Therefore, whereas the same proportion of coding genes was found to be testis- or ovary-biased, TE expression appears clearly biased toward testis.

We further searched for TE families enriched in copies with testis- or ovary-biased expression. We found 22 families with global testis-biased expression ($\chi^2 - p < \frac{0.05}{n_{families}}$ and >50% of testis-biased copies), and 19 families with global ovary-biased expression ($\chi^2 - p < \frac{0.05}{n_{families}}$ and >50% of ovary-biased copies, **supp. data 3**, **supp. Fig. 2.7.**). Interestingly, the majority of the sex-biased families corresponds to class I Long Terminal Repeats (LTR) elements: ovary-biased families comprise 13/14 LTRs and 1/14 Long Interspersed Nuclear Elements (LINEs), and testis-biased families 10/16 LTRs, 4/16 DNA transposons, 1/16 LINE, and 1/16 Rolling Circles (RC). We generated a phylogeny of the LTR reverse transcriptase (RT) from family consensus sequences to confirm our annotation of the families, and to test if the expression pattern of the sequences is linked to their evolutionary relationships (**supp. Fig. 2.8**). As most of the biased families are Gypsy elements, we more specifically generated a phylogeny of expressed *Gypsy* TE copies, irrespective of being sex-biased or not (**Fig. 2.1**). We only used expressed TE copies for which we were able to detect an RT sequence. Plotting of expression patterns on the molecular phylogeny of TE copies showed that related TE copies often have similar sex-biased expression. To test if this could be explained by shared insertional preferences that would target similar TEs to similar expression environments, we analyzed sequences surrounding insertions of four subtrees (two mainly male-biased, one mainly female-biased, and one mainly non-biased; (**supp. Fig. 2.9**). The GC content was not significantly different between subtrees, and no insertion sequence specificity or preference could be detected (**supp. Fig. 2.9 and 2.10**). These observations suggest that the expression bias may be explained by the sequence of the TE itself, *i.e.* that TEs would harbor regulatory sequences shared between phylogenetically related copies that co-regulate their expression. Analyzing *Gypsy* copies of a male-biased subtree of the previous phylogeny, we observed that at least 10 of them (42%) present LTRs and complete ORFs, suggesting that they are only mildly corrupted and might correspond to recent and autonomously transcribed insertions (**supp. Fig. 2.10**). However, the phylogeny also presents some examples of closely related copies with different expression patterns. This might be explained either by the loss of regulatory sequences by some copies, and/or by the influence of regulatory sequences from neighboring genes.

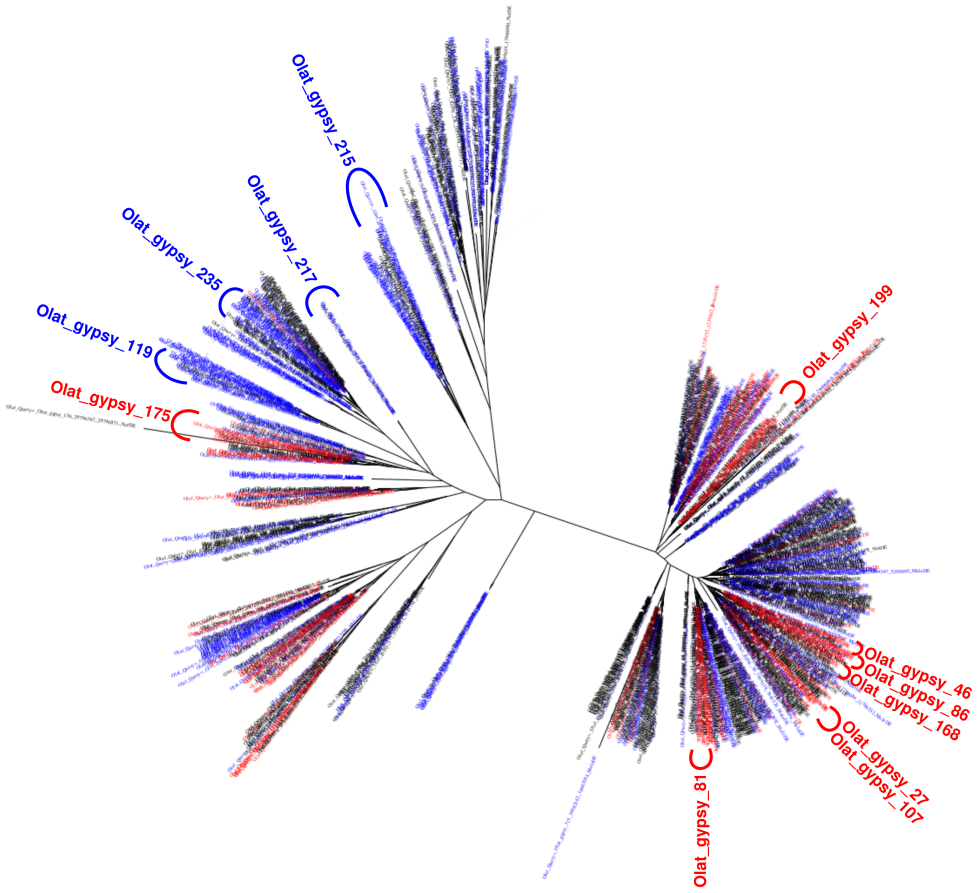

Figure 2.1 – **Phylogeny of expressed *Gypsy* TE copies of medaka using the amino acid sequence of the reverse transcriptase (RT).** Tip colors correspond to the expression bias (red: ovary; blue: testis; black: locus expressed but not biased). The location of biased Gypsy families are indicated. Note that copies with the same expression bias locally group together.

2.4.3 Sex-biased genes and TEs form clusters in the genome

To characterize the genomic distribution of sex-biased genes and TEs, we first asked if they physically clustered in the medaka genome. To unravel possible clusters of co-expressed genes, we applied a method that previously demonstrated the existence of large clusters of co-expressed genes in the Drosophila genome (BOUTANAEV *et al.*, 2002). Briefly, stretches of adjacent genes or TEs with the same expression bias are counted across the genome. A stretch stops as soon as a gene or TE is found with a different expression bias: features within a stretch all present the same bias. For example in the *O. latipes* genome, the biggest stretch of consecutive male-biased genes is 10 genes long (and spans 36kb), and a maximum of 8 consecutive female-biased genes is found (spanning 133 kb). For TEs, a maximum of 32 and 15 male and female-biased copies are found, respectively. The observed number of stretches was compared to an expected number computed from random distributions of genes in the genome (**Fig. 2.2.**, brown bars). We observed that testis- and ovary-biased genes are both not randomly distributed in the genome since they do

not follow the expected random distribution in terms of stretches of adjacent genes with similar expression bias. We found more clusters of at least 3 genes than expected if genes were randomly distributed (**Fig. 2.2.**). Moreover, for testis-biased genes we observed stretches of 9 or 10 genes, while such arrangement is never predicted among 1000 random genomes. Results are given for coding and non-coding genes grouped together in the top panel of Figure 2.2 (for coding and non-coding genes treated independently see **supp. Fig. 2.11**). We performed the same analysis with TE copies (**Fig. 2.2.**, bottom panel). We observed a non-random distribution for both testis and ovary sex-biased TEs as already found for genes. However, TE clusters contained markedly more elements than gene clusters: many comprised more than 8 TE copies, and up to 32 for male-biased TEs, while a maximum of 11 copies were predicted in a consecutive location in the random genomes. Since there are fewer ovary-biased TEs, the highest stretch sizes are lower than for male-biased TEs (15 and 32 consecutive biased TEs, respectively), but still higher than the maximum of 5 consecutive biased TEs expected at random under the null hypothesis. Globally, genes and TEs are thus not randomly distributed in the genome and tend to group into clusters with the same sex-biased expression. This is also true for coding and non-coding genes if analyzed separately (**supp. Fig. 2.11**).

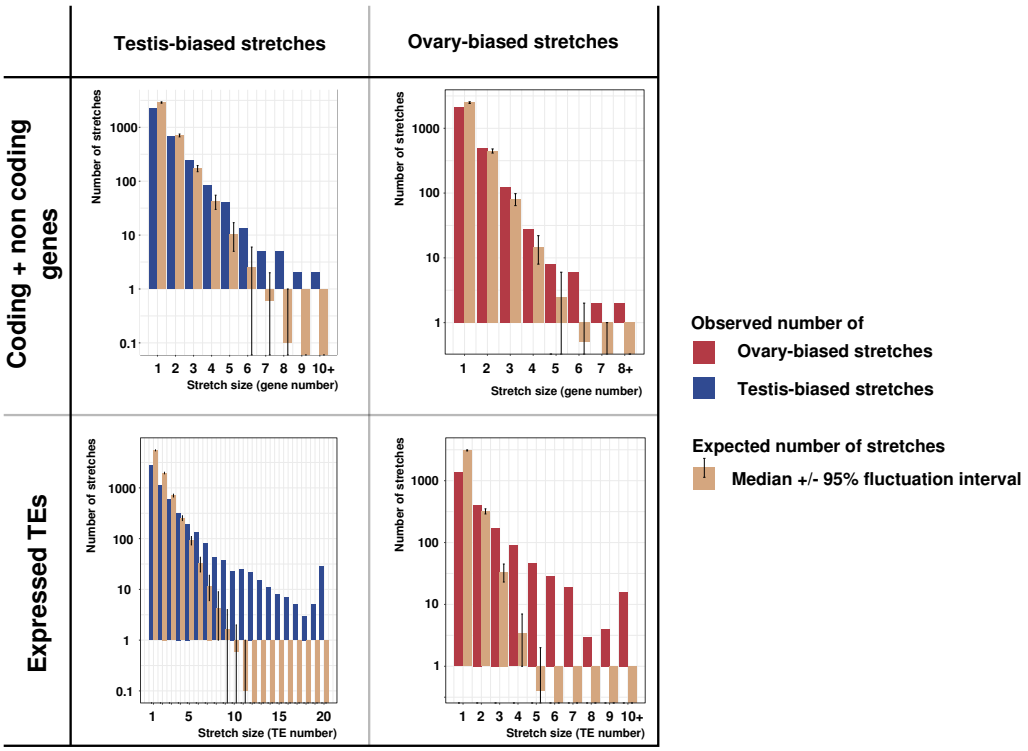

Figure 2.2 – **Stretches of consecutive sex-biased genes and TEs along the genome of *O. latipes*.** First row: stretches of sex-biased genes. Second row: stretches of sex-biased TEs, as identified by SQuIRE. The first column represents testis-biased sequences, and the second ovary-biased sequences. The observed number of stretches of a given length (in gene number) in the genome are shown in red and blue for males and females, respectively. 1000 random genomes were generated using the same proportion of male- and female-biased genes or TEs to estimate the stretch sizes if genes or TEs were randomly distributed. The median number of expected stretches is shown in brown. Error bars represent the fluctuation interval containing 95% of the values generated by the bootstraps.

This first method to identify clusters of co-expressed genes being quite inflexible, we applied a

second and complementary search to identify regions showing a high density of sex-biased genes or TEs. We used the Gene clusters method previously developed by our team (TOUBIANA *et al.*, 2020), which calculates the local mean log_2FC of the transcripts in a sliding window, and detects significantly biased regions using a bootstrap approach. We applied this to all gene transcripts, including coding and non-coding ones. We were able to identify 32 male-biased regions spread over 17 (out of 24) chromosomes (**supp. Table 2.1, Fig. 2.3.a., supp. Fig. 2.12**), and covering 3.94% of the genome. The method also uncovered 18 female-biased regions, spread over 13 chromosomes and covering 2.48% of the genome (**supp. Table 2.1, Fig. 2.3.a.**). Hence, about 6% of the genome of the medaka consists of sex-biased regions with respect to gene expression. In order to investigate their functions, we inspected manually all genes present in the sex-biased clusters in the medaka genome. We recovered genes known to be involved in male sexual functions in male-biased clusters, including *dmrt1a* – the autosomal paralog of the master sex-determining gene *dmrt1by*, *morn3*, *frizzled-4*, *ucp2* and *lrguk*, and genes with known female sexual functions in female-biased clusters, such as *zonadhesin*, *bucky ball*, *bokb* and *hsd17b1* (**supp. Table 2.2**). We also tested the association of genes in sex-biased clusters with Gene Ontology (GO) terms. At the genome-wide scale, male-biased genes were significantly associated with 'spermatogenesis', 'cilium assembly and function', and 'protein polyglutamylation' (BOBINNEC *et al.*, 1999) (**supp. Fig. 2.14**), and female-biased genes with 'acrosome reaction', 'oogenesis' and 'binding of sperm to zona pellucida' (**supp. Fig. 2.15**), indicating a global link in our data between sex-biased expression and gonadal/germ cell function. In contrast, only the term 'spermatogenesis' was significantly associated with genes in male-biased clusters (**supp. Fig. 2.16**), and no significant GO term linked to sexual function and reproduction was found for genes in female-biased clusters (**supp. Fig. 2.17**). This might indicate that genes in sex-biased clusters, compared to sex-biased genes in general, are enriched in genes with so far uncharacterized sexual functions (possibly lineage-specific and evolving more rapidly), with functions that are less sex-specific than germ cell functions, or with functions that have been more recently recruited to the gonads.

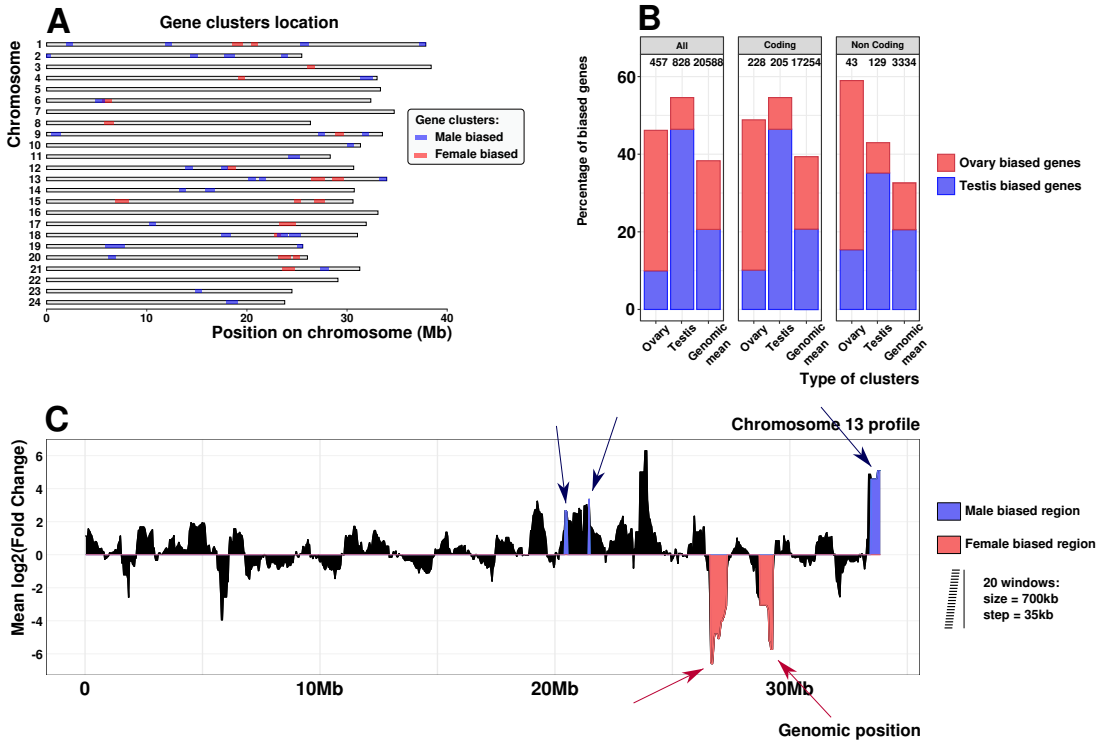

Figure 2.3 – **Sex-biased gene expression clusters in *O. latipes* genome.** A. Genomic location of gene clusters built from coding and non-coding genes. B. Percentage of testis- or ovary-biased genes in each type of clusters. C: Chromosome 13 profile obtained with Gene clusters representing the expression bias of all genes across sliding windows along the genome. The mean log2FC is represented for each window along the chromosome. The significantly male- and female-biased regions are shown in blue and red, respectively. The size of one window is shown, at scale, on the right of the plot.

We called "gene clusters" the regions presenting a significantly higher mean differential expression of genes. 810 sex-biased genes (3.9% of all genes, and 10.1% of all sex-biased genes) are located in a cluster with the same expression bias. The size and gene composition of these regions calculated using coding genes only, non-coding genes only, or both are described in supp. table 2.1. Among the 828 genes located in a testis-biased region, 353 (42.6%) are testis-biased, while 78 (9.4%) are ovary-biased (**Fig. 2.3.B.**). Ovary-biased regions contain 457 genes with 165 (36.1%) female-biased genes and 47 (10.3%) male-biased genes (**Fig. 2.3.B.**). As an example, we observe on chromosome 13 of *O. latipes* two main ovary-biased regions of 1Mb with a mean log_2FC of -5 (**Fig. 2.3.C.**). For the detailed gene cluster profiles for each chromosome see supplementary data 4-7.

In contrast to the situation observed in Drosophila (BOUTANAEV *et al.*, 2002; ELLISON et BACHTROG, 2013), we did not observe any particular trend on chromosome 1, which is the X chromosome. The Y chromosome of *O. latipes* differs from the X by 250kb that include only one gene, *dmrt1by*. This region is not represented in the reference genome we used, but present on NCBI for a different strain. However, if we look at the Gene clusters profile obtained from TE expression, and not from genes, we observe a testis-biased cluster surrounding the insertion breakpoint of the Y-specific region (**supp. data 7**, chromosome 1; **supp. Fig. 2.18**). This cluster was thus present next to the region of insertion, and the insertion seems to have occurred in a region that was already

male-biased.

2.4.4 Neighboring genes and TEs share correlated expression bias between male and female gonads

As we could determine the expression of TEs at the copy level, we next asked if neighboring genes and TEs shared similar expression patterns that would reflect a co-regulation (originating from an enhancer of the gene or from the TE). To do so, we assessed whether the sex bias in expression of adjacent genes (coding and non-coding: **Fig. 2.4.**, coding or non-coding: supp. **Fig. 2.19.**) and TEs is correlated. Only genes and TEs expressed in gonads were used to test the hypothesis. We calculated the Pearson correlation coefficient of log_2FC of gene-TE pairs with different distance categories (border to border gene-TE distance 10pb-1kb, 1-5kb, 5-50kb, and 50-500kb). A given TE copy can be associated to all genes within the distance selected, and conversely a given gene can be associated to all TEs in the distance selected. We calculated the correlation for each TE family containing at least 5 TE copies, and making at least 10 gene-TE pairs (*i.e.* 10 TEs with 1 gene, or 5 TEs with 2 genes). The correlation coefficient of each family is represented depending on the distance considered to create the gene-TE pairs (**Fig. 2.4.**). We observed that the closer TEs and genes are located, the higher the correlation coefficient is (ANOVA 1 way, $p < 1e^{-10}$). The same result was obtained using only the coding or the non-coding genes (**supp. Fig. 2.19.**).

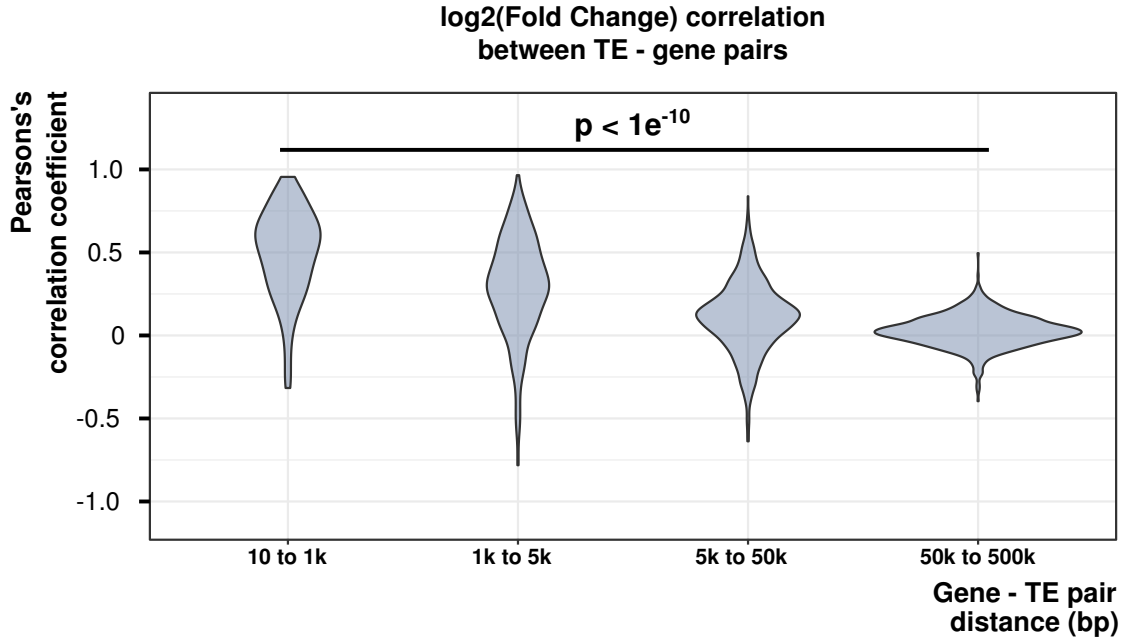

Figure 2.4 – **Correlation coefficients calculated on expression bias between adjacent genes and TE copies for different intervening distances, for each TE family.** Each violin corresponds to the distribution of correlation coefficients of the different TE families. Fewer gene-TE pairs are found for the shortest distance categories, explaining why the variance is higher for these categories. The closer TEs and genes are located, the higher the correlation coefficient of expression bias between genes and TEs.

We complementarily asked if clusters of sex-biased genes could concentrate TEs with the same sex bias in gonadal expression. No clear common pattern of gene and TE distribution emerged from the inspection of gene clusters, indicating the diversity of their structures. In some clusters hot-spots of TEs were observed (*e.g.* **supp. Fig. 2.12**), while in others TEs showed no marked local enrichment (*e.g.* **supp. Fig. 2.13**). On average, TE copies located in testis-biased regions are over-expressed in testis (mean log_2FC = 2.96), and significantly more than TEs not located in sex-biased clusters (mean log_2FC = 1.35, Student test $p < 1e^{-16}$). TEs located in female-biased regions also present on average over-expression in testis (mean log_2FC = 0.77), but at a lower level than TEs not located in sex-biased clusters (mean log_2FC = 1.35, Student test $p = 4.2e^{-6}$) (supp. Fig. 2.20). As the number of copies compared between both groups is large, a low p-value is not really informative, and it is important to check for the size of the effect (mean log_2FC 0.77 vs 2.95 vs 1.35). TEs located in ovary-biased regions are still more testis-biased than ovary-biased, which reflects the previous observation that the general trend of the genome is a testis-biased expression of TEs. 3.94% and 2.48% of the genome are considered as male and female sex-biased gene clusters, respectively. 5.84% of expressed TEs and 4.11% of all TEs (expressed or not) are located in male-biased gene clusters. 1.81% of expressed TEs and 2.40% of all TEs are located in female-biased gene clusters. Overall, the results indicate that globally TEs do not preferentially integrate into sex-biased clusters (this was also observed considering only recent insertions, to take into account the possibility that clusters change their position during time; data not shown). A slight enrichment might be observed for expressed TEs in male-biased clusters, which might reflect preferential insertion (maybe due to open chromatin in testis) and/or positive selection of insertions in these regions – but this minor effect requires further investigation to assess its significance. To double check for the presence of sex-biased TEs in sex-biased regions, we investigated if sex-biased TE copies were more likely to be inserted in regions with a similar sex-bias gene expression than somewhere else in the genome. For that, we tested the possible relationship between the localization of the copy and its expression (**supp. Fig. 2.21**). Most of the 37,038 TE copies analyzed are located in unbiased regions and present unbiased expression (supp. fig. 15, 18,884 copies). We computed the expected copy numbers if there were no association between both localization and expression (**supp. Fig. 2.21**, grey values), and tested the difference between observed and expected counts using a χ^2 test of independence ($p < 1e^{-125}$). Again, as the number of copies is very high, it is important to consider the size of the effect (**supp. Fig. 2.21**, ratio values) and not only the p-value. We observed a fold of 1.72 (1293/773) for testis-biased TE copies present in testis-biased regions, showing that there are more testis-biased TEs in testis-biased regions than expected at random. For ovary-biased copies within ovary-biased regions the enrichment is of 1.58, which means that there are more ovary-biased TEs located in ovary-biased gene clusters than expected at random. On the opposite, testis-biased regions should contain approximately 217 ovary-biased copies, but contain only 114, making these regions depleted in ovary-biased copies (fold 0.53). Conversely, only 163 testis-biased copies are found in ovary-biased gene clusters (vs 215 expected). Thus, these regions are depleted for testis-biased TEs (fold 0.76).

Finally, we tested if TE families harboring a large proportion of sex-biased copies are more likely to be located in sex-biased regions. Independently of the expression bias of their copies, and taking into account both expressed and non-expressed copies, 19 TE families were enriched in testis-biased regions, while nine families appeared enriched in ovary-biased regions (**Fig. 2.5.a.**).

On figure 2.5.b. we show the same data but coloring TE families if the expression of their copies is significantly biased in one sex. Overlapping the two graphic representation of the data (**Fig. 2.5.c.**), we observe that four of the families enriched in male-biased regions correspond to families for which copies present a strong biased expression toward males (supp. data 3). These families include a family of Helitrons, which are class II TEs involved in dosage compensation in Drosophila males (ELLISON et BACHTROG, 2013, 2015), and are preferentially localized in the sex-determining region of the platyfish Xiphophorus maculatus (ZHOU *et al.*, 2006). Another of these four families is an Unknown TE family that is highly testis-biased, with 51 out of 56 expressed copies being more expressed in testis than in ovary, the remaining five being non-biased. This family is enriched in testis-biased regions, with 16% of the copies located in such regions (15/93 copies in total, and 8/56 expressed copies). We analyzed the sequence of the copies from this family. We found two putative transcription factor binding sites enriched in expressed copies compared to non-expressed ones (**supp. Fig. 2.22**). These transcription factors are involved in male gonad development, Sertoli cells development and spermatogenesis (SOX8), and in response to testosterone and male genitalia development (HOXD13). Among the six TE families preferentially located in female-biased regions (**Fig. 2.5.a.**), there is no family significantly harboring female-expressed copies (**Fig. 2.5.c.**).

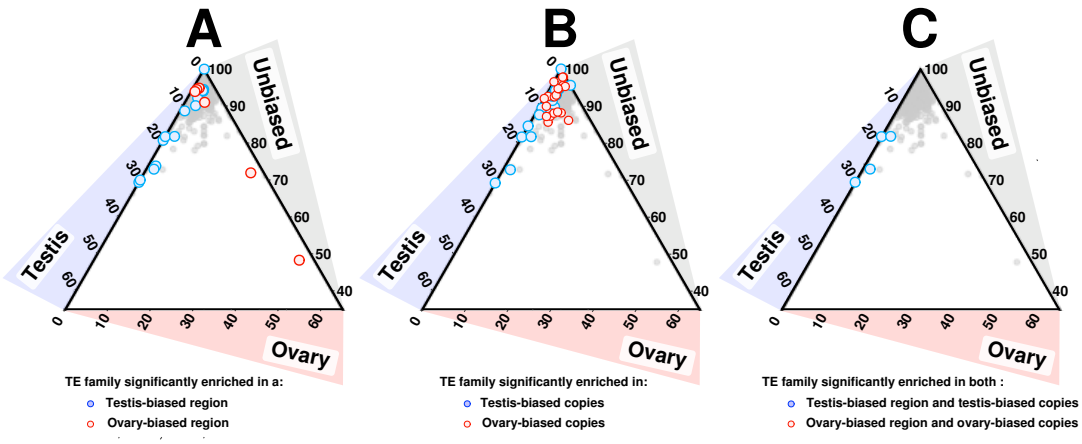

Figure 2.5 – **TE copies from specifically testis-biased TE families are preferentially located in testis-biased gene clusters.** A. Comparison of location and expression of TE families. Each data point corresponds to one TE family. The grey scale gives the percentage of copies of the family located in an unbiased genomic region; on the left the percentage of copies over-expressed in testis is indicated, and on the bottom, the percentage of copies over-expressed in the ovaries. TE families significantly enriched in the biased regions are shown in red or blue. B. The same data as in A panel but with a different coloration, TE families are highlighted in red and blue if their copies are significantly more expressed in one sex compared to the other. Families enriched in male-biased regions are often enriched in male-biased copies. In contrast, families enriched in female-biased regions are not necessarily made up of copies over-expressed in ovaries. C. Intersection between A and B. The 5 TE families with both significant testis-biased expression and testis cluster localization are shown in blue.

2.5 Discussion

2.5.1 Global gene expression analysis reveals a similar proportion of testis- and ovary-biased coding genes in the medaka

In this study we investigated gene and TE expression in the gonads of the Japanese medaka *O. latipes*. Compared to other tissues, gonads are the most sex-biased organs in terms of gene expression (BÖHNE *et al.*, 2014; TSAKOGIANNIS *et al.*, 2018b). These organs constitute thus a good model to study sex-dependent gene expression regulations. As the dissection of medaka gonads does not allow to physically separate the germline from the soma, our data are made of both types of cells and should be interpreted as a gonadal and not as a germline expression analysis. Several studies previously analyzed RNA-seq data in teleost fish, mainly from gonads (BAR *et al.*, 2016; BÖHNE *et al.*, 2014; LIU *et al.*, 2015; ROBLEDO *et al.*, 2015; TAO *et al.*, 2018; TSAKOGIANNIS *et al.*, 2018b; WANG *et al.*, 2017; ZENG *et al.*, 2016) or from brain (BEAL *et al.*, 2017; BÖHNE *et al.*,

2014; LIU *et al.*, 2015; SAARISTO *et al.*, 2017; SHEN *et al.*, 2020; TSAKOGIANNIS *et al.*, 2018b; WU *et al.*, 2019). Here we found that 40% of protein-coding genes present a sex-biased expression in medaka gonads, with 21% being testis- and 19% ovary-biased. In cichlid fish, by analyzing 4 species, 66% of the coding genes were found differentially expressed, with slightly more testis-biased genes (BÖHNE *et al.*, 2014). In the Japanese fugu, only 3.7% of the coding genes were found differentially expressed between gonads, but still with more testis-biased genes (WANG *et al.*, 2017). The proportion of gonad-biased genes thus greatly varies between fish species and probably also depends on the conditions and parameters of analysis. Many technical biases could explain these variations, since RNA-seq data analysis relies on several parameters including the type of pipeline used, the availability of a reference genome of good quality, and the thresholds applied to decide if a gene expression is sex-biased or not. The observed variations could nevertheless also result from true biological variability, notably because of differences in gonad maturity between species: medaka is a constitutive spawner, while fugu and cichlids have breeding periods and the bluehead wrass is a consecutive hermaphrodite. A common trend that persists is the slight bias towards testis over-expressed genes. However this effect is not that important in medaka, and we can consider that genes are overexpressed in similar proportion between sexes in this species.

2.5.2 A higher proportion of TEs and non-coding RNA genes are over-expressed in testis compared to ovary

As TEs are known to be particularly involved in the regulation of vertebrate non-coding RNAs (KAPUSTA *et al.*, 2013), we included non-coding RNA genes in our analyses, either merged with protein-coding genes, or as a distinct class of genes. It should be noted that, because our sets of transcripts were generated using poly-A purification and since a large part of non-coding RNAs is not poly-adenylated (CHENG, 2005; KAPRANOV *et al.*, 2010; SUN *et al.*, 2018), non-coding genes are likely to be under-represented in our dataset. While RNA genes were globally found to be slightly less differentially expressed than coding genes (33.2% vs 40%, respectively), they appeared more biased toward testis (63% of differentially expressed non-coding RNA genes are testis-biased, vs 52% of differentially expressed protein-coding genes). Some of them might correspond to lncRNAs important for spermatogenesis, as reported in tetrapods (NECSULEA *et al.*, 2014). PolyA-enriched RNA-seq data are also relevant to study TE expression (LANCIANO et CRISTOFARI, 2020; LERAT *et al.*, 2016), since both class 1 and class 2 TEs express poly-adenylated RNAs (DEININGER et BATZER, 2002). Almost half of expressed TE insertions were found differentially expressed (46.3%), with a major bias toward testis compared to ovary (78% vs 22% of biased TEs, respectively). It is expected that TEs are particularly expressed in the germline, since their spreading and fixation in the genome rely on their activity in these cells (DECHAUD *et al.*, 2019). The question may arise if the higher proportion of testis- versus ovary-biased TEs might be a consequence of the different proportions of germline and somatic cells in the male and female gonads. These proportions are not available so far from literature, and we can only roughly estimate from our own unpublished observations in medaka a ratio of soma vs germline cells of ca. 1:100 in testis and 100:1 in ovary. Male and female germ cells however do not contain the same amounts of RNA. Again, there are no data available concerning this point, but we can very roughly estimate the ratio of soma vs. germ cell RNA amount to be of 1:100 in both types of gonads. This estimate is reinforced by the similar proportions of

male- vs. female-biased gene transcripts determined in gonads, suggesting that expression levels are not much biased by their intrinsic structure. The higher proportion of testis-biased TEs could rather suggest that in medaka the male gonad is more permissive than the female gonad in term of TE expression. This idea is somehow in opposition to the observations concerning the control of TEs in fish by piRNAs, a class of non-coding RNAs molecules that induce TE mRNA degradation. Indeed, both in zebrafish (HOUWING *et al.*, 2007) and medaka (KNEITZ *et al.*, 2016), piRNAs are more expressed in testes than in ovaries. Consequently, We might expect lower levels of TE expression in testis compared to ovary. However, the steady-state TE mRNA levels we observe result from both transcription and repression. The higher TE expression measured in testis could therefore result from much higher initial transcription, somehow counterbalanced by piRNA-mediated repression. More precise investigations are needed to conclude on this point. In addition, some retrotransposon families were shown to escape methylation in testis in mammals (MOLARO et MALIK, 2016). Similar experiments are needed in fish to evaluate if differential methylation of TEs might be linked to their higher level of expression in testis compared to ovary.

2.5.3 Sex-biased genes and TEs form clusters in the medaka genome

We focused particularly on the relative localization of genes and TEs with respect to their sex-linked expression bias. Many studies already reported that sexual genes (BOUTANAEV *et al.*, 2002; DORUS *et al.*, 2006; LI *et al.*, 2005) and genes in general (LERCHER *et al.*, 2002; ROY *et al.*, 2002; SINGER *et al.*, 2005) are not randomly distributed on chromosomes. Regulatory constraints such as the presence of specific enhancers or local chromatin states that affect neighboring genes can for instance result in the conservation of syntenies, even between distant species (FERRIER et HOLLAND, 2001; KIKUTA *et al.*, 2007; PASCUAL-ANAYA *et al.*, 2013). These functional constraints are now known to translate into Topologically Associated Domains (TADs) (JORDAN ROWLEY et CORCES, 2018; DE WIT, 2019), which are regions preferentially interacting in the 3D organization of the genome. A way to define clusters is by counting successive genes presenting the expression pattern (or any other feature) of interest (BOUTANAEV *et al.*, 2002; DORUS *et al.*, 2006; LERCHER *et al.*, 2002; LI *et al.*, 2005; ROY *et al.*, 2002; SINGER *et al.*, 2005). Such a procedure is quite stringent, but can reveal candidate regions where large stretches of co-regulated genes exist. Applying this approach, we showed that the *O. latipes* genome is enriched in stretches of more than 3 sex-biased genes, this observation being valid for both coding and non-coding genes. When considering TEs, the enrichment is even stronger, with stretches of more than 20 testis-biased elements. This suggests that some regions contain a higher density of sequences with sex-biased expression.

In a second and complementary attempt, we developed a new method that evaluates the global expression bias of genes along the chromosomes, and identifies regions with a mean expression bias significantly diverging from random expectations. This method can be applied to any set of differentially expressed genes, and not only to sex-biased genes, the only requirements being the localization of the genes and the associated log_2FC of their transcripts. Applying this sliding-window approach, we detected 32 and 18 clusters enriched in genes with male- or female-biased expression, respectively. These clusters contain about 10% of all sex-biased genes. We can hypothesize that such organization is linked to a common regulation of the clustered genes. The notion of "synexpression groups" has been proposed in eukaryotes, which corresponds to sets of co-regulated

genes (HERPIN *et al.*, 2019; NIEHRS et POLLET, 1999; RAMIALISON *et al.*, 2012). In the medaka, a pilot study including 560 genes expressed during embryonic development showed that co-regulated genes tend to share particular DNA motifs in their cis-regulatory regions (RAMIALISON *et al.*, 2012). These motifs allow the genes to be tightly controlled in space and time during development. About a third of the developmental syn-expression groups presented pairs of genes distant from less than ten genes, which is rarely observed by chance (RAMIALISON *et al.*, 2012). This revealed a slight tendency of co-regulated genes to group on the chromosomes. With the finding of 50 sex-biased gene clusters encompassing 25 biased genes in the mean, our genome-wide analysis put a big step further this observation. It also demonstrates that this trend not only concerns developmental genes, but also genes functioning in mature organs. It would be interesting to compare the domains we identified with data of Hi-C, for now unavailable for medaka gonads (NAKAMURA *et al.*, 2018).

2.5.4 A cluster of sex-biased TEs could have favored the birth of sexual chromosomes

In mammals, X and Y chromosomes stopped to recombine 210 Mya (WATERS *et al.*, 2007), leading to an accumulation of TEs and a loss of genes on the Y chromosome. In *O. latipes*, the sex chromosomes are relatively young. X and Y chromosome are still recombining on their almost complete length and their single main structural difference is a short Y chromosome-specific region of 250kb containing the master sex-determining gene dmrt1by and its transcriptional regulatory region, a copy of the *Izanagi* transposon (KONDO, 2006). One theory predicts that an important early step in the evolution of the Y chromosome is the linkage of sexually antagonistic genes that are beneficial to males but not to females in the Y-specific region (CHARLESWORTH *et al.*, 2005). This leads to a loss of recombination between the X and the Y chromosomes in this region, and in consequence to the accumulation of TEs that can no longer be purged by crossing-over with the homologous region lacking the TEs (CHARLESWORTH *et al.*, 2005). We found on the X chromosome a testis-biased cluster of TEs surrounding the insertion breakpoint of the Y-specific region, suggesting that the future Y chromosome of *O. latipes* already accumulated male-biased TEs in this region before the new master sex-determining gene was inserted. It is thus tempting to speculate that this region enriched in sex-biased TEs could have eased the recruitment and evolution of the new male master-sex determining gene *dmrt1by* of *O. latipes* by providing a favorable male transcriptional environment.

2.5.5 Disentangling the possible functional links between TEs and sexual genes

It has been already established that TEs are generally not randomly distributed in the genome. Some retrotransposons for instance are able to target regions (either precise nucleotide sequences or larger particular chromatin environments) where they can insert without generating deleterious mutations, thus limiting their counter-selection (SULTANA *et al.*, 2017). Some Ty retrotransposons target the upstream region of pol-III transcribed genes in *S. cerevisiae*, which allows both their expression and their location in a 'safe' environment, with no risk of disrupting essential genes (CHEUNG *et al.*, 2018; GUO *et al.*, 2015; SPALLER *et al.*, 2016). TEs spreading and fixation is intrinsically linked with their activity in gonads, and more precisely in germ cells (DECHAUD *et al.*, 2019). They could be positively selected if they insert in a region allowing their expression in this

tissue. Post-integration selection is also an important force modulating the location of TEs. Strongly deleterious insertions, such as insertions disrupting essential genes, are rapidly removed from the genome as individuals carrying them are strongly disadvantaged (MEDSTRAND *et al.*, 2002). In contrast, insertions with a positive impact on the host fitness will be retained by selection. TEs are now known to be able to modulate gene expression and to rewire entire regulatory networks (CHUONG, 2013; FESCHOTTE, 2008). It has been already proposed that TEs harbouring transcription factor binding sites (TFBS) allowing their germline expression could serve as "taxi" for regulatory elements to also control the expression of surrounding genes (DECHAUD *et al.*, 2019; REBOLLO *et al.*, 2012b; SUNDARAM *et al.*, 2014). They thus constitute good candidates to be involved in the fast evolution of sexual pathways in medaka. If such a positive selection concerns several insertions of a same TE family, these insertions can concomitantly appear enriched in different regions where they bring advantages. Finally, most insertions have limited impact on host fitness; they generally undergo genetic drift and are eliminated through random mutations. The combination of all these mechanisms can at the end lead to an enrichment of TEs in particular genomic regions, either due to insertional preferences, to their low impact in these regions, or in contrary to acquired positive functional role for instance in host gene regulation. To get more insights into these different evolutionary processes, we asked if the location of sex-biased TEs in *O. latipes* could be related to that of sex-biased genes.

We showed here that the sex-biased expression of neighboring genes and TEs is correlated: the closer TE copies are to genes, the higher is the correlation of their expression. This observation allows hypothesizing a co-regulation of both types of sequences by shared cis-regulatory elements, provided either by enhancers of the sexual genes and/or by the sex-biased TEs themselves. Both hypotheses are not mutually exclusive and further analyses will be necessary to understand the origin of this correlation. In the first hypothesis, TEs inserting close to sexual genes could co-opt regulatory sequences favoring their expression and by this way their transposition in gonads, particularly in germ cells for transmission to the next generation. Our observation that expressed TEs are slightly enriched in male-biased clusters could indeed reflect a preference of insertion of TEs in these regions. The effect, however, was minor, and not observed for female-biased clusters.

In our data, TE families enriched in copies mainly expressed in one sex are LTRs, for both testis and ovary. ERVs were previously shown to frequently give birth to enhancers in fast evolving tissues (SIMONTI *et al.*, 2017), and more particularly their LTRs that contain many TFBS (TENG *et al.*, 2014; THOMPSON *et al.*, 2016). Furthermore, LTR elements are known to escape repression in tissues such as testis or placenta, which are hypomethylated and thus allow a higher global transcriptional activity (CHUONG, 2013; MOLARO et MALIK, 2016). However, further experiments would be needed to demonstrate such a recruitment of LTR elements for regulatory purpose. Even if we could not identify any sequence insertion preference between related Gypsy elements with the same sex-biased expression, from our data we cannot completely eliminate a purely neutral model where TEs preferentially insert in regions of open chromatin and subsequently follow the expression of neighboring genes. About three times more TEs were found overexpressed in testes than in ovaries. This observation agrees with the transcription of various genomic elements known to occur in testis (SOUMILLON *et al.*, 2013). If the majority of these transcripts are probably non-functional, this high level of expression could also favor the birth of new genes or regulatory elements in this organ, in particular from TEs (SIMONTI *et al.*, 2017).

Finally, we demonstrated in our work that sex-biased TE copies are enriched in gene clusters with the same sex-biased expression. We particularly identified an Unknown TE family strongly biased toward testis expression, and preferentially localized in regions considered as male-biased. Of note, expressed copies of this family are enriched in binding motifs for SOX8 and HOXD13, two factors involved in male sexual function. This TE family constitutes thus an interesting candidate to investigate for a potential role in the evolution of sex chromosomes and/or the regulation of sexual development of the medaka fish.

2.6 Methods

Experimental animals

Laboratory reared medaka (*Oryzias latipes*) of the Carbio strain were used. All fish were kept under standard photoperiod cycle of 14 hr/10 hr light/dark at 26°C (±1°C). Animals were kept and sampled in accordance with the applicable EU and national German legislation governing animal experimentation, in particular all experimental protocols were approved through an authorization (568/300-1870/13) of the Veterinary Office of the District Government of Lower Franconia, Germany, in accordance with the German Animal Protection Law (TierSchG).

Sampling and sequencing

The gonads of *O. latipes* were dissected. As testis are small, the testes of three males were pooled in one replicate. We generated 3 testis replicates (3x3 fish) and 3 ovary replicates (3x1 fish). Total RNA was isolated using RNeasy®Mini kit (Qiagen) following the manufacturer's instructions. RNA quality was assessed by measuring the RNA Integrity Number (RIN) using an Agilent 2100 Electrophoresis Bioanalyzer Instrument (Agilent Technologies 2100 Expert). RNA samples with RIN > 8 were used for sequencing. RNA sequencing libraries were constructed following the standard TruSeq Illumina mRNA library preparation protocol (www.illumina.com; Illumina Inc., BGI, Hong Kong), with a read length of 100 and sequencing depth for paired end of 62-72 million reads.

Genome and transposable elements annotation

The genome of *Oryzias latipes* strain *Hd-rR* was sequenced and assembled with chromosome length scaffolds (https://www.ncbi.nlm.nih.gov/assembly/GCF_002234675.1). This genome is also annotated (ftp://ftp.ncbi.nlm.nih.gov/genomes/refseq/vertebrate_other/Oryzias_latipes/latest_assembly_versions/GCF_002234675.1_ASM223467v1).

Transposable elements were annotated using the following protocol. First, a TE consensus database was generated using RepeatModeler (http://www.repeatmasker.org/RepeatModeler/) (1596 consensi). To avoid false positives we removed short consensi under 80nt (1400 resulting consensi), we self-blasted each consensus to find potential satellite sequences (1398 resulting consensi), we removed non-TE genes by blasting the consensi against NCBI (1382 resulting consensi), and we removed the redundant consensi by blasting the bank against itself (947 consensi). We crossed the bank with LTRharvest (Ellinghaus *et al.*, 2008) output to reannotate some ERV TEs, and added

Gypsy, ERV and Copia elements (1262 resulting consensi). We also added 2 Helitron sequences from HelitronScanner (XIONG *et al.*, 2014) that were not found by other programs (1264 consensi). Some SINE sequences were reannotated using SINE_scan (Mao and Wang 2017). We finally ran MITE-hunter (HAN et WESSLER, 2010) but after manual checking we did not find any good consensus to add. The bank was used to annotate the genome with RepeatMasker (http://repeatmasker.org/). All TE copies annotated by the same consensus sequence are considered part of the same TE family.

Gene expression analysis

A detailed description of our protocol along with all parameters applied here is listed in supplementary file *supp_data.odt*. Read mapping was performed with Hisat2 version 2.1.0 (KIM *et al.*, 2019). As we wanted to exclude from the assembly, at this step, potentially expressed TEs, we discarded multimapped reads. We then used StringTie (PERTEA *et al.*, 2015) to assemble the transcripts using the genomic coordinates of the aligned reads, and to quantify transcript expression in each sample. We used the ballgown R package (Frazee et al. 2015) to estimate the TPM (Transcript Per Million) expression of each gene or transcript. Transcripts with low expression were filtered out as recommended in the new tuxedo procedure (PERTEA *et al.*, 2016). Genes and transcripts reconstructed by StringTie (PERTEA *et al.*, 2015) were compared to the reference gene annotation of the genome. Each reconstructed transcript was assigned a class code depending on its similarity to a reference transcript, allowing to know if it was already present in the reference or if it is new.

Filtering of assembled transcripts

45,444 transcripts were identified with the previous pipeline. Many of them might correspond to TEs, mapping errors, or missassembled transcripts due to low expression. We applied different filters to get a clean set of coding or non-coding, expressed, and non-TE transcripts. The details of these filters are given in *supp_data.odt* file. Briefly, we selected those with a mean expression level higher than 0.5 FPKM (Fragment Per Kilobase per Million). We then removed transcripts when more than 40% of their sequence was masked by our bank of TEs. Finally, we separated coding from non-coding transcripts with an ORF detection, to keep only transcripts having an ORF of more than 300 nucleotides. Selecting genes with at least one coding transcript and considered as non-TE, we ended up with 17,254 coding genes and 3,334 non-coding genes (supp. Fig 2.23).

Differential expression analysis

Details of the differential expression analysis are given in *supp_data.odt*. Differentially expressed genes were identified using the R package DESeq2 (LOVE *et al.*, 2014). The method used was apeglm (ZHU *et al.*, 2019) and we applied a threshold of log_2FC=1. We considered genes as differentially expressed between testis and ovary when the s-values were lower than 0.005 (FSOS = 0.5%).

TE expression quantification

We used SQuIRE (YANG *et al.*, 2019) (https://github.com/wyang17/SQuIRE) to estimate TE expression at copy level resolution. This program does not count several times multimapped reads

that could be assigned to several (highly similar) TE copies. Using different parameters such as local proportions of uniquely mapped reads, it at the end attributes each read to a specific locus, or "share" it between different loci but with a divided score. SQuIRE thus does not overestimate the expression of young TE families containing several highly similar copies. SQuIRE is divided in different steps to perform its analysis. The "Fetch" step retrieves the genome of interest along with gene and TE annotations on UCSC. As the O. latipes genome available on UCSC is not up to date, and as we build our own TE library, we did not use this step and generated the corresponding files using the latest *O. latipes* genome and our TE library. We then ran SQuIRE "clean", "map", "count" and "call" steps to estimate TE expression.

Reverse transcriptase phylogeny

We first defined a reference set of RT amino-acid sequences from a subset of the different LTR consensus sequences identified in the medaka genome, for the different LTR superfamilies (Gypsy, Copia, ERV, and BEL/Pao). Using *ORFfinder* https://www.ncbi.nlm.nih.gov/orffinder/ and conserved domain detection https://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi we obtained the amino acid sequence of their RT. We stored these sequences in `Reference_RT.fa` (supp. data 8). We then compared all LTR elements with these reference RTs by blastx. Using *recup_prot_query_-Blastx.py* we retrieved for each TE the best hit of more than 50 amino acids. We added some known RT from other species found on NCBI: `RT_ref.fa` (supplementary data 20). These sequences were aligned with mafft (version v7.450) https://mafft.cbrc.jp/alignment/software (KATOH, 2002; KATOH et STANDLEY, 2013), and stored in `RT_only.mafft` (supp. data 10). This alignment was fed to *FastTree* (version 2.1.11) (PRICE *et al.*, 2010) with LG model to build the shown phylogeny (supp. data 11). The different codes for blastx or mafft used to generate the phylogeny are described in `Phylogeny.md` on the gitlab repository (see below). We used the exact same approach to design the tree using expressed Gypsy copies (supp. data 12, 13).

Building gene and TE stretches

The method used to generate Figure 2.3. is the same as described in Boutanaev *et al.*, 2002 (BOUTANAEV *et al.*, 2002). The genes of TEs are represented by their expression bias, either testis, "**T**", either ovary, "**O**", either non-biased, "**N**". They are represented as a sequence on each chromosome, like, as an example:

```
-- NNNOONNTNNTTTNNNTONNTOTNNNTTTTTNONNONNTN --
```

For each bias, the number of stretches is then counted. In this example, for testis-biased genes, we observe 5 stretches of 1 gene, 1 stretch of 3 genes, and 1 stretch of 5 genes. This will be represented by the blue bars in Figure 4. Then the genome if then shuffled 1000 times (not 50 times as it is done in Boutanaev *et al.*, 2002 (BOUTANAEV *et al.*, 2002)). For each shuffling, the same counts are computed and we finally plot the median value with the 95% fluctuation interval. This is not a confidence interval of the mean, but this represents the distribution of 95% of the values obtained through the shuffles. The same approach was applied for ovary-biased genes. The script used to generate such barplot is available from the gitlab repos (see below): *stretch_of_genes.R*.

Detection of gene clusters

The search for such stretches however is not fully suitable to detect genomic regions enriched in biased genes. If two consecutive genes are separated by a long gene desert for example, they can still belong to a common stretch, in spite of their important intervening distance. Additionally, a single gene with a different expression can split a cluster, hindering its identification in spite of the global common expression bias of surrounding genes. It is thus interesting to study gene clustering on chromosomes in a more relaxed manner. We thus developed a new method to determine if genes are randomly distributed on chromosomes, or if they are grouped according to their expression, by designing a bootstrapping approach. The pipeline is available at https://gitlab.com/Corend/gene_-clusters_pyth, and was already used in a study on waterstriders (TOUBIANA *et al.*, 2020).

Step 1: Design of the expression profile

First, a sliding window is designed on the genome. The size and step of the sliding window can be set by the user through the -step and -window parameters in the pipeline. The window and step sizes have to be chosen carefully according to the gene density of the studied genome, to ensure a sufficient statistical power. We tested different values and retained a step size of 25kb, and a window size of 500kb. Hence, each window overlaps with the next 20 windows. Then, using the bed file of genes provided by the user (-b) and the expression of each gene (-e), the mean log2Fold change of the genes is calculated in each window of each chromosome. The result can be used to display the fold change across each chromosome (**Fig. 2.3.C.**).

Step 2: Bootstrap analysis

Randomly distributed genes can form clusters just by chance. To test whether a cluster is observable by chance, we designed a bootstrap approach. The bootstrap number can be adjusted by the user (-boot). We used 10,000 bootstraps for our analysis. For each bootstrap, all the genes in the genome are randomly redistributed at each locus. Then the mean log_2FC is calculated again in each window. After the bootstraps, we have for each window the observed mean log_2FC, and 10,000 theoretical mean log_2FC. We calculate how many times the observed value is superior to the bootstrap, and how many times it is inferior. If the observed value is always superior to the bootstrap values, then this region can be considered as "significantly testis-biased". On the opposite, if the observed value is inferior to the bootstraps, it is considered as "ovary-biased". The output of the pipeline corresponds to the mean log_2FC for each window and its associated bootstrap value.

Step 3: Statistical analysis

This part is not included in the pipeline so that the user can choose its own bootstrap threshold. We converted the bootstrap value in a p-value: $\frac{(min(Bootstrap\_superior,\ Bootstrap\_inferior)\times 2)}{10000}$. We then converted the p-values (1 p-value per window) in q-values using Benjamini-Hochberg FDR correction from the R package qvalue (STOREY *et al.*, 2019). We use a q-value threshold of 5%, meaning that among the windows considered as enriched in biased genes, 5% are false positives. Using this threshold, we colored the regions on the plot (**Fig. 2.3.C.**).

Enrichment of TEs in biased regions

We investigated the localization of the TE copies of each family (**Fig. 2.5.**). We tested if each family was significantly enriched in the ovary- or testis-biased regions using Fisher's exact test according to the method described in Karakülah and Suner, 2017 (KARAKÜLAH et SUNER, 2017). As we tested all the TE families (1164), we applied a Bonferroni correction to the test by taking 0.05/1164 as a p-value threshold. The ternary plot was generated using the script *Ternary_plot.R*. The data used in the script is available in supplementary data 14.

Figures

Scripts used to generate the figures of the paper are available on the following gitlab repository: https://gitlab.com/Corend/te_gene_expression.

2.7 Acknowledgements

This work was supported by a grant from the French National Research Agency (ANR-16-CE92-0019 – EVOBOOSTER) to JNV and by the Deutsche Forschungsgemeinschaft (Scha 408/13-1) to MS in the ANR/DFG cofunding program. We would like to thank the French Institute of Bioinformatics - IFB CNRS UMS 3601 - (funded as part of Investissement d'avenir program managed by Agence Nationale pour la Recherche, contract ANR-11-INBS-0013) for providing life science data and tools, storage and computing resources on the IFB national service infrastructure in bioinformatics.

2.8 Supplementary informations

Gene expression analysis details

Read mapping

The alignment of all read samples on the reference genome was performed with *Hisat2* version 2.1.0 (KIM *et al.*, 2019) using following command line options: `hisat2-build $Genome $GenomeIndex` to build an index of the genome; `hisat2 -p 7 -k 2 -dta -x $GenomeIndex -1 $Sample_1.fq.gz -2 $Sample_2.fq.gz -S $Sample.sam` to align each sample on the reference genome. The `-dta` option is used as recommended in the New tuxedo pipeline to perform an assembly of the transcripts (PERTEA *et al.*, 2016). The `-k 2` option allows to report a maximum of 2 alignments per sequencing read. As we wanted to exclude potentially expressed TEs from the assembly, we discarded multimapped reads using the following command line: `samtools view -h -@ 7 -f 0x2 $Sample.bam |awk '{if(substr($1, 0, 1)=="@" || $5==60){print $0}}' > $Sample.sam`. The option `-f 0x2` extracts the properly aligned paired reads from the bam. The `awk` command also keeps the header section of the SAM file as well as the reads with a MAPQ = 60. In *Hisat2*, a MAPQ = 60 means that the read has been uniquely aligned regardless of the mismatches / indels number. The filtered SAM file was finally converted in BAM and sorted: `samtools view -h -@ 7 -b -S $Sample.sam > $Sample_filtered.bam && samtools sort -@ 7 $Sample_filtered.bam`

`sorted.bam && mv sorted.bam $Sample_filtered.bam`.

Transcript assembly

We used *StringTie* (PERTEA *et al.*, 2015) to assemble the transcripts using the genomic coordinates of the aligned reads, for each sample: `stringtie -p 7 -o $Sample_assemb.gtf -j 2 $Sample_filtered.bam && echo "$Sample$_assemb.gtf" » mergelist.txt`. The -j option specifies that at least 2 reads are needed to support an exon-exon junction and create a new transcript. This is thus more stringent than the default value of 1 used by *StringTie*. One assembly per sample was created; we used the following command line to merge these assemblies: `stringtie –merge –o merged.gtf mergelist.txt`.

Transcript expression quantification

StringTie (PERTEA *et al.*, 2015) was also used to quantify transcript expression in each sample: `stringtie –e –B –p 7 –G merged.gtf –o $Sample/$Sample.gtf $Sample_filtered.bam`. -e option is used to estimate the expression of transcripts of the GTF file given to the -G option. -B option is used to create an output needed to run ballgown (FRAZEE *et al.*, 2015) later.

Count matrix generation

We used the `prepDE.py` (https://github.com/gpertea/stringtie/blob/master/prepDE.py) script from *StringTie* (PERTEA *et al.*, 2015) to generate the count matrix from the output of the previous step.

Gene expression level quantification

We used the *ballgown* R package (FRAZEE *et al.*, 2015) to estimate the TPM (Transcript Per Million) expression of each gene or transcript. Data were loaded with ballgown (FRAZEE *et al.*, 2015) from the output of `stringTie -e -B`. Transcripts with low expression were filtered out as recommended in the new tuxedo procedure (PERTEA *et al.*, 2016) using: `subset(data,"rowMeans(texpr(data)) > 0.5", genomesubset=TRUE)` (supplementary fig 1). FPKM were then calculated using *texpr* function from *ballgown* (FRAZEE *et al.*, 2015). TPM (PACHTER, 2011; WAGNER *et al.*, 2012) were finally computed from this table by normalizing the sum of each sample to 10^6.

Comparison between the new annotation and the reference gene annotation

Genes and transcripts reconstructed by *StringTie* (PERTEA *et al.*, 2015) were compared to the reference gene annotation of the genome using gffcompare (https://ccb.jhu.edu/software/stringtie/gffcompare.shtml). Each new transcript was assigned a class code depending on its similarity to a reference transcript, as described in gffcompare documentation. With this step we are able to know if each transcript found in our data is already present in the reference or if it is new.

Correlation between TE and genes proximity, links between gene clusters and TE expression

The scripts used to generate the violin plot (**Fig. 2.4., supp. Fig 2.19**) are available on the gitlab (see below). First, we ran *Correlation_violin.py*. The input files are a bedfile of the genes selected (coding, non-coding or both), and a bedfile of the TEs. Then, we ran *Correlation_violin.R*, with as input the expression of genes and TEs (supp. data 15 and 16), along with the output of the previous script. Supp. figure 2.21 was generated using the script *Violin_mosaic.R* from the gitlab (see below). On the mosaic plot, the sum of the black value (37,038) is not equal to the total number of TE copies (37,108): some copies were removed as they were overlapping different gene clusters.

Supplementary data

— **Supplementary data 1:** TE bank generated from *O. latipes* genome in FASTA format.

— **Supplementary data 2:** TE annotation of O. latipes genome using the TE bank.

— **Supplementary data 3:** Table of the most expressed TE families.

— **Supplementary data 4-7:** Gene clusters profiles for all chromosomes using coding genes only, non coding genes only, all genes or transposable elements.

— **Supplementary data 8:** Fasta file of the reference reverse transcriptase used to generate the phylogeny.

— **Supplementary data 9:** Fasta file of the reference reverse transcriptase retrieved from NCBI.

— **Supplementary data 10:** Reverse transcriptase alignment using TE consensi.

— **Supplementary data 11:** Phylogenetic tree generated using reverse transcriptase alignement from TE consensi.

— **Supplementary data 12:** Reverse transcriptase alignment using Gypsy TE copies.

— **Supplementary data 13:** Phylogenetic tree generated using reverse transcriptase alignement from Gypsy TE copies.

— **Supplementary data 14:** Number of sex-biased TE copies per TE family.

— **Supplementary data 15:** Gene expression data used to generate the Maplot.

— **Supplementary data 16:** TE copies expression data.

Supplementary tables

		All genes (coding + non coding)	**Coding genes only**	**Non-coding genes only**
Number of testis-biased clusters detected		32	9	3
Number of ovary-biased clusters detected		18	10	2
Size of testis-biased clusters	**Total**	28.945Mb	9.053Mb	20.46Mb
	Mean	905kb	1.006Mb	6.82Mb
	Minimum	455kb	850kb	5.5Mb
	Maximum	1.96Mb	1.360Mb	7.7Mb
Size of ovary-biased clusters	**Total**	18.2Mb	10.46Mb	12.98Mb
	Mean	1.011Mb	1.046Mb	6.49Mb
	Minimum	700kb	850kb	4.4Mb
	Maximum	1.715Mb	1.658Mb	8.58Mb
Number of genes in testis-biased gene clusters	**Total**	828	205	129
	Mean	26	23	43
	Minimum	7	8	33
	Maximum	56	54	55
Number of genes in ovary-biased gene cluster	**Total**	457	228	43
	Mean	25	23	22
	Minimum	4	6	13
	Maximum	64	40	30

Table 2.1 – **Total number and size of detected gene clusters using all genes, coding genes only, or non-coding genes only.**

Bias	Gene	Function	Reference (doi)
Male	*dnah9*	Dynein axonemal heavy chain 9. This gene encodes the heavy chain subunit of axonemal dynein, a large multi-subunit molecular motor. Axonemal dynein attaches to microtubules and hydrolyzes ATP to mediate the movement of cilia and flagella.	10.1007/ s003359900202
Male	*morn3*	Membrane occupation and recognition nexus repeat containing 3. Regulator of spermatogenesis.	10.4103/ 1008-682X.138186
Male	*nsmce1*	Non-structural maintenance of chromosomes element 1. Meiotic chromosome segregation.	10.1093/ dnares/dsaa019
Male	*pdgfd*	Platelet-derived growth factors D. Cell proliferation.	10.1210/ er.2010-0004
Male	*dync2h1*	Cytoplasmic dynein 2 heavy chain 1. Functions in cilia biogenesis.	10.1016/ S0378-1119(97) 00417-4
Male	*fzd4*	Frizzled-4. Involved in adult spermatogenesis.	10.1095/ biolreprod.112. 105809
Male	*numa1*	Nuclear mitotic apparatus protein 1. Binding partner of BRAP2 in human testis.	10.1016/ j.bbamcr.2013.05. 015
Male	*cfap54*	Cilia And Flagella Associated Protein 54.	10.1091/ mbc.E15-02-0121
Male	*dnajc18*	DnaJ homolog subfamily C member 18. Might play a role during germ cell maturation in adult rat testis.	10.12717/ DR.2017.21.3.237
Male	*dnajb13*	DnaJ homolog subfamily B member 13. Plays a role in the formation of the central complex of ciliary and flagellar axonemes.	10.1016/ j.ajhg.2016.06.022
Male	*ucp2*	Mitochondrial uncoupling protein 2. Regulation of human spermatozoa motility.	10.1159/ 000494479
Male	*armc3*	Armadillo repeat-containing protein 3.	10.1186/ s12863-016-0356-7
Male	*cfap221*	Cilia- and flagella-associated protein 221. May play a role in cilium morphogenesis.	10.1128/ MCB.00354-07
Male	*ccnblip1*	E3 ubiquitin-protein ligase CCNB1IP1. Limiting factor for crossing-over during meiosis.	10.1128/ MCB.23.6.2109- 2122.2003
Male	*lrguk*	Leucine-rich repeat and guanylate kinase domain-containing protein. Involved in multiple aspects of sperm assembly including acrosome attachment, shaping of the sperm head and in the early aspects of axoneme development.	10.1371/ journal.pgen. 1005090
Male	*ythdc2*	3'-5' RNA helicase YTHDC2. Plays a key role in the male and female germline by promoting transition from mitotic to meiotic divisions in stem cells.	10.1371/ journal.pgen. 1006704
Male	*cdkn2c*	Cyclin-dependent kinase 4 inhibitor C.	
Male	*cep350*	Centrosome-associated protein 350. Required for ciliation.	10.1098/ rsob.170114
Male	*tdrd12*	Putative ATP-dependent RNA helicase TDRD12. Probable ATP-binding RNA helicase required during spermatogenesis to repress transposable elements and preventing their mobilization, which is essential for the germline integrity.	10.1073/ pnas.1316316110
Male	*lamb1*	Laminin subunit beta-1. Expressed in the developing male and female gonads and mesonephros.	10.1046/ j.1432-0436. 1997.6230129.x
Male	*dmrt1a*	Doublesex- and mab-3-related transcription factor 1A. Transcription factor that plays a key role in male sex determination and differentiation by controlling testis development and germ cell proliferation.	10.1242/ dev.048751
Male	*zan*	Zonadhesin. Binds in a species-specific manner to the zona pellucida of the egg. May be involved in gamete recognition and/or signaling.	10.1086/ 508473
Male	*bobk*	Bcl-2-related ovarian killer protein homolog B. May play a role in apoptosis.	10.1038/ sj.cdd.4402016
Male	*buc*	Prion-like protein required for the formation of Balbiani body (electron-dense aggregates in the oocyte) and germ plasm assembly, and for the establishment of oocyte polarity during early oogenesis.	10.1016/ j.ydbio.2008. 05.557
Male	*hsd17b1*	Estradiol 17-beta-dehydrogenase 1. Favors the reduction of estrogens and androgens.	10.1096/ fj.02-0026fje

Table 2.2 – **Genes with sexual-related function found in male- and female-biased gene clusters.**

Supplementary figures

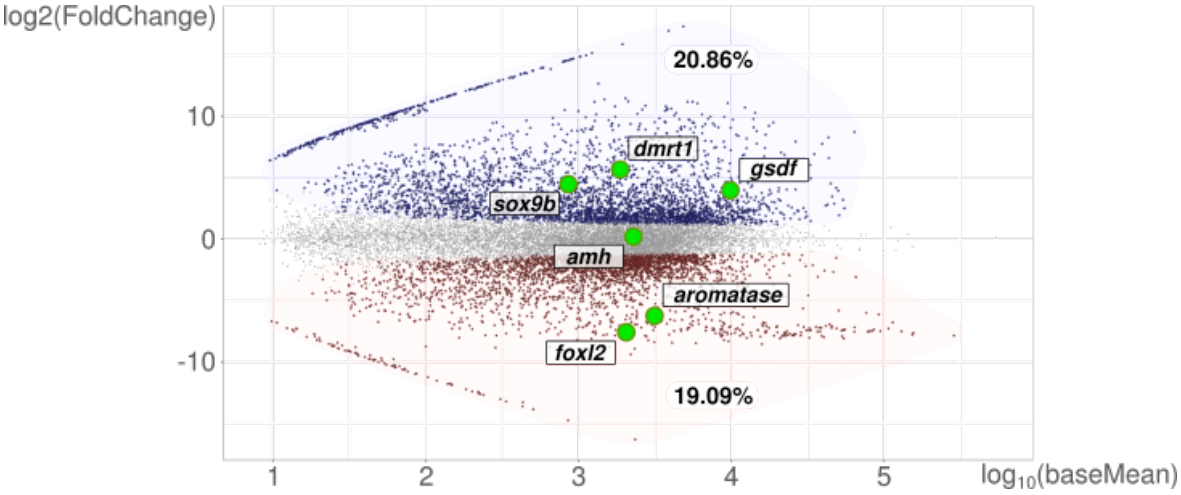

Figure 2.6 – **MAPlot of coding gene expression in the gonads of *O. latipes*.** Each dot represents a coding gene. The x-axis corresponds to the signal intensity averaged across all replicates, and the y-axis to the log2FC of expression between testis and ovary. The higher the log2FC of a coding gene is, the more it is over-expressed in testes (in blue, significantly, 3,600 genes), and the lower it is, the more it is over-expressed in ovaries (in red, significantly, 3,293 genes). The more the gene is on the right, the more it is overall expressed across all replicates. In green are displayed genes described in the literature as being involved in medaka sexual development and function (HERPIN *et al.*, 2013; HORIE *et al.*, 2016; KOBAYASHI *et al.*, 2017; NAKAMOTO *et al.*, 2006).

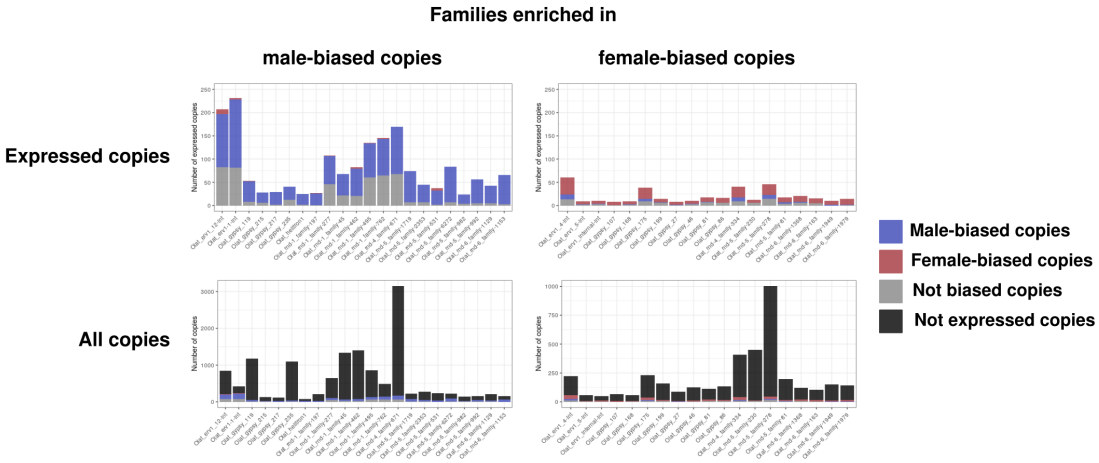

Figure 2.7 – **Number of sex-biased, expressed (but non-biased) and non-expressed TE copies in TE families enriched in sex-biased copies.**

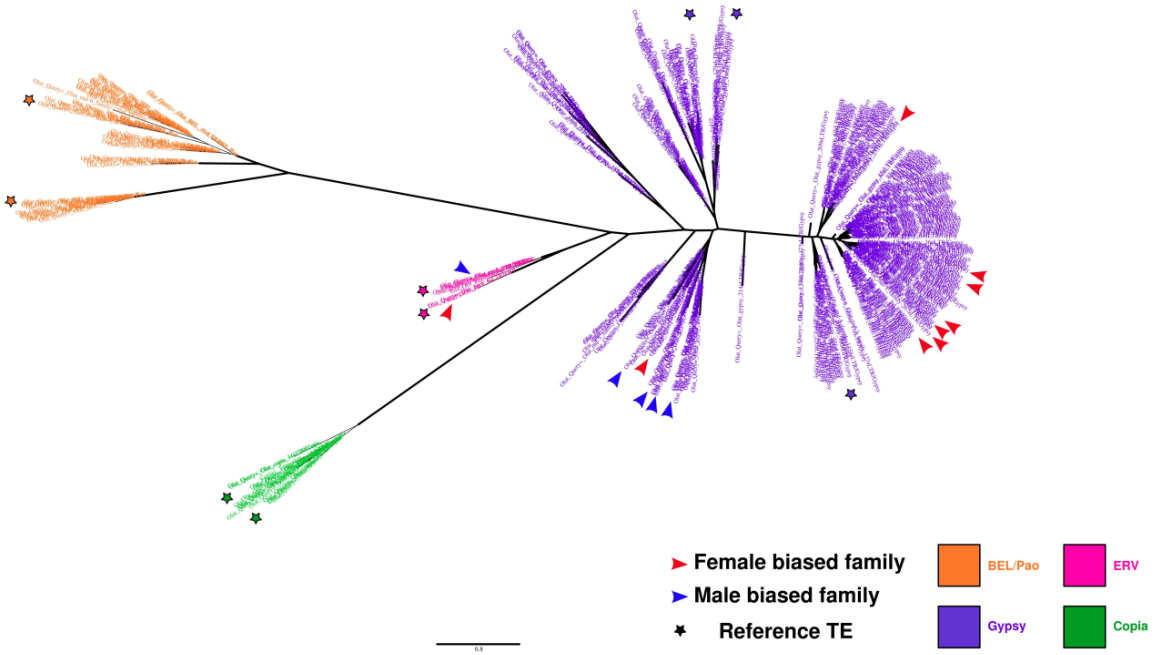

Figure 2.8 – **Phylogeny generated using the consensus of each LTR retrotransposon family.**

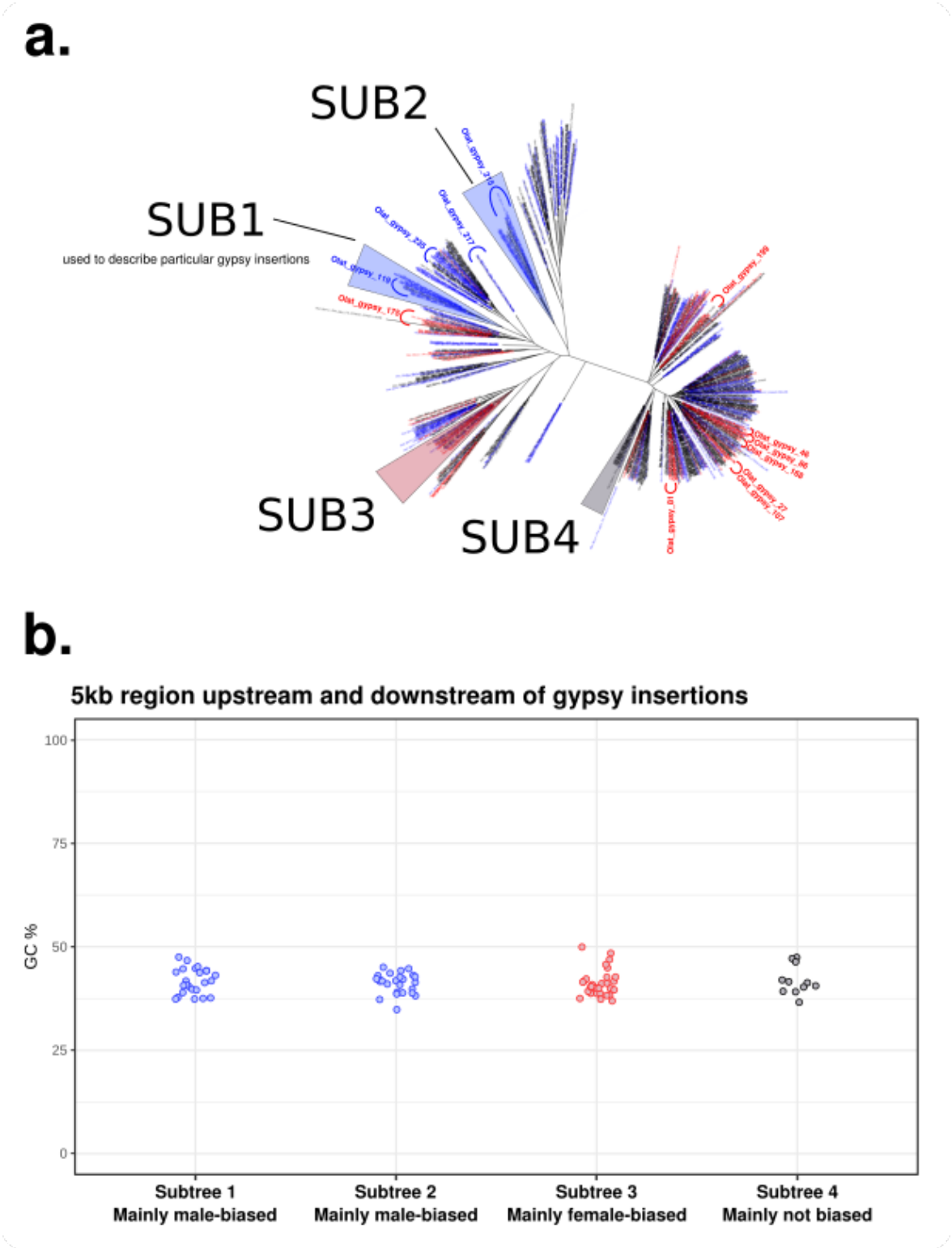

Figure 2.9 – **Phylogeny and surrounding regions GC% of *Gypsy* copies.** a. Phylogeny of expressed *Gypsy* TE copies of medaka using the amino acid sequence of the reverse transcriptase (RT). Tip colors correspond to the expression bias (red: ovary; blue : testis; black: expressed but not sex-biased). Four subtrees are highlighted (SUB1-4) for which we analyzed copy insertion sites. b. Distribution of GC% in 5kb upstream and downstream *Gypsy* insertion regions, for insertions of subtrees SUB1-4. GC% are not significantly different between subtrees ($p-val = 0.790$, ANOVA1). Blue: clusters of mainly male-biased copies; red: cluster of mainly female-biased copies; black: cluster of mainly non-biased copies.

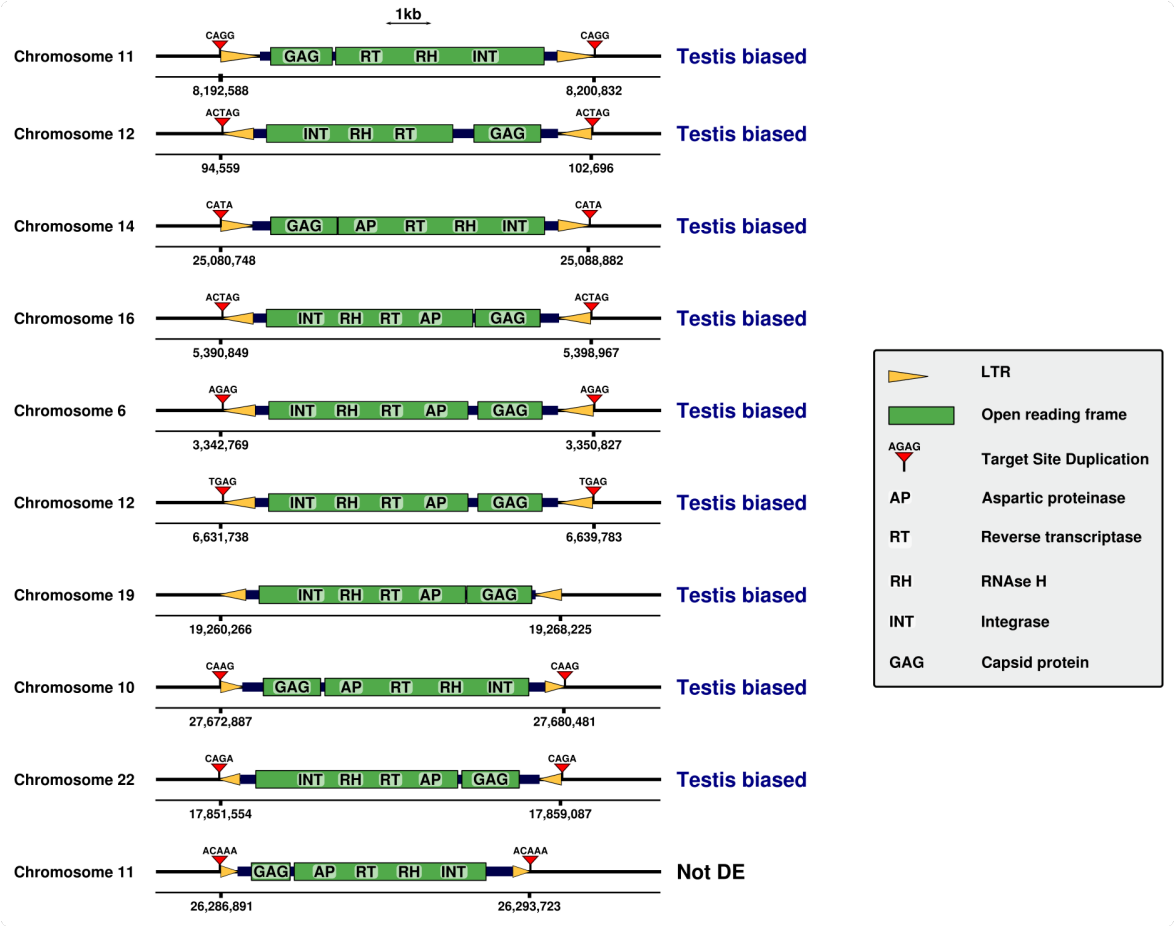

Figure 2.10 – **Structure of the 10 longest *Gypsy* insertions of the male-biased subtree 1 (Fig. 2.9.).** Green boxes : Open Reading Frames ; yellow arrows : LTRs ; AP : Aspartic Protease ; RT : Reverse Transcriptase ; RH : RNAse H ; INT : Integrase. DE: Differentially Expressed.

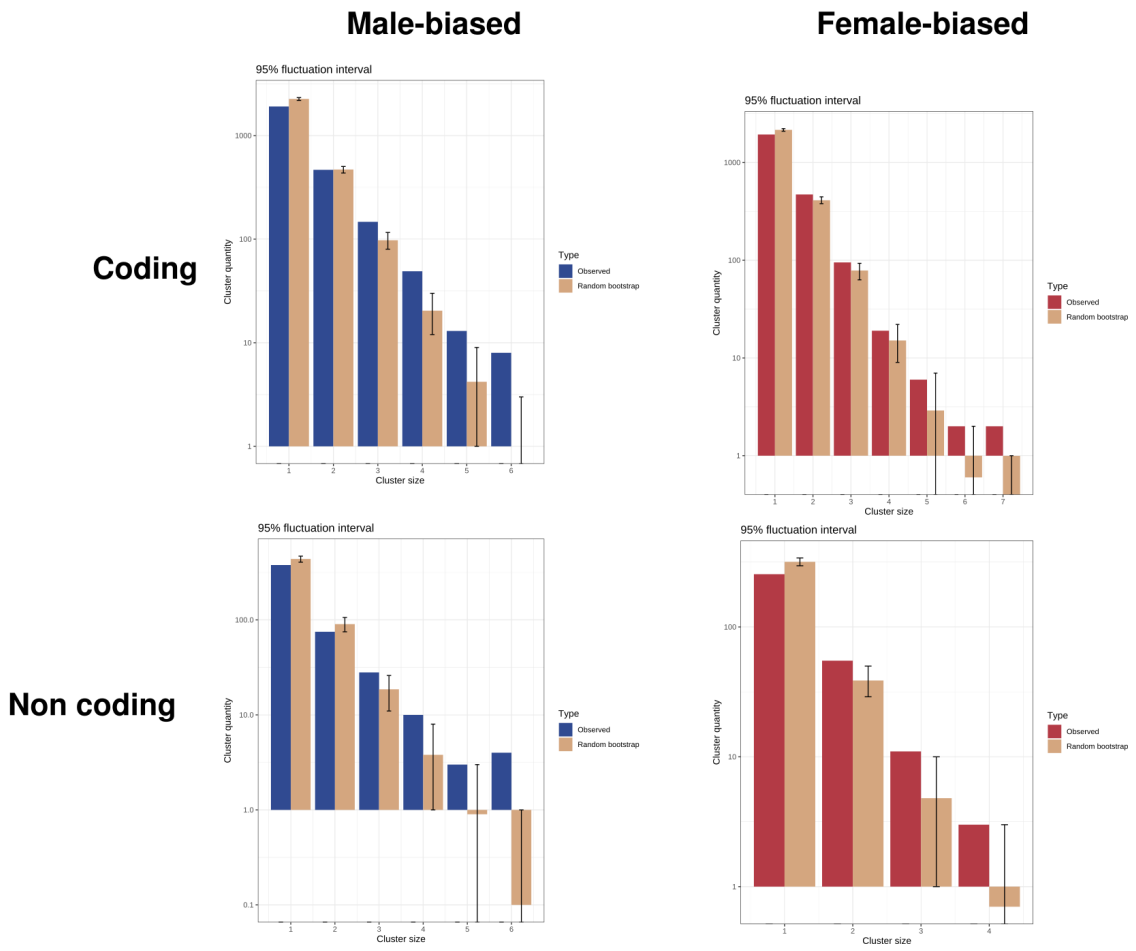

Figure 2.11 – **Stretches of genes obtained using only male-biased or female-biased coding genes or non-coding genes.**

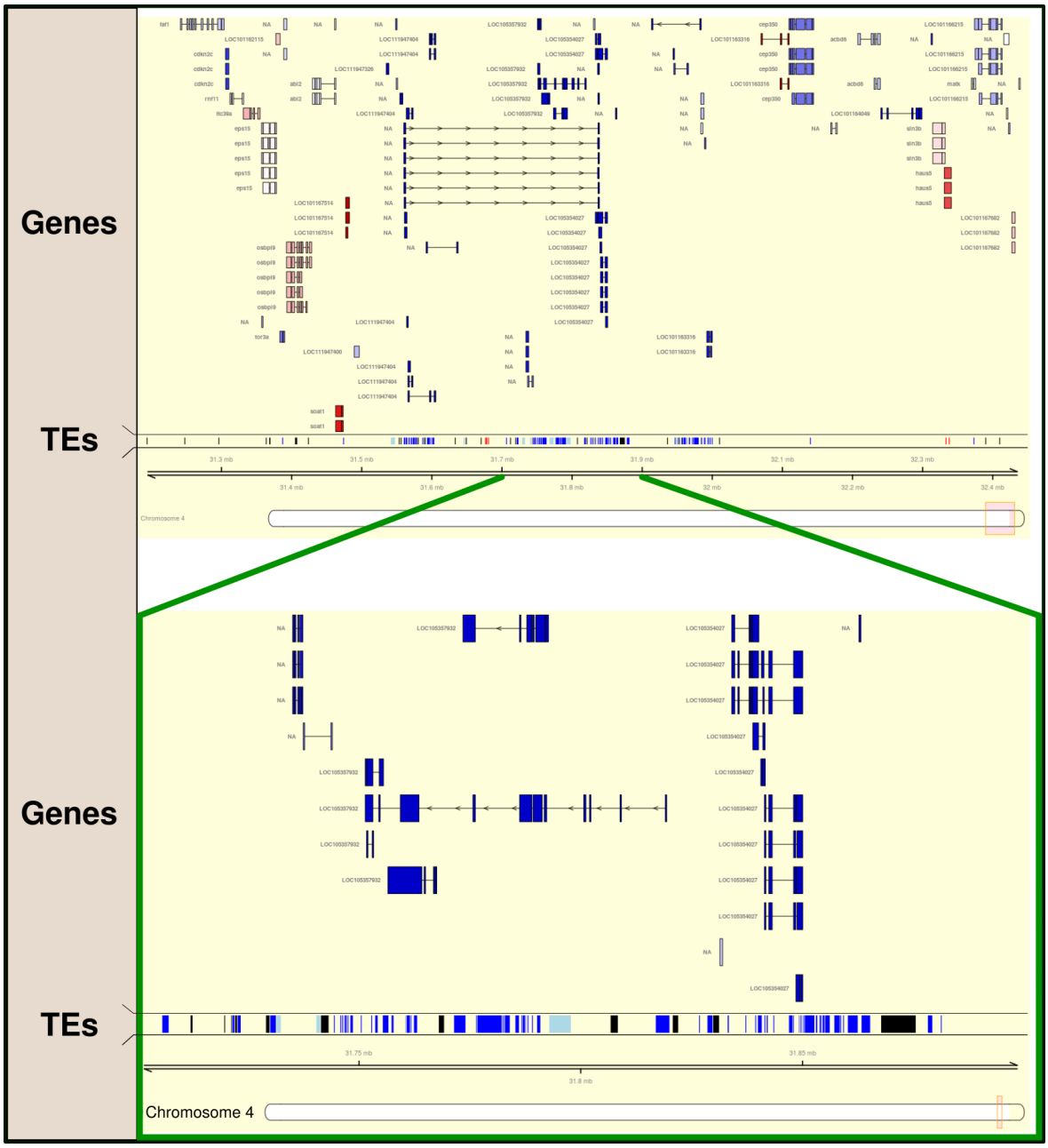

Figure 2.12 – **Representation of the cluster from chromosome 4.** Male-biased gene clusters located on chromosome 4. This cluster contains numerous TEs that are particularly concentrated in the central regions, both in intronic and intergenic regions.

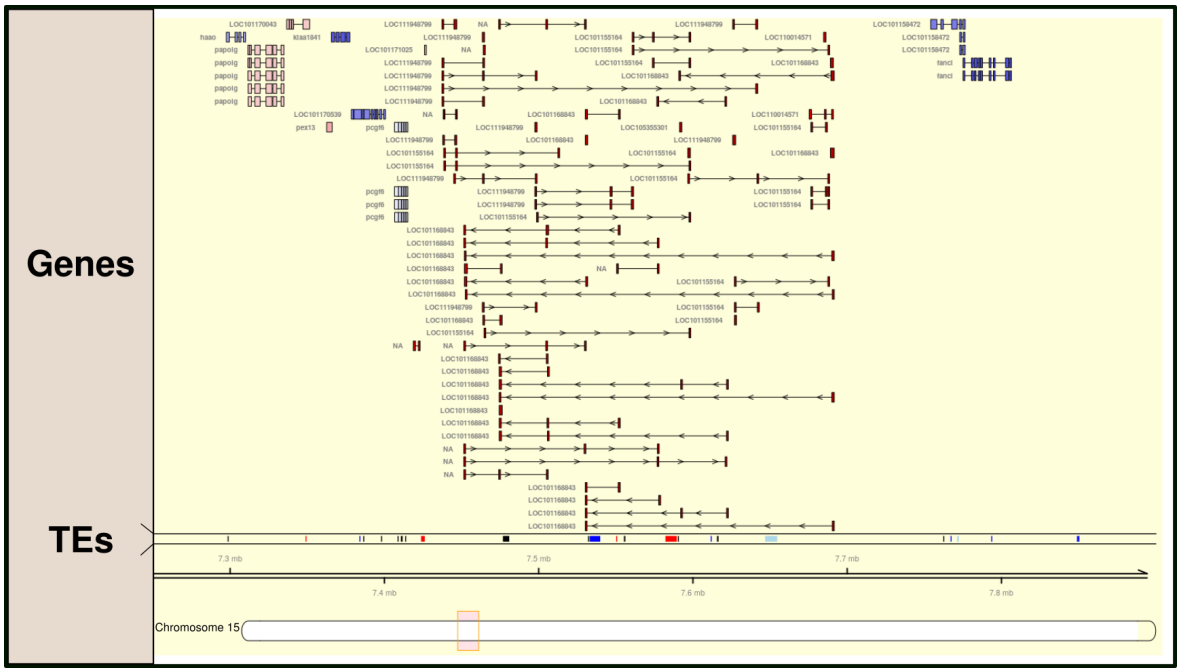

Figure 2.13 – **Representation of the female-biased gene cluster from chromosome 4.**

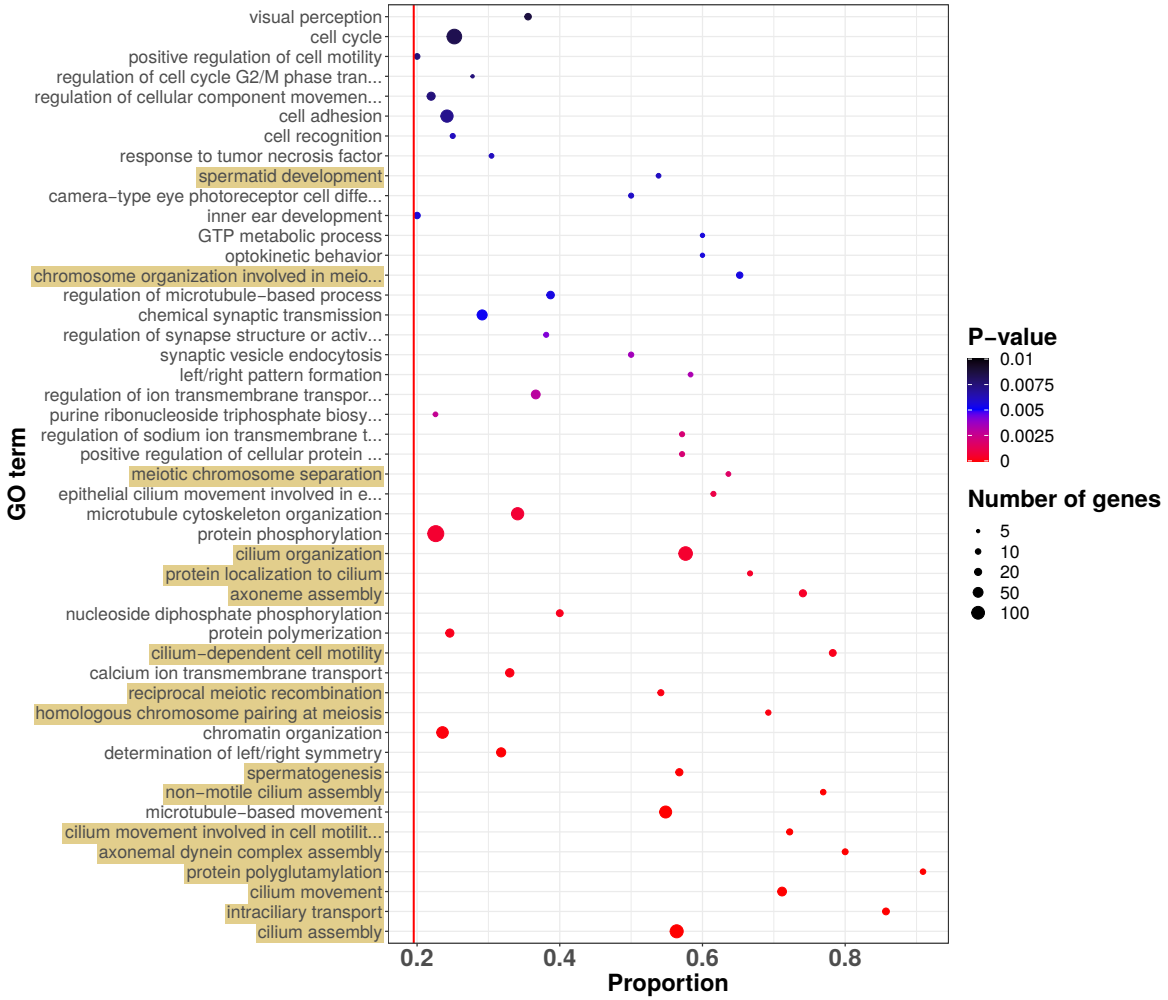

Figure 2.14 – **GO term enrichment of male-biased genes ($p < 0.01$).** The red line indicates the proportion of GO terms expected by chance. The x-axis shows the observed proportion, with higher proportions on the right. Any proportion at the right of the red line is higher than expected by chance. The size of the points is proportional to the number of genes associated with the annotation. The color of the point indicates the associated P-value, with the lowest p-values in red. Most significant GO terms are at the bottom. Terms highlighted in beige are related to male sexual function.

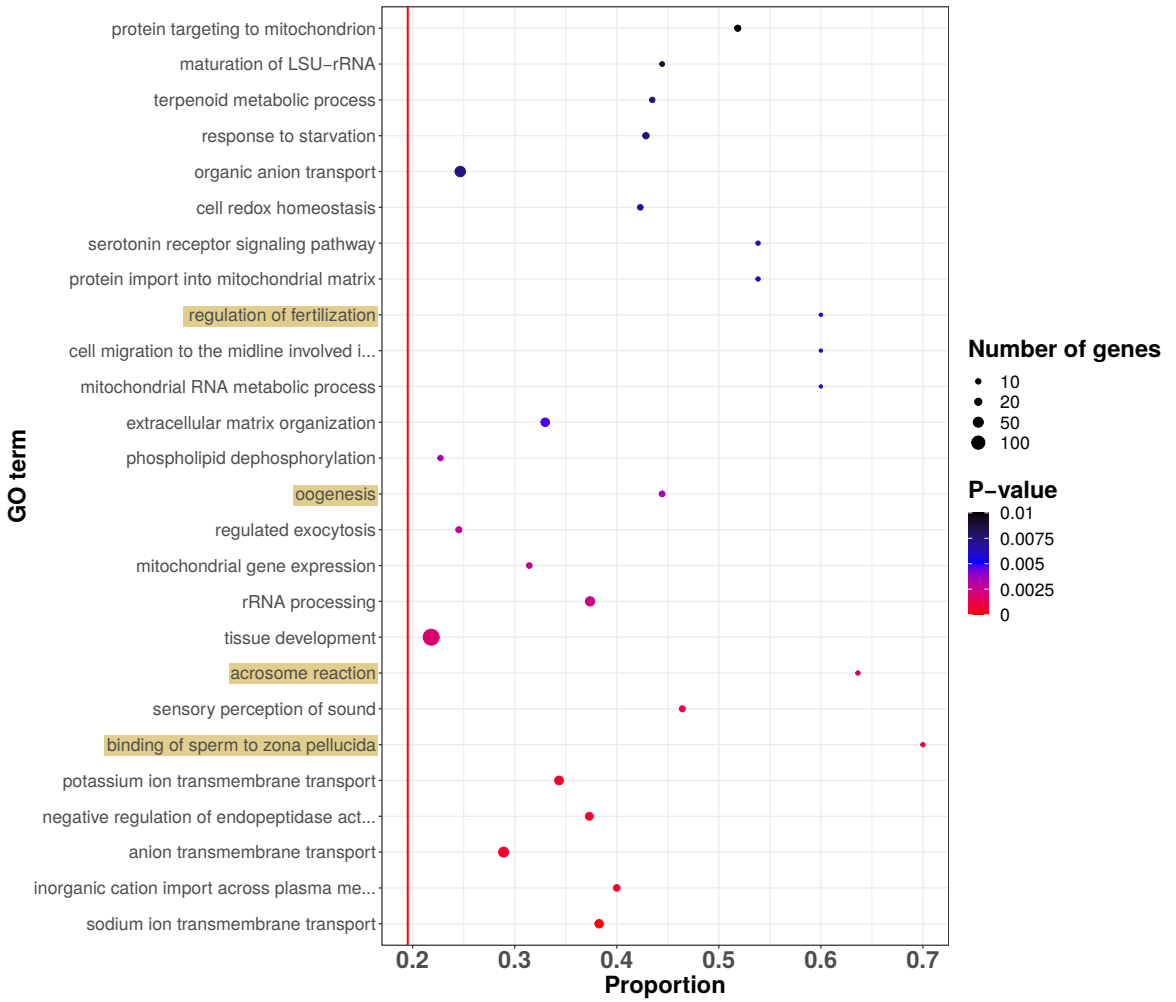

Figure 2.15 – **GO term enrichment of female-biased genes** ($p < 0.01$)**.** The red line indicates the proportion of GO terms expected by chance. The x-axis shows the observed proportion, with higher proportions on the right. Any proportion at the right of the red line is higher than expected by chance. The size of the points is proportional to the number of genes associated with the annotation. The color of the point indicates the associated P-value, with the lowest p-values in red. Most significant GO terms are at the bottom. Terms highlighted in beige are related to female sexual function.

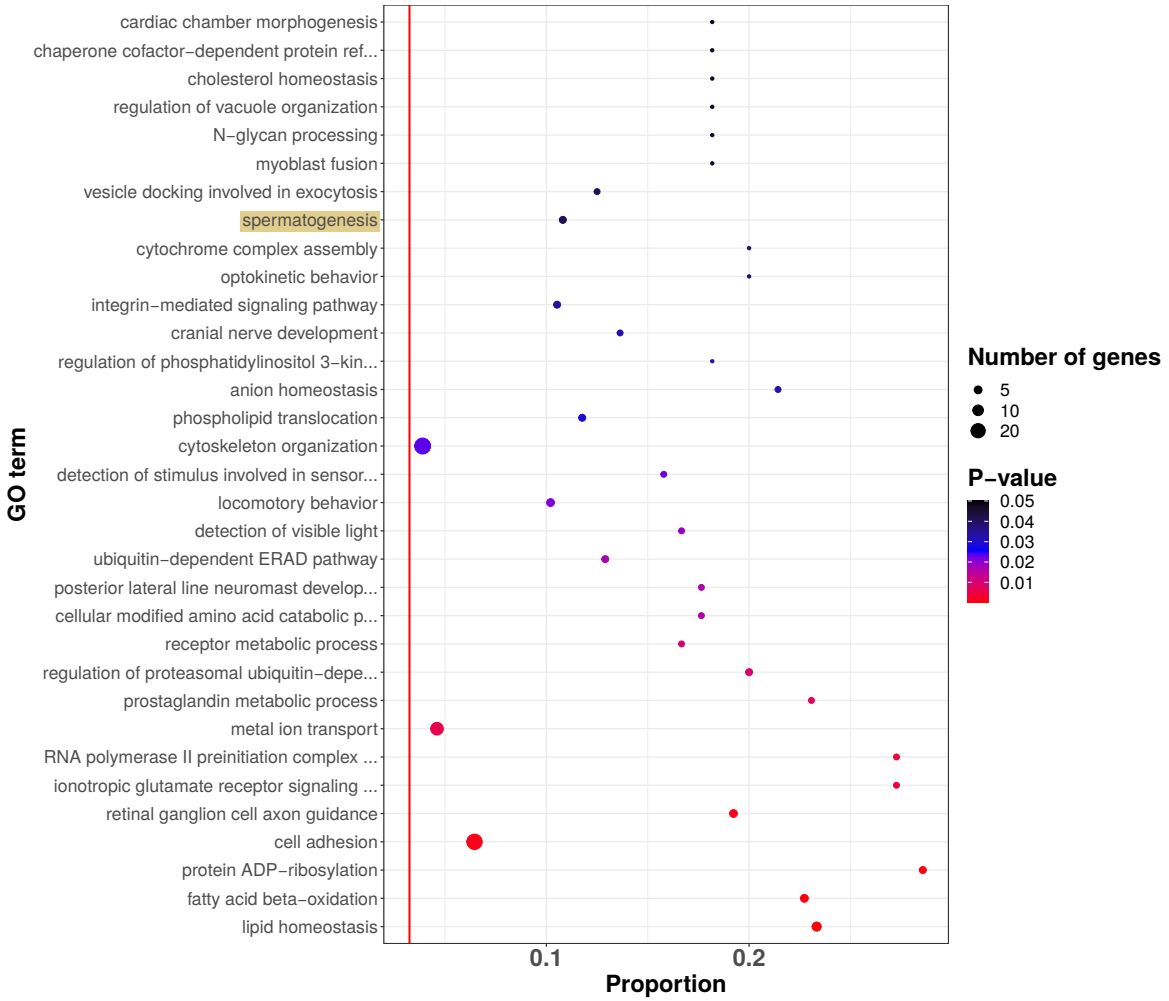

Figure 2.16 – **GO term enrichment of genes in male-biased clusters** ($p < 0.05$)**.** The term 'spermatogenesis' is significantly found associated to these genes. Apart from this term, the other enrichments do not specifically refer to gonadal function. The red line indicates the proportion of GO terms expected by chance. The x-axis shows the observed proportion, with higher proportions on the right. Any proportion at the right of the red line is higher than expected by chance. The size of the points is proportional to the number of genes associated with the annotation. The color of the point indicates the associated Pvalue, with the lowest p-values in red. Most significant GO terms are at the bottom. Terms highlighted in beige are related to male sexual function.

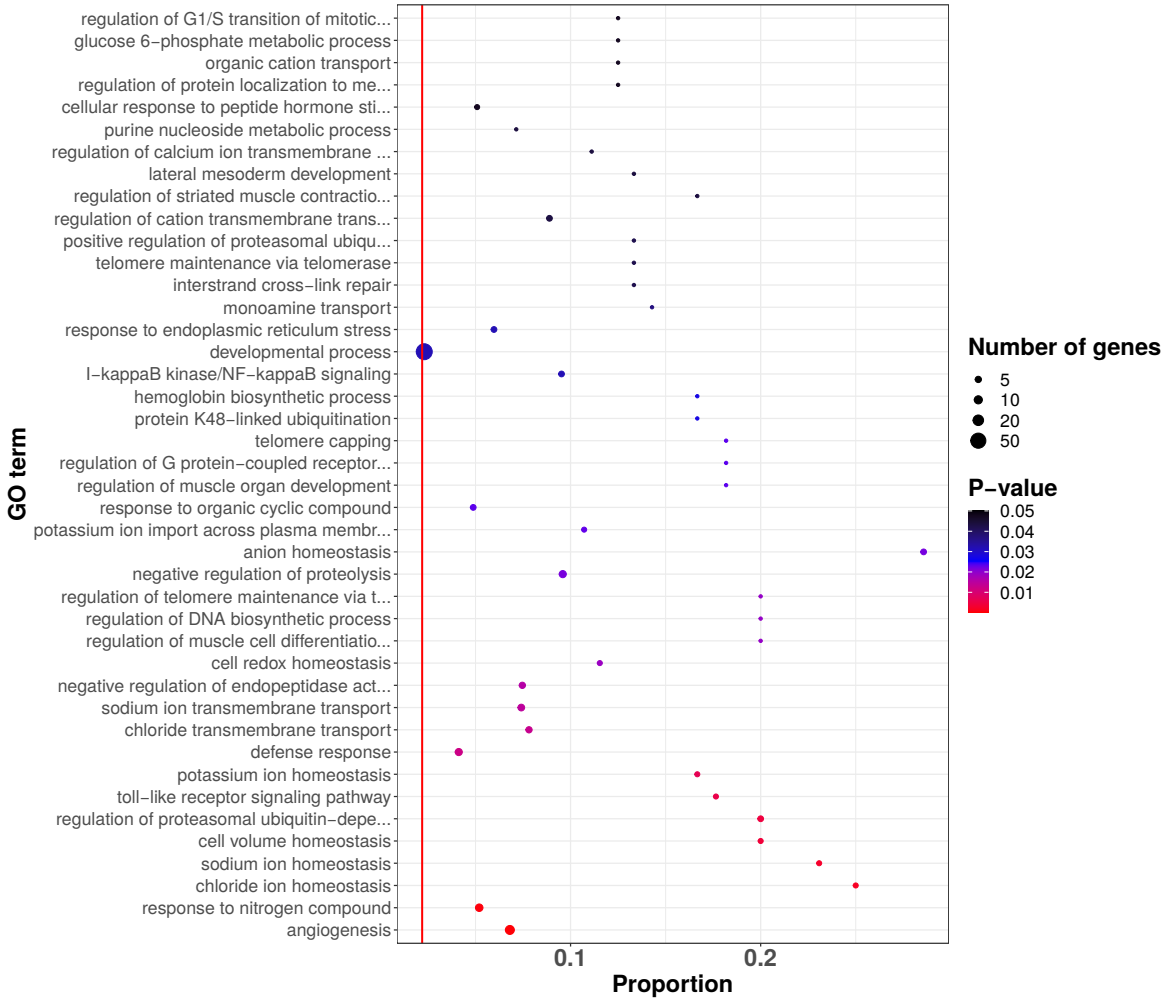

Figure 2.17 – **GO term analysis of genes in female-biased clusters ($p < 0.05$).** GO term analysis of genes in female-biased clusters ($p < 0.05$). No function obviously linked with female sexual function was found as enriched here. The red line indicates the proportion of GO terms expected by chance. The x-axis shows the observed proportion, with higher proportions on the right. Any proportion at the right of the red line is higher than expected by chance. The size of the points is proportional to the number of genes associated with the annotation. The color of the point indicates the associated p-value, with the lowest p-values in red. Most significant GO terms are at the bottom.

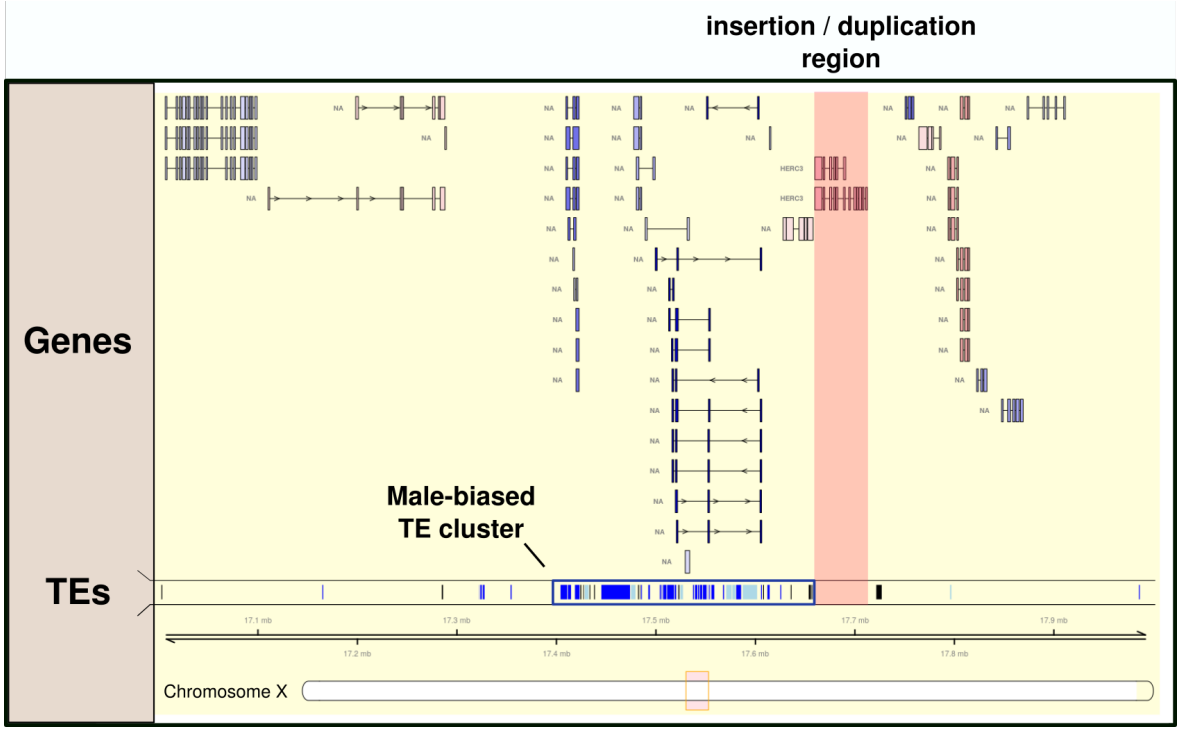

Figure 2.18 – **Genomic organization of the X chromosome region surrounding the Y-specific insertion carrying the *dmrt1by* master sex-determining gene of *O. latipes*.** The region that was duplicated on chromosome Y concomitantly with the insertion of the Y-specific region is framed in pink. A cluster of malebiased TEs is located close to the duplicated region on both X and Y chromosomes. The color code reflects the expression bias (blue: male-biased, red: female-biased).

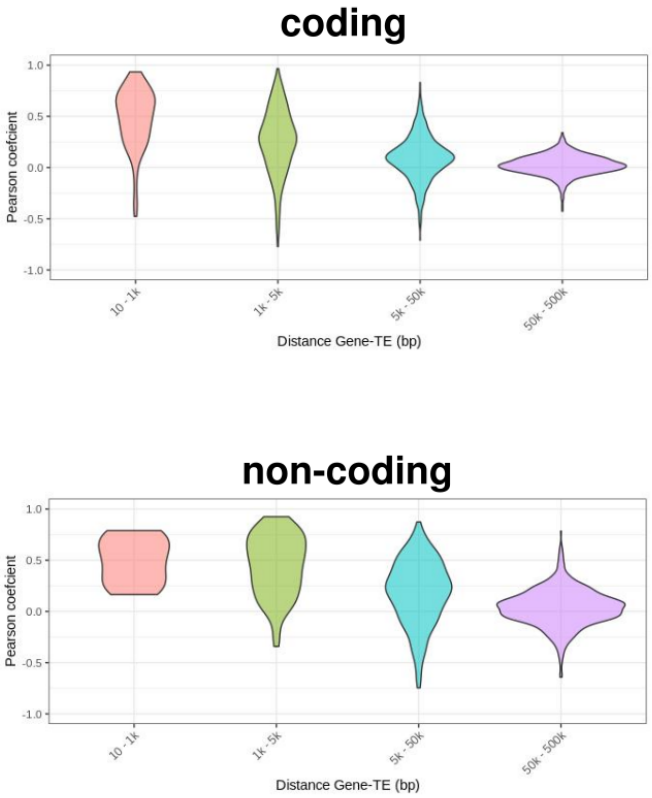

Figure 2.19 – **Gene-TE expression correlation using coding or non-coding genes only.**

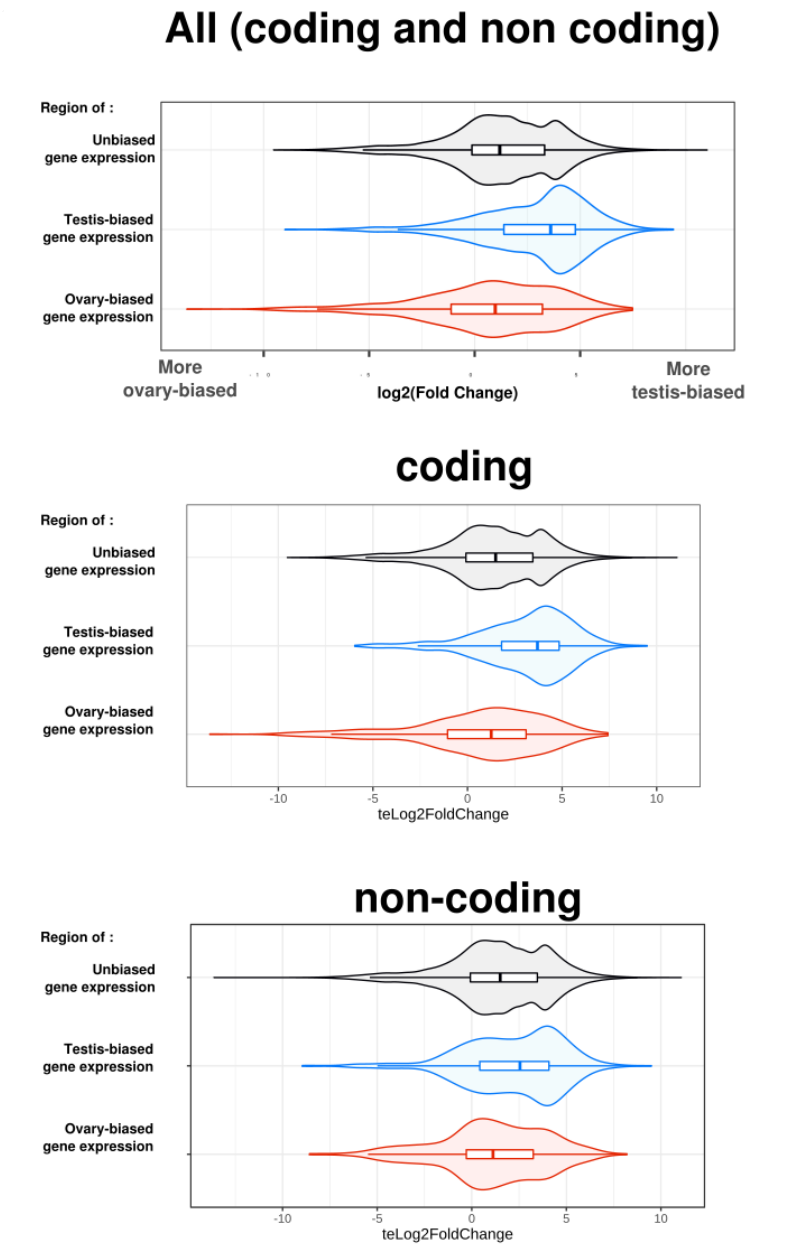

Figure 2.20 – **TE expression depending on their location.** Sex-biased TE expression is dependent from the genomic location of sex-biased gene clusters. Expression differential of TE copies depending on their location in the genome, using clusters made from coding genes or non-coding genes only, or from both coding and non-coding genes.

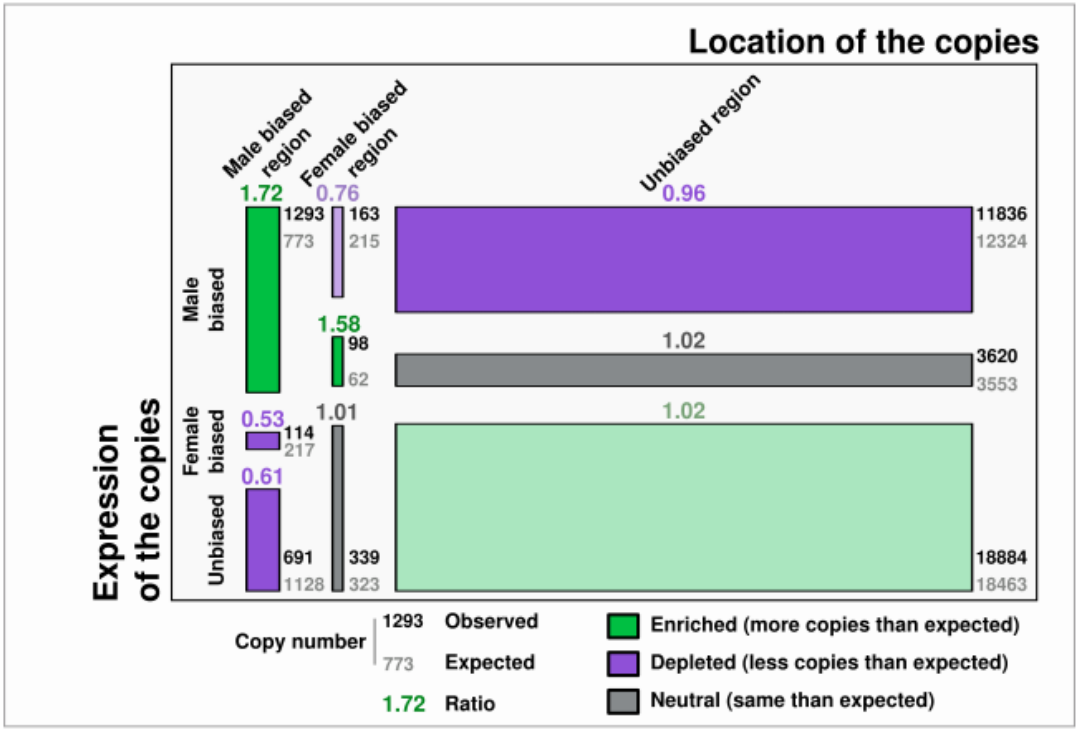

Figure 2.21 – **Mosaic plot representing the location and expression of TE copies.** Mosaic plot representing the number of observed sex-biased vs. non-biased TE copies being located in sex-biased vs non-biased regions (as assessed by gene expression). The surface of each rectangle corresponds to the number of TE copies in each type of region, indicated in black (observed). The width of the rectangles corresponds to the proportion of copies located in the different types of regions, while the height of the rectangles corresponds to the proportion of testis- vs. ovary-biased vs. unbiased copies located in each type of region. The expected number of copies calculated if there was no association between region and TE expression is indicated in grey (expected). When the number of copies is significantly higher than expected by chance (ratio > 1), the category is filled in green; when it is significantly lower, the category is filled in purple (ratio <1). As some copies are present in overlapping sexbiased regions, we selected the 37,038 copies unambiguously associated to a region (sum of the black values).

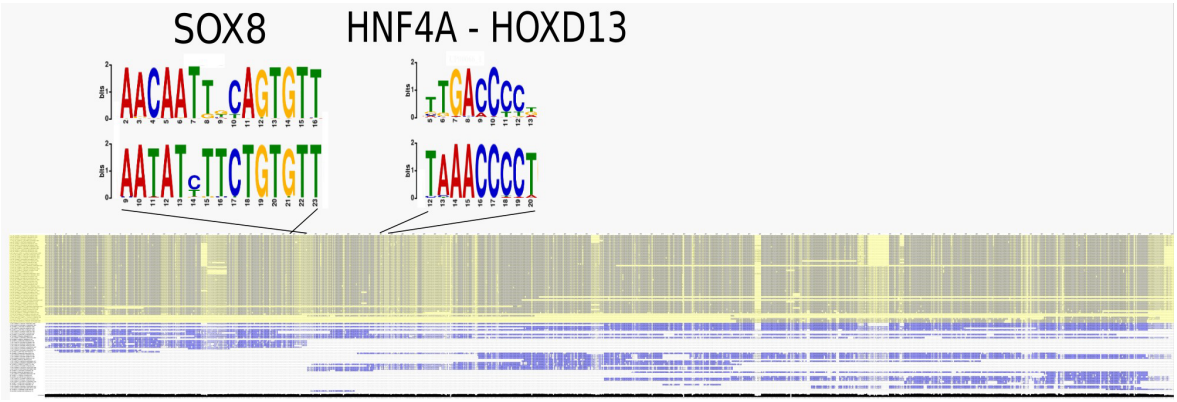

Figure 2.22 – **Alignment of copies from a candidate TE family.** rnd-5_family-992 TE copies alignment and localization of the predicted SOX8 and HOXD13 binding sites. Copies localized in male-biased clusters in O. latipes are framed in yellow.

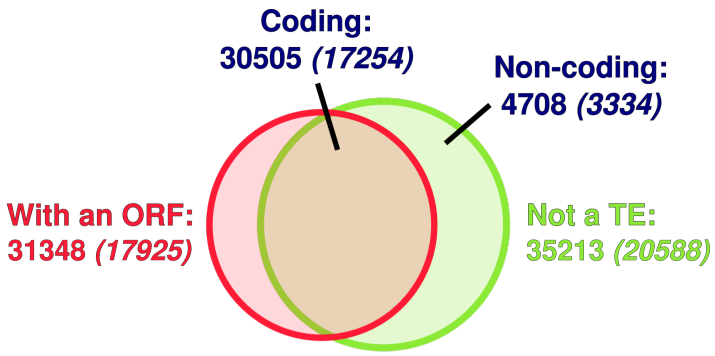

Figure 2.23 – **Venn diagram showing the number of transcripts and genes considered as coding and non-coding.** Venn diagram showing the number of transcripts and genes considered as coding and non-coding. A coding gene has at least one coding transcript with an ORF longer than 300 nucleotides. A non-coding gene has only non-coding transcripts.

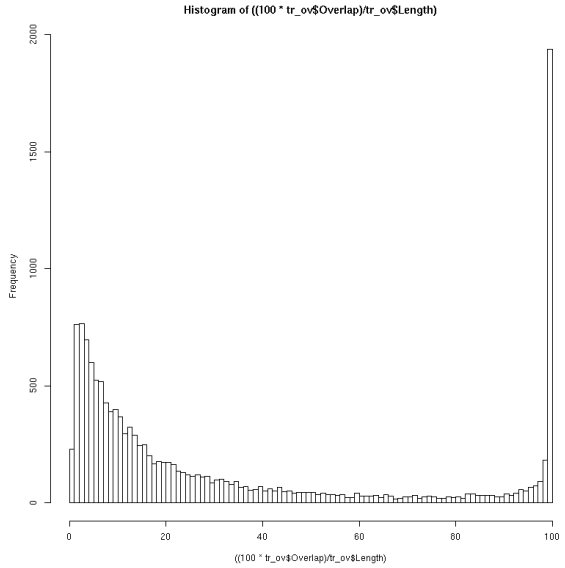

Figure 2.24 – **Percentage of the transcripts covered by a TE Percentage of the transcripts covered by a TE (% length) after removing transcripts that do not overlap a TE.**

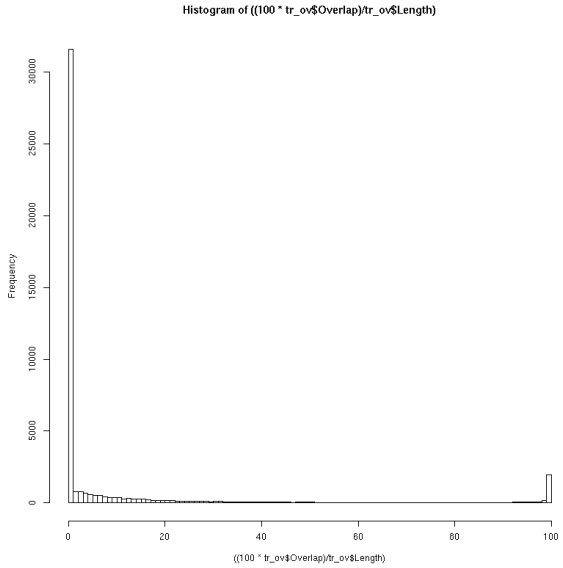

Figure 2.25 – **Percentage of the transcripts covered by a TE (% length) using all transcripts.**

3 Specific families of transposable elements are associated with sex-biased genes in the genome of the medaka fish

Corentin Dechaud[1], Anabel Martinez-Bengochea[2], Frederik Helmprobst[2], Manfred Schartl[2,3], Jean-Nicolas Volff[1], Magali Naville[1].

Corresponding author : Magali Naville - magali.naville@ens-lyon.fr

Affiliations :

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

2 : Entwicklungsbiochemie, Biozentrum, Universität Würzburg, Würzburg, Germany.

3 : The Xiphophorus Genetic Stock Center, Department of Chemistry and Biochemistry, Texas State University, San Marcos, TX, USA.

Sommaire

3.1 Avant-propos

Ce travail présente les résultats que j'ai obtenus à partir de données transcriptomiques issues de cinq espèces de poissons téléostéens : *Oryzias latipes*, *Oryzias luzonensis*, *Oryzias curvinotus*, *Oryzias javanicus* et *Xiphophorus maculatus*. Je recherche ici des exemples précis d'éléments transposables potentiellement impliqués dans la régulation de l'expression des gènes. Les principales conclusions de cette partie sont issues de prédictions bioinformatiques. Cependant des analyses fonctionnelles et expérimentales par CRISPR-Cas9 sont en cours dans l'équipe de Manfred Schartl à Würzburg (travaux de Frederik Helmprobst). Elles permettront de tester ces résultats préliminaires. Cette partie a été rédigée au format d'un article non définitif qui n'a pas encore été soumis, dans l'attente des derniers résultats.

3.2 Abstract

Teleost fish harbour a large diversity of sex determination mechanisms, which relies on a fast evolution of underlying genetic systems. As transposable elements (TEs) have been shown to be main actors in the evolution of regulatory networks, and as TE families are highly diverse in the clade, we hypothesized that they play a major role in the modulation of fish sex development pathways. To test this, we analysed gonadal transcriptomic data from four medaka species and a platyfish. We first identified sexually biased genes, and analysed the relative location of TE copies and genes with respect to their expression levels to estimate the global impact of TEs on gene expression. Our approach revealed 8 TE families significantly located near genes with sex-biased expression. We focused on a particular TE family that we coined *Finja* and that appeared associated to genes overexpressed in testis. Elements of this family harbour a binding site for the transcription factor RFX2, which is involved in spermatogenesis in mice. We show that *Finja* copies overlapping testis-biased gene 5'UTRs are more likely to harbour this binding site, which is more conserved than the rest of the TE sequence. Our approach is able to detect candidate TEs involved in the control of gene expression, and can be applied to other tissues or conditions. This study highlights TEs as a natural source of regulatory diversity for genomes that can impact fast evolving biological pathways such as sex in teleost fish.

3.3 Introduction

Teleost fish constitute the largest group of extant vertebrates (NELSON *et al.*, 2006) and harbour an important diversity. This diversity affects many traits of their development, morphology, physiology or behaviour, including sexual development and function (KOBAYASHI *et al.*, 2013; VOLFF *et al.*, 2007). Different types of reproduction exist in teleost fish, with hermaphrodites (two sexes in the same individual), parthenogenetic (clonal reproduction) or gonochoristic species (two sexes in separate individuals). In gonochoristic species, sex can be determined by environmental factors (water temperature or pH), or genetically (by one master sex-determining gene and/or by multiple minor genetic factors) (BACHTROG *et al.*, 2014). In this study we focus on teleost fish with genetic sex determination and that present a master sex-determining gene. Compared to mammals, in which *sry* is widely conserved and ancient (180-210My) (WATERS *et al.*, 2007), such a ubiquitous gene is not observed in teleost fish. Different sex determination genes are present in different species and are still unknown for most of them (KIKUCHI et HAMAGUCHI, 2013). In the genus *Oryzias* for instance, *dmrt1by* triggers male differentiation in *O. latipes* and *O. curvinotus* but is absent from all other *Oryzias* and teleost species (KIKUCHI et HAMAGUCHI, 2013). In contrast, the master sex determinating gene in *O. luzonensis* is *gsdfy* (KIKUCHI et HAMAGUCHI, 2013). These closely related species illustrate the large diversity of sexual differentiation and development that is observed in fish.

Teleost fish genomes also harbour a large diversity in terms of transposable element (TE) families, compared to other vertebrates (CHALOPIN *et al.*, 2015). TEs are sequences able to insert in the genome. They are often repeated and found in the genome of all animals analysed to date. TEs are classified depending on their transposition mechanism. Class 1 TEs, or retrotransposons, spread through a "copy-and-paste" mechanism involving the reverse transcription of an RNA intermediate, while class 2 TEs, or DNA transposons, mostly spread through a "cut-and-paste" mechanism without reverse transcription (WICKER *et al.*, 2007). Teleost fish harbour a genomic content of TEs of 10-50% that is similar to other vertebrates, but they present a larger diversity of TE families (CHALOPIN *et al.*, 2015). This diversity is a potential source of regulatory motifs for host genomes. Indeed, if the majority of TE insertions are neutral or deleterious for the host, some can also be selected for interesting functions. Examples are known of TEs that completely rewired gene regulatory networks, such as MER20 transposons in placental mammals that are found co-opted in a transcriptional network associated to pregnancy (LYNCH *et al.*, 2011, 2015; SUNDARAM *et al.*, 2017). Interestingly, the expression of *dmrt1by*, the master sex-determining gene of *O. latipes*, is controlled by a TE called *Izanagi*. The TE-derived promoter allows a tightly regulated expression of *dmrt1by* limited to a short period of time before hatching. *dmrt1by* resulted from a duplication of the former *dmrt1* gene that is on an autosome. This duplication gave rise to *dmrt1a* and *dmrt1b*, the last one became *dmrt1by* in *O. curvinotus* and *O. latipes*, and acquired *Izanagi* in *O. latipes*. The chromosome on which it was duplicated became the Y chromosome in both species (HERPIN *et al.*, 2010).

The example of *dmrt1by* concerns one gene only, but we speculate that TEs, with their ability to rapidly spread and bring new regulatory elements, could have participated more widely in the fast evolution of sexual regulatory networks. Furthermore, functional and evolutionary links

between sex and TEs have already been described in the literature (DECHAUD *et al.*, 2019). Previous observations on sexual development and TE diversity make the *Oryzias* genus a good model to study the potential impact of TEs on the control of sexual gene expression.

To address this question, we generated and analysed RNA-seq data from adult gonads of *O. latipes* and of four related species allowing us to detect sex-biased genes and genes that do not present conserved patterns of expression between species. We already described the expression correlation between TEs and close genes (See chapter 2), and some clues suggested that TEs were responsible for this correlation. Here we want to go further in assessing the impact of TEs on gene expression by the intermediate of *cis*-regulatory sequences they carry. We thus describe a new method to identify candidate TE families to be involved in the rewiring of gene regulatory networks based on these RNA-seq data and TE annotation. By analysing the relative localization of differentially expressed genes and TEs, we were able to identify TE families located close to or overlapping genes with male- or female-biased expression in the gonads. Using this approach we found different candidate TE families potentially involved in the control of expression of sex-biased genes. One of these TE families was found enriched in the 5'UTR of genes over-expressed in testis compared to ovary. Elements from this particular family, that we coined *Finja*, present a transcription factor binding site for RFX2, which is known to be involved in spermatogenesis in other species such as mouse (KISTLER *et al.*, 2015). Furthermore, some genes are testis-biased specifically in species where they are found associated to *Finja*.

3.4 Results

3.4.1 Similar proportion of genes are sex-biased in *O. latipes*, *O. luzonensis*, *O. curvinotus*, *O. javanicus* and *X. maculatus*

A differential expression analysis of the gonadal transcriptomic data of five species (*O. latipes*, *O. luzonensis*, *O. curvinotus*, *O. javanicus* and *X. maculatus*) revealed that a substantial part of genes are sex-biased in the gonads, for all of these species (**Fig. 3.1.**). Around 40% of genes are differentially expressed between male and female gonads, with the exception of *O. javanicus* where only 11.9% of genes are male-biased and 4.7% are female-biased. This lower proportion seems mainly due to one of the female samples being different from other samples (probably as a result from a sequencing problem) (supp. data). A common feature between species is the slightly higher number of male-biased genes compared to female-biased genes. The largest difference is observed in *X. maculatus* where 22.7% of genes are male-biased while 16.3% are female-biased (**Fig. 3.1.**). We also observed in all five species the presence of genes being specific to one sex, *i.e.* specifically expressed either in testis or in ovary (upper and lower diagonals in Fig. 3.1., 8% and 2% of testis- and ovary-biased genes in *O. latipes*, respectively). The expression of some genes known to be involved in sexual function (*dmrt1*, *gsdf*, *foxl2*, *amh*, *sox9b* and *aromatase*) was already estimated in *O. latipes* using RT-qPCR (KUROKAWA *et al.*, 2007). We looked at the expression of these genes in our data in all species. The expression bias in the four *Oryzias* species was the same for *gsdf*, *foxl2*, *sox9b* and *aromatase* (**Fig. 3.1.**, **Table 3.1.**). The expression of *dmrt1* is male-biased in *O. latipes* and *O. luzonensis*, but surprisingly very low (and thus not significantly different) in *O. curvinotus* and *O. javanicus*. *amh* presents a male-biased expression in *O. curvinotus* and *O. luzonensis* but is not

		O. latipes	***O. curvinotus***	***O. luzonensis***	***O. javanicus***	***X. maculatus***
dmrt1	Gene ID	MSTRG.9933	MSTRG.32815	MSTRG.13347	MSTRG.13417	MSTRG.12471
	LFC	5.22	-0.20	9.78	0.88	5.46
	BaseMean	1677	83	1542	78	5018
gsdf	Gene ID	MSTRG.13717	MSTRG.18136	MSTRG.697	MSTRG.18965	MSTRG.8698
	LFC	3.92	4.23	4.21	3.12	0.22
	BaseMean	50148	64098	52699	43838	12186
foxl2	Gene ID	MSTRG.15499	MSTRG.11615	MSTRG.12356	MSTRG.19710	MSTRG.18754
	LFC	-7.60	-7.89	-10.07	-6.26	0.16
	BaseMean	2014	1503	1754	717	1.95
amh	Gene ID	MSTRG.3964	MSTRG.7993	MSTRG.25587	MSTRG.5035	MSTRG.9603
	LFC	-0.06	2.48	1.59	0.41	1.31
	BaseMean	2339	4051	5420	522	2129
sox9b	Gene ID	MSTRG.9128	MSTRG.45510	MSTRG.22579	MSTRG.12552	MSTRG.17032
	LFC	4.75	2.64	6.10	2.13	0.92
	BaseMean	912	414	1274	646	3136
aromatase	Gene ID	MSTRG.2667	MSTRG.8004	MSTRG.2141	MSTRG.3725	MSTRG.4338
	LFC	-7.14	-9.52	-8.29	-6.78	-1.56
	BaseMean	3165	2007	3187	3332	2066

Table 3.1 – **Comparison of the expression of genes known to be involved in sexual function obtained from RNA-seq data in four *Oryzias* and one *Xiphophorus* species.** LFC = Log2 Fold Change: expression differential between males and females. If LFC is positive, the gene is male-biased; if LFC is negative, the gene is female-biased. The baseMean is the expression level in both tissues; it represents the intensity of the signal.

differentially expressed in *O. javanicus* and *O. latipes*.

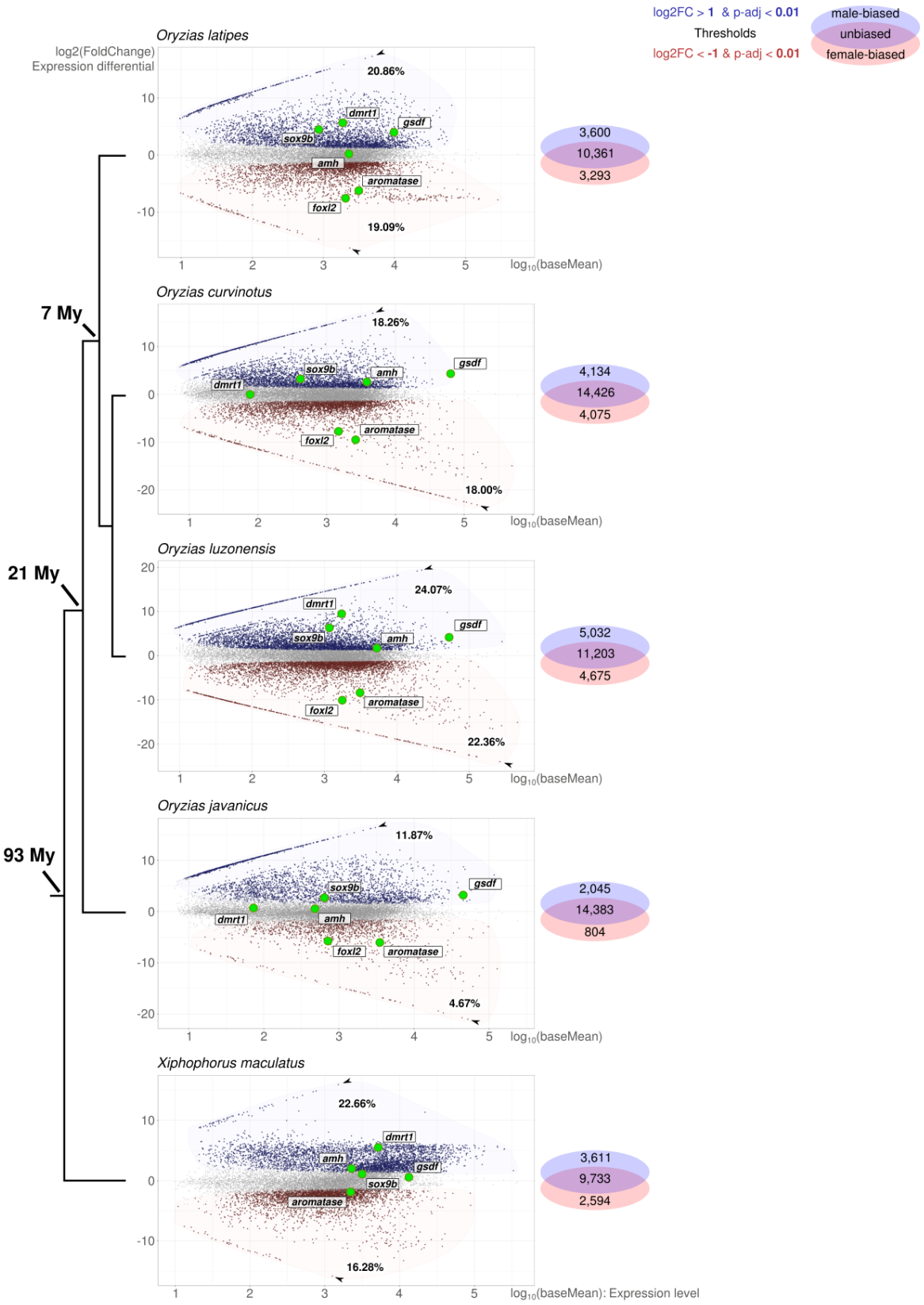

Figure 3.1 – **MAplots representing gene expression patterns of transcriptomic data from the gonads of four teleost fish species of the genus *Oryzias* and one *Xiphophorus* species.** In blue and red are represented genes significantly more expressed in testis and in ovary, respectively. We observed a similar proportion of differentially expressed genes in the different species except in *O. javanicus*. This might be due to one *O. javanicus* sample that presents a sequencing issue. In green are represented reference genes known to be involved in sexual function in fish. The black arrows show the upper and lower diagonals corresponding to genes specifically expressed in either testis or ovary.

3.4.2 Eight TE families are physically associated to genes with a sex-biased expression

We first aimed at detecting TE families physically associated to genes differentially expressed between male and female gonads in *O. latipes*. Even if regulatory sequences can be distant from the controlled gene, in a first attempt we tested associations between TE families and close genes. We looked for TEs overlapping the 500nt upstream region of the TSS (TSS -500), and within 500bp around the TSS (TSS +/- 500). We also divided genes in different parts: 5'UTR, CDS, 1st intron, other introns and 3'UTR (**Fig. 3.2.A.**). We distinguished first introns from subsequent ones as they are known to harbour transcriptional regulatory signals (PARK *et al.*, 2014). For each TE family, we counted how many genes contain a copy of the family in each particular region. For example, the TE family that is present at the higher number in 5'UTRs is present in 98 5'UTRs, while most TE families (580/1228) are never found in a 5'UTR region. As a TE family with a lot of copies is expected to have more copies located in 5'UTR or in other gene regions just by chance, we normalized the number of 5'UTR or other gene regions containing a TE from a particular family by the total number of copies in the family. We call this normalized value the "prevalence" of a TE family in a given type of region. The relation between the prevalence and the absolute count was calculated for each TE family and each gene region (**Fig. 3.2.B.**). For example, the TE family *Olat_r-5_f-2836* is present in the coding sequence of 86 genes, representing around 22% of all copies from this family. More globally we observed that TEs tend to be more often inserted in first introns than in other gene regions, maybe due to insertions preferences, selective pressures or larger size of first introns compared to other gene regions.

We then looked at the expression bias of genes containing a particular TE family (**Fig. 3.2.C.**). Using only TE families located at least in 5 genes we generated ternary plots. Each data point on such plots represents a TE family. On the different axes is shown the percentage of male-biased, female-biased and unbiased genes associated to this TE family. The more the TE family is on the bottom left of the plot, the more male-biased are the genes associated to it. The more the TE family is located on the bottom right, the more female-biased are its associated genes. Using a χ-square test, we identified TE families with significant association to a particular sex-bias of genes. For example, in 5'UTR, 4 TE families are significantly associated to a sex-biased expression of genes (2 associated to female-biased expression, 2 associated to male-biased expression) (**Fig. 3.2.C.** in blue). TE families associated to an expression bias are mainly located in 5'UTR and 3'UTR regions (7/8). Two TE families, when localized in first introns, are significantly associated to genes without a sex-biased expression. These TE families maybe control the expression in both organs at the same level. In TSS-500 regions only one TE family presents a significant association, but with a small size effect. All in all, our method revealed different candidate TE families such as *Finja*, found associated to a testis-biased expression in 5'UTR (**Fig. 3.2.C.**), and that will be further investigated.

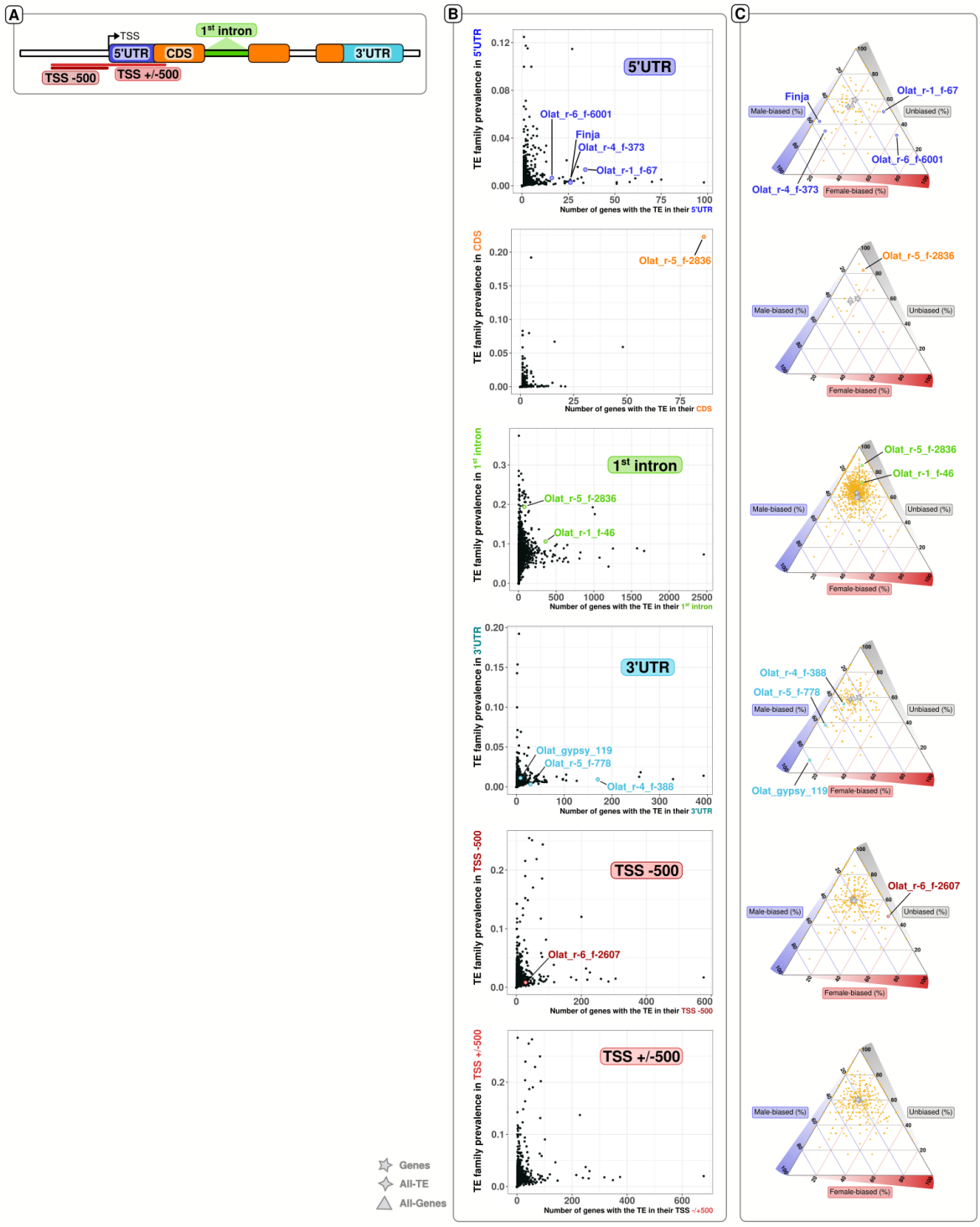

Figure 3.2 – **Some TE families are associated to genes with sex-biased expression.** A. Schematic representation of a gene divided in different parts used to evaluate gene-TE associations. B. Scatter plots corresponding to each type of gene region. Each data point corresponds to a TE family. The X-axis represents the number of genes physically associated to each TE family. The Y-axis represents the prevalence of the TE family in a particular gene region. The prevalence corresponds to the number of genes physically associated to the TE family normalized by the total copy number of the family. C. Ternary plots showing association between gene sex-biased expression and TEs. Each data point represents a TE family. Each axis represents the percentage of sex-biased or unbiased genes that present a particular TE-derived sequence. TE families indicated with text are significantly associated to a sex-bias in gene expression. In grey, the *Genes* point represents the genomic mean, *i.e.* the expression of all genes. The *All-Genes* data point represents the mean of all genes that contain any TE-derived sequence, and *All-TE*, the mean of all TE families. These data points allow to compare what is expected if a TE family is distributed randomly and does not impact gene expression.

3.4.3 Several *Finja* copies are localized in 5'UTR region of genes overexpressed in medaka testes

Finja copies are found in the 5'UTR region of 26 genes, among which 14 (54%) are overexpressed in the testes of *O. latipes*, while we would expect 5.2 testis-biased genes at random (**Fig. 3.2.C.**). Among the 12 remaining genes, 11 are not sex-biased and 1 is ovary-biased (**Fig. 3.2.C.**). One of the testis-biased genes with *Finja* in its 5'UTR region is *p2rx1*, which encodes a calcium ion channel involved in platelet function and male fertility in mice. *Finja* is located in the first exon of *p2rx1* at the junction with the first intron (**Fig. 3.3.**). As observed on the MAplot (**Fig. 3.3.**), *p2rx1* is among the most testis-biased genes in *O. latipes*, and qPCR (**Fig. 3.4.**) revealed that *p2rx1* expression is restricted to testis in *O. latipes*. We investigated the expression of p2rx1 in three other *Oryzias* and a *Xiphophorus* species as well. We found that in species where *Finja* is present close to the 5'UTR of *p2rx1*, *i.e.* in *O. curvinotus* and *O. luzonensis*, *p2rx1* is also testis-biased. In contrast, in species where *Finja* is absent, *i.e.* in *O. javanicus* and *X. maculatus*, *p2rx1* is not testis-biased. The presence of *Finja* near the 5'UTR region of *p2rx1* is thus associated to a male-biased expression pattern. In *O. javanicus Finja* is absent in the 5'UTR region of *p2rx1*, while it is present in the genome at other genomic locations. In *X. maculatus* in contrast, *Finja* is completely absent from the genome. We further investigated if such observation of the presence of *Finja* associated to a male-biased expression pattern was valid for the 25 other genes of *O. latipes* containing *Finja* in their 5'UTR region (supp. data). Some genes like *ttbk2*, coding for the tau-tubulin kinase 2, were globally not differentially expressed in *O. latipes* despite the presence of *Finja* in their 5'UTR region. However, looking at isoform-specific expression, we found that transcripts with *Finja* in their 5'UTR region were male-biased, while other transcripts missing *Finja* were not. We similarly compared the isoform-specific expression data of all 26 genes and the interspecific insertion polymorphism of *Finja* in the 5 species, and found that *Finja* is specifically associated to testis-biased isoforms in all species for 18 of the 26 genes (supp. data).

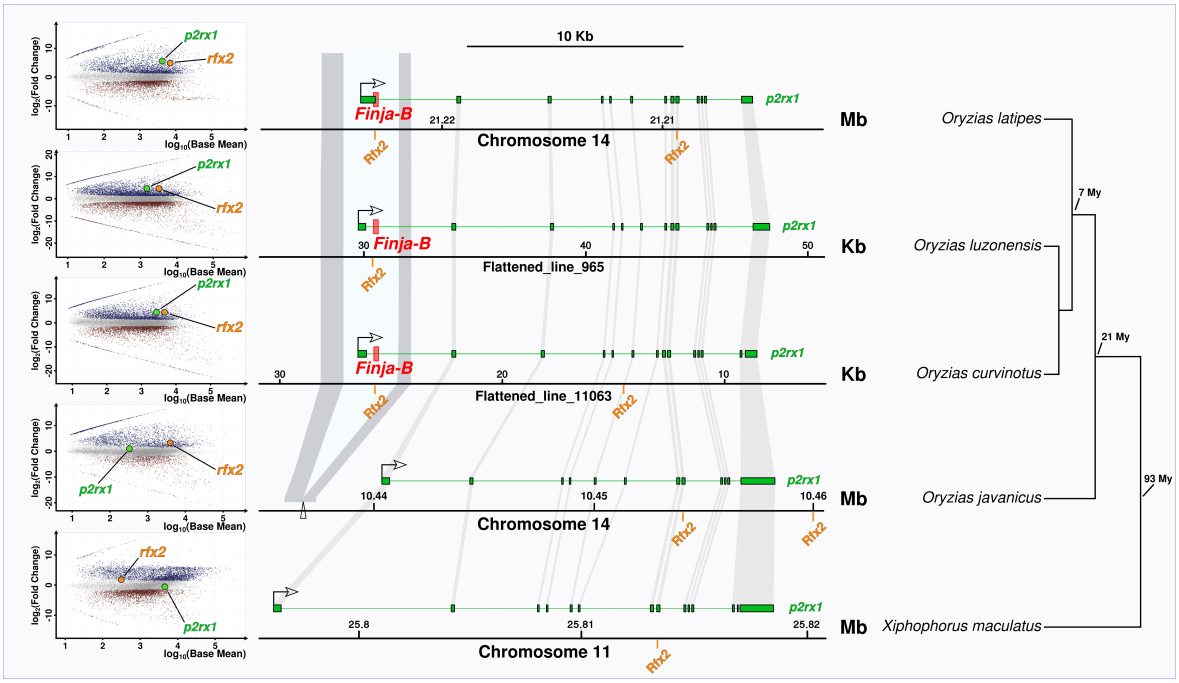

Figure 3.3 – ***p2rx1* male-biased expression correlates with the presence of a *Finja* transposable element carrying a RFX2 transcription factor binding site in the 5'UTR region.** The exonic structure of *p2rx1* (in green) in *Oryzias* and *X. maculatus* is similar. In *O. latipes*, *O. curvinotus* and *O. luzonensis*, a TE, *Finja*, is present at the 3' border of the first exon (5'UTR). It is absent from this locus in other species, along with the first exon. *Finja* contains a predicted TFBS for RFX2. The expression of *p2rx1* correlates with the presence of the TFBS in the vicinity of the first exon in these species. Expression of *p2rx1* and *rfx2* is represented on the MAplot on the left.

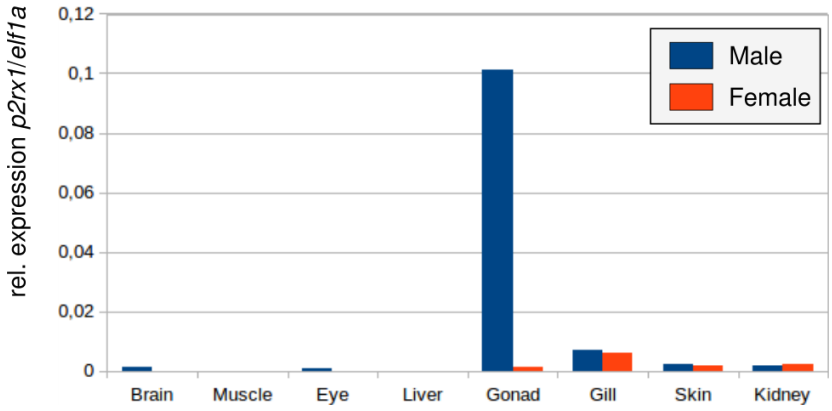

Figure 3.4 – ***p2rx1* is specifically expressed in testes.** The relative expression of *p2rx1* compared to *elf1a* was estimated by qPCR in different organs of *O. latipes* for each sex.

3.4.4 *Finja* is a repeated sequence of 2600bp harbouring a putative binding site for a transcription factor involved in spermatogenesis

We next characterized *Finja* elements. The longest *Finja* copies (2300-2600 nt) present a mean pairwise sequence similarity of 94.74%, while the 26 copies located in 5'UTR regions are shorter (between 100 and 1500 nt) (**Fig. 3.5.A.**) and present a lower pairwise sequence similarity (68.71%). We were able to detect a putative TFBS on the long *Finja* copies for RFX2, NR2F6 and ZNF410 (**Fig. 3.5.A.** and **B.**). In human, RFX2 acts as a key regulator of spermatogenesis

(https://www.uniprot.org/uniprot/P48378), NR2F6 is predominantly involved in transcriptional repression of hormonal genes (https://www.genecards.org/cgi-bin/carddisp.pl?gene=NR2F6), and ZNF410 activates the transcription of matrix remodelling genes (https://www.genecards.org/cgi-bin/carddisp.pl?gene=ZNF410).

We tried to classify *Finja* in a transposable element group but we were not able to detect any open-reading-frame (ORF), pol III-binding sites (a feature of SINE elements), terminal inverted repeats (a feature of DNA transposons), or long terminal repeats (a feature of LTR retrotransposons). The non-coding nature of *Finja* suggests that it is not autonomous and might be mobilized by other TEs. We also self-aligned *Finja* and detected at the 5' end of *Finja* a satellite region (**Fig. 3.5.C.**) of 100bp repeated 0 to 8 times depending on the copy. We also found a tandemly repeated region of 600 nt in the central part of the element (grey boxes, **Fig. 3.5.A.**).

Interestingly, *rfx2*, for which product a putative TFBS is found on *Finja*, is highly testis-biased in *O. latipes*, supporting the potential role of this TFBS in the control of male-biased expression by *Finja* (**Fig. 3.3.**). *znf410* and *nr2f6* on the contrary are not differentially expressed between male and female gonads. Furthermore, *Finja* copies located in the 5'UTR region of genes harbour the region containing the TFBS for RFX2 (**Fig. 3.5.A.**). This observation also allows to hypothesize that *Finja* promotes male-biased expression in gonads through the presence of this TFBS (**Fig. 3.3.**).

We then asked if the sub-region of *Finja* containing RFX2 TFBS was particularly retained in 5'UTR *Finja* copies. To do so we first evaluated the structure of the different *Finja* copies in the genome by computing the coverage of the *Finja* element by its different copies along its whole length (**Fig. 3.5.D.**, black curves). We observed two coverage peaks at the end of the tandemly duplicated regions (**Fig. 3.5.D.**, black arrows), with at the border of the second one the TFBS for ZNF410. This region corresponds to a satellite that is present at other genomic loci, independently of *Finja*, thus creating an artificial up-representation of this region of the consensus. We also represented the consensus coverage using only *Finja* copies present in gene 5'UTRs (**Fig. 3.5.D.**, blue curves). The most represented region for these specific copies is the one containing the TFBS for RFX2 (22/26 copies, 85%), while only 60% of genomic *Finjas* cover it, quantifying what is observed in Fig. 3.5.D. Since they are similar, both tandem repeats appear covered by the 5'UTR *Finja* copies, giving the impression of elements longer than what they are in reality; this is only an artefact due to these internal repeats. We further used a bootstrap approach to assess if the observed enrichment of particular subregions in 5'UTR *Finja* could be due to chance. To do so, we randomly sampled the same number of copies and aligned them to the consensus. After 1000 random samplings, we plotted at each position of the *Finja* consensus the number of bootstraps for which the coverage was higher than the one observed using *Finja* from 5'UTRs (**Fig. 3.5.D.**, green curve). Again, the region containing the TFBS for RFX2 was more significantly retained in 5'UTRs than the rest of the TE. These results strongly suggest that the presence of the TFBS has been selected and retained in the proximity of genes, allowing to hypothesize that *Finja* brought a new regulation leading to a male-biased expression of the gene, a regulation that might have been positively selected as beneficial for the host.

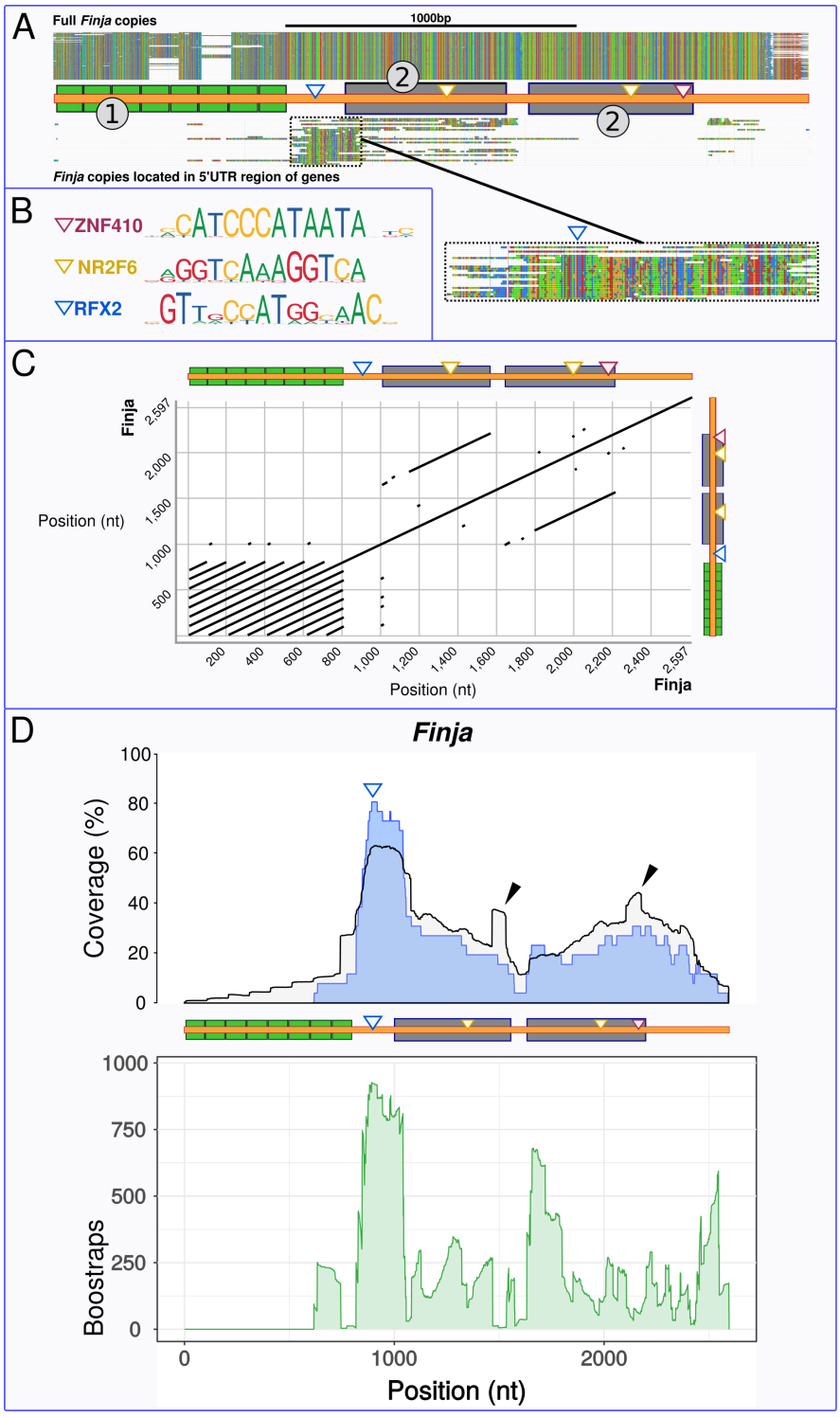

Figure 3.5 – **RFX2 transcription factor binding site has been selected and retained in *Finja* copies from 5'UTR gene regions.** A. Sequence alignment of the longest genomic *Finja* copies and copies located in the 5'UTR region of genes. Each colour corresponds to a given nucleotide. Triangles correspond to the different predicted TFBS. The copies located in the 5'UTR region of genes contain the region harboring the TFBS for RFX2. Green and grey colour boxes are described in the C part. B. Logo sequences of the TFBS for RFX2, ZNF410 and NR2F6 retrieved from Jaspar database. C. *Blastn* self-alignment of *Finja* showing its structure with a satellite region at one extremity (panel A, 1, green boxes), and a tandem repeat in the middle (panel A, 2, grey boxes). D. Bootstrap approach to detect regions overrepresented in *Finja* copies located in gene 5'UTRs compared to all genomic *Finja* copies. In black is shown the coverage on the consensus of *Finja* using all genomic copies. A peak is observed at the right of the tandem repeats (grey boxes) that corresponds to a satellite found in multiple copies in the genome. The most represented region is the region containing the RFX2 TFBS (blue triangle). In blue is shown the coverage measured taking only TE copies that are found inside gene 5'UTRs. To test if the enrichment in the RFX2 binding site region was likely to be observed by chance, we randomly sampled 26 TE copies 1000 times and computed the same coverage. The green curve represents the number of times (between 0 and 1000) where the observed coverage (using TE copies in gene 5'UTRs) is higher than the coverage calculated by random sampling (from all *Finja* copies). The green peaks thus represent regions enriched in gene 5'UTRs. The most enriched regions in 5'UTR contain the predicted RFX2 binding site.

3.4.5 *Finja* presents interspecific polymorphic insertions

Finja is a putative transposable element that we were able to detect in *Oryzias* genomes only. To address the possibility of a recent activity of the element, we searched for potential insertion polymorphisms between *Oryzias* species. As the best genome assemblies in *Oryzias* are from *O. latipes* and *O. javanicus*, we compared both genomes to detect such insertions. *Blast* comparisons allowed to identify four *Finja* insertions present in *O. latipes* and absent from *O. javanicus* at the same genomic loci (**Fig. 3.6**). Analysing the insertion junctions, we detected a target site duplication (TSD) of 7 nucleotides (**Fig. 3.6**) flanking three of the four insertions. The TSD sequences are conserved and identical between the polymorphic insertions found on chromosome 11 and 9 (ATAATTA), while some points mutations affected the TSDs of the chromosome 14 polymorphic insertion (ATAATTA / ATAATAA, **Fig. 3.6**). The presence of interspecific polymorphic insertions indicates an activity and transposition of *Finja* after the *O. latipes* / *O. javanicus* speciation. What is more, the polymorphic insertions from chromosome 9 and 14 correspond to three and two tandem *Finja* insertions, respectively, forming a satellite-like structure. A TSD sequence is also present at the junction of each of these tandem insertions. This configuration suggests that *Finja* could serve as a substrate for the formation of satellite regions.

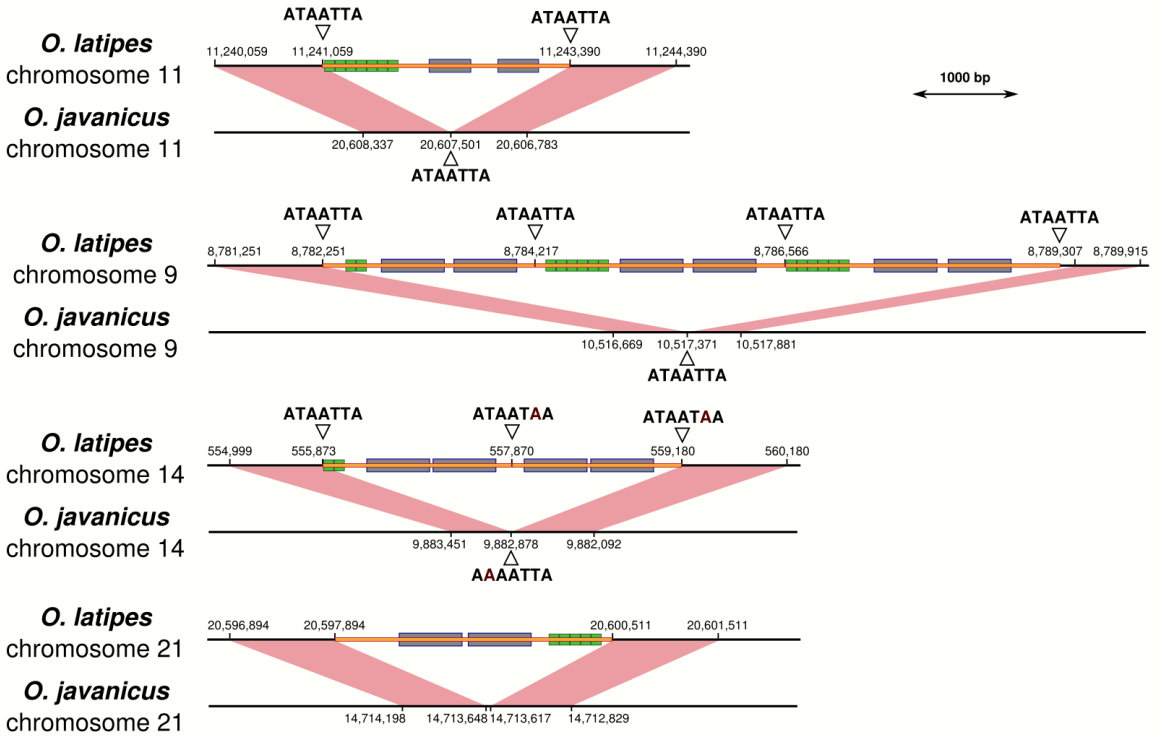

Figure 3.6 – ***Finja* polymorphic insertions between *O. latipes* and *O. javanicus*.** Polymorphic insertions of *Finja* were found inserted in *O. latipes* and absent from the same locus in *O. javanicus*. Orthologous flanking genomic regions are shown in red. The site of integration, present as a single copy in *O. javanicus*, has been duplicated and flanks the insertions in *O. latipes* (Target Site Dulplication, TSD). In chromosome 9 and 14, multiple sequential *Finja* insertions are found in tandem and separated by the TSD.

3.5 Discussion

3.5.1 The proportion of sex-biased genes are similar among *Oryzias* species

Sex determination mechanisms are variable among the five species studied. *O. latipes*, *O. luzonensis*, and *O. curvinotus* present a XY-XX determination system, with either *dmrt1by* (*O. latipes* and *O. curvinotus*) or *gsdfy* (*O. luzonensis*) as master sex-determining gene, whereas *O. javanicus* uses a ZZ-ZW sex determination system (KIKUCHI et HAMAGUCHI, 2013). The master sex-determining gene in *O. javanicus* is still unknown. It has been previously suggested that sex chromosomes in *Oryzias* repeatedly evolved from autosomes as sexual chromosomes are not homologous between some closely related *Oryzias* species (TAKEHANA *et al.*, 2008). In *X. maculatus*, three sexual chromosomes exist: X, Y, and W. In our data, females are either WY of XX, while males are YY or XY (SCHULTHEIS *et al.*, 2009). These differences illustrate the great diversity existing in teleost fish in term of sexual determination and differentiation. In addition, *gsdf* is involved in male sex maintenance in *O. latipes* (ZHANG *et al.*, 2016), while in *O. luzonensis* it is the master sex-determining gene, which expression is restricted to sex-determination (MYOSHO *et al.*, 2012), also illustrating the variability of the gonad function in adults. It is thus expected to observe differences in adult gonadal gene expression between species. A conserved feature of sex-biased genes is their fast evolution, as described in different lineages such as Drosophila (ASSIS *et al.*, 2012), *C. elegans* (CUTTER et WARD, 2005), fish (YANG *et al.*, 2016) and primates (KHAITOVICH *et al.*, 2005). Many other factors can influence the number of genes detected as differentially expressed between sexes, such as the maturity of the gonad sequenced, or the type of algorithms and parameters used. To limit the number of biases possibly influencing the results, the same pipeline was used to generate gene annotations and analyse their expression in the five species. This allowed us to compare the percentages of differentially expressed genes between species. Gonads and brain are the most studied organs in teleost fish sex-biased transcriptome studies, and gonads harbour a strongly larger number of sex-biased genes (between 4% and 66% depending on the species and the study) compared to brain, where 0% to 8% of genes are found sex-biased (BAR *et al.*, 2016; BEAL *et al.*, 2017; BÖHNE *et al.*, 2014; LIU *et al.*, 2015; ROBLEDO *et al.*, 2015; SAARISTO *et al.*, 2017; SHEN *et al.*, 2020; TAO *et al.*, 2018; TSAKOGIANNIS *et al.*, 2018a; WANG *et al.*, 2017; WU *et al.*, 2019; ZENG *et al.*, 2016). Our result of around 40% differentially expressed genes between male and female gonads corresponds to what is observed in other fish species, even if all biases detailed before can influence this number. We investigated the expression of six reference genes known to be involved in male or female sexual development in *O. latipes*, and compared our result between species and with what is known from previously published RT-qPCR experiments (KUROKAWA *et al.*, 2007). For 4 out of 6 genes, despite the differences in term of sex determination between the *Oryzias*, the results were similar. *dmrt1* shows different expression biases, and has a very low expression in *O. curvinotus* and *O. javanicus*. In *O. latipes*, we know that *dmrt1* expression is related to the silencing of *dmrt1by* in adults, mediated by the presence of *izanagi*. As this regulation of *dmrt1by* is absent in *O. curvinotus*, it could explain why we do not observe any expression of *dmrt1* in this species. In *O. javanicus*, *dmrt1by* is absent as sex is determined by a ZZ-ZW system, which could also explain the absence of sex-biased expression of *dmrt1*.

3.5.2 Particular TEs are physically associated to sex-biased gene expression in *O. latipes* gonads

Somatic expression and transposition of TEs is not sufficient for them to be transmitted to the next generation. TEs spreading and fixation in genomes indeed require germline expression (DECHAUD *et al.*, 2019). Furthermore, TEs are known to carry their own regulatory sequences that allow their expression and transposition (REBOLLO *et al.*, 2012b), such as TFBS, POL II and POL III binding sites, splicing signals and polyA sites. In mammals, TEs were shown to harbour a large proportion of TFBS that actively participate to their spreading in genomes (SUNDARAM *et al.*, 2014). Selection thus favours TEs that carry regulatory sequences allowing a germline expression: TEs present in extant genomes must have transposed in the germline of previous generations. Here we wanted to assess the impact of TEs on the expression of sexual genes in *O. latipes* gonads. As it is not possible to dissect germline from soma in the gonads of the species we sequenced, the results should be interpreted as gonadal and not germline expression data.

We managed to detect several candidate TE families physically associated to sex-biased genes; these families should be further investigated. To do so we looked for association between TEs and six gene regions, namely 5'UTR, CDS, 1st intron, 3'UTR, 500bp TSS upstream region, and TSS +/- 500bp. We also tested other regions (data not shown) such as subsequent introns (other than first intron) and different distances to TSS (1kb, 5kb, 10kb or 20kb), but do not show the results as they did not reveal any additional candidate TE families. Gene regulation can be effective at a longer range, but this type of association is harder to investigate, as many possible gene-TE associations are taken into account and dilute the true functional signal. Furthermore, a given TE could be involved in controlling gene expression at long range for one gene, but at shorter distance for other genes, making its identification harder because it will not be enriched in any gene region. Our method has thus some limits. Detecting enhancers at longer range would require other data, such as Hi-C sequencing, which allows to retrieve DNA regions interacting with each others. Another caveat of our approach is that a controlling TE at very few copies will not be statistically detected due to lack or statistical power. Thus, it could be interesting to investigate some TE families that do not show a statistical association to a bias maybe because of their low copy number, but for which all copies are associated to such a bias. Also, some TE families can regroup both insertions carrying a regulatory sequence and others that do not, which would hinder their identification. In human, the potential of TEs to control gene expression has been largely investigated. Exaptation of TEs in regulatory regions is probably rare and more likely to happen for old TEs (SIMONTI *et al.*, 2017), even if younger TEs are more likely to be co-opted in gonads than in other organs (SIMONTI *et al.*, 2017). The small number of TE families detected by our method (8 families associated to sex-biased gene expression) could reflect (in addition to a possibly low resolutive power of our approach) such scarcity in the recruitment of TE families. We observed that the number of TE copies located in different gene regions is very different according to the region. CDS present the lowest number of TE-derived sequences, and first introns contain the most, which can not be explained by a different global size only, as CDS cover around 27Mb of the genome, and introns 1 66Mb. This corresponds to the observation that TEs disrupt gene function by inserting into coding exons. Similar results were observed in other species (human, mouse and zebrafish), where CDS were shown to contain few TE-derived sequences (KAPUSTA *et al.*, 2013). These analyses also showed that 5'UTR contain

less TEs than 3'UTR. It is proposed that TEs are preferentially acquired at the 3' end of transcripts because the region is more permissive to their potential exonisation (KAPUSTA *et al.*, 2013). We did not perform the analysis on the other *Oryzias* species as their genome assembly is of lower quality and would not allow to properly match genes and TEs. In the previous chapter we described clusters of sex-biased genes in *O. latipes*. The TE families or genes that we find potentially regulated are not located in these gene clusters. This suggests that the candidates that we described here are not at the origin of the sex-biased gene clusters.

3.5.3 *Finja* might serve as a « taxi » that spreads binding sites for RFX2 transcription factor

Finja was one of the candidate TEs detected by our method. It retained our attention as it carries a TFBS for RFX2, a transcription factor involved in spermatogenesis in mice (KISTLER *et al.*, 2009; WU *et al.*, 2016). We compared the presence and absence of *Finja* in the 5'UTR region of genes in different species. Different clues support the control of gene expression by *Finja* through the presence of RFX2 TFBS. First, the gene *p2rx1*, coding a calcium ion channel involved in platelet function and male fertility in mice, is overexpressed in testis only in species where *Finja* is present in its 5'UTR. In species without *Finja*, the TFBS is also absent. Second, *rfx2* is highly testis-biased in the five species investigated. Third, we observe the selection of the region of *Finja* containing the TFBS for RFX2 in copies located in the 5'UTR of genes. Our current model is that *Finja* carries different TFBS, including one for RFX2, and spreads in the genome. Some genes became overexpressed in testis after the insertion of *Finja* as it is observed in *O. latipes*, *O. curvinotus* and *O. luzonensis* through the recruitment of RFX2. It probably happened in the common ancestor of these three species after the divergence with *O. javanicus*, in which *Finja* is absent at this locus. This suggests that interspecific polymorphic insertions of TEs induce differences in the regulation of gene expression between closely related species. It also supports the role of TEs in the evolution of fast evolving pathways such as sex. We retrieved ChIP-seq data from mice RFX2, and compared the location of the binding regions with the annotation of TEs (data not shown), but we were not able to detect any enrichment of particular TE families in RFX2 regions in this species. This suggests that the spreading of RFX2 TFBS by a TE could be specific to the *Oryzias* lineage. The TE possibly brought an evolutionary novelty and served as a "taxi" for a regulatory sequence driving male-biased expression. To further test our hypothesis, we are now generating CRISPR-cas9 fish with a deletion of *Finja* upstream of *p2rx1* to check if its expression is silenced or equivalent in both gonads. We will generate ChIP-seq data for RFX2 to reveal all its binding sites in the genome, and confirm its binding on *Finja* upstream of *p2rx1*. In a previous article (chapter 2), we showed that sex-biased genes and sex-biased TEs are organised in clusters across chromosomes in *O. latipes*. We also should check if the sex-biased genes we find potentially regulated by TEs have a known function related to gonads in other organisms. In teleost fish TEs could be involved in the birth of new regulations, the replacement of existing regulations, or the reinforcements of existing regulations. To discriminate which scenario is true for the genes we studied, and better understand the evolution of gene expression regulation, it would be important to look at gonadal expression of orthologous genes, potentially controlled by TEs in *O. latipes*, in more distant species.

3.6 Methods

Genome and transposable element annotation

O. latipes: ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/002/234/675/GCF_002234675.1_ASM223467v1/GCF_002234675.1_ASM223467v1_genomic.fna.gz

O. javanicus (TAKEHANA *et al.*, 2020): ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/003/999/625/GCA_003999625.1_OJAV_1.1/GCA_003999625.1_OJAV_1.1_genomic.fna.gz

X. maculatus: ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/002/775/205/GCF_002775205.1_X_maculatus-5.0-male/GCF_002775205.1_X_maculatus-5.0-male_genomic.fna.gz

O. curvinotus: Unpublished genome

O. luzonensis: Unpublished genome

The transposable elements annotation was generated on *O. latipes* genome only (See chapter 2 for methods). In other species, we only looked for TEs locally using *Blast* approaches.

Gene expression analysis

The exact same approach as in chapter 2 (*NewTuxedo*, see chapter 2) was used for all species.

Gene correspondence between species

As we generated independent gene annotations based on RNA-seq data for each species, we were not able to compare identical gene sets between species. Hence, we used a "Reciprocal Best Hit" (RBH) approach (MORENO-HAGELSIEB et LATIMER, 2008) to define probable orthologs. Each gene from *O. latipes* was blasted against the genes of other species and *vice versa*. Each species was compared to *O. latipes*, and for each comparison we retrieved the genes with a reciprocal best blast hit. This is a very conservative approach. For some genes, we were not able to retrieve a RBH in all species. We generated a conversion table that allowed us to retrieve a given gene in other species. The *Blast* command lines used are the following :

```
blastn -query $OjavGenes -db Olatipes/$LatipesDB -out java_vs_latipes.blasted
-evalue 1e-10 -outfmt "6 qseqid sseqid pident mismatch gapopen qstart qend sstart
send evalue bitscore qcovs qcovhsp qlen slen length" -num_threads 8
blastn -query $OlatGenes -db Ojav/$OjavDB -out latipes_vs_java.blasted -evalue
1e-10 -outfmt "6 qseqid sseqid pident mismatch gapopen qstart qend sstart send
evalue bitscore qcovs qcovhsp qlen slen length" -num_threads 8
```

Gene annotation

Using the reference genome annotation of *O. latipes* we were able to associate each gene generated using our RNA-seq data to a gene of the reference annotation (See chapter 2 for methods). We then used the conversion table previously generated to find the corresponding gene in other species.

Ternary plots

To find TE families potentially involved in sex-biased expression of genes, we generated ternary plots for different types of gene regions: 5'UTR, CDS, 3'UTR, 1st intron, TSS-500 and TSS+/-500. The *NewTuxedo* approach creates a genome annotation for each species based on the RNA-seq data. From this annotation we generated files containing the 5'UTR, CDS, 3'UTR, 1st intron, TSS-500 and TSS+/-500 regions for each gene. Using *bedtools* intersect we overlapped each TE copy with these region files, and retrieved the genes overlapping with each TE family. Then we generated both ternary plots using the R package *ggtern* (HAMILTON et FERRY, 2018), and the histograms using *ggplot*. The script used to generate these figures is available on the gitlab (*Cross_all.sh*, see below).

Conservation of *Finja* copies

We aligned the longest *Finja* TE copies to look at sequence conservation. We retrieved all copies from RepeatMasker with a length between 2300 and 2600 nucleotides. We performed the alignment with *MAFFT* (v7.419) (https://mafft.cbrc.jp/alignment/software) (KATOH, 2002; KATOH et STANDLEY, 2013) using default parameters. We manually curated the alignments to remove non-informative regions. Visualisation was performed using *Jalview* (www.jalview.org).

Transcription factor binding site detection

We looked for TFBS in *Finja* using AME online from MEME suite using default parameters (http://meme-suite.org/tools/ame) (MCLEAY et BAILEY, 2010). The TFBS detected by AME were downloaded in Transfac format and localized on the consensi using RSAT matrix scan online (http://rsat.sb-roscoff.fr/matrix-scan-quick_form.cgi) (TURATSINZE *et al.*, 2008).

Structure of different *Finja* copies in the genome

To understand the structure of *Finja* copies, we generated a consensus coverage approach from the output from RepeatMasker (http://www.repeatmasker.org/). This output allows to retrieve which part of the consensus sequence was used to mask a particular copy, and can be used to determine the consensus coverage. We generated a coverage of *Finja* consensus using all *Finja* copies, and then a second one using only copies located in gene 5'UTRs. To test if copies from gene 5'UTRs were enriched in a particular region of the consensus, we randomly selected 26 *Finja* copies (the same number of copies than in the one overlapping the 5'UTR gene regions) and generated the same coverage approach. After repeating 1000 times this random sampling, we were able to calculate the enrichment on copies overlapping 5'UTR of genes compared to copies randomly sampled. The scripts used are available on the gitlab (*Draw_coverage.sh*, see below).

Detection of interspecific polymorphic insertion of *Finja* in *Oryzias*

To identify polymorphic *Finja* insertions, we focused on full length copies in the *O. latipes* that we suspected to be recent insertions as they have a similar sizes and a high sequence similarity between each other. We retrieved 1kb of flanking genomic DNA on each side and blasted them on

the *O. javanicus* genome. If both sides of the TE insertion in *O. latipes* were found side by side in this genome (empty site without insertion), we considered it as a polymorphic insertion.

Gitlab repository

The scripts used to generate the figures of the article can be found at https://gitlab.com/Corend/Finja_-paper. All scripts from the repository that are not described here are called by *Draw_coverage.sh* or *Cross_all.sh*.

]

4 Conclusion et perspectives

Sommaire

4.1 Conclusion et perspectives générales

L'objectif global de ma thèse était de mieux comprendre l'impact des ET sur les réseaux de gènes à évolution rapide. Les poissons téléostéens sont un excellent modèle pour apporter de nouveaux éclairages dans ce domaine. Ils associent une grande diversité en termes de reproduction, détermination, développement et maintien du sexe, à une grande diversité en termes d'ET trouvés dans leurs génomes. Le médaka japonais *Oryzias latipes* a été choisi en particulier car il possède d'autres avantages. C'est une espèce modèle étudiée, facile à élever en laboratoire, avec un génome séquencé de bonne qualité. De plus, son système de détermination sexuelle est contrôlé par la présence d'un ET jouant le rôle de séquence régulatrice.

Pour ces raisons, des données transcriptomiques ont été générées à partir de gonades mâles et femelles d'*Oryzias latipes* adultes. Des données d'ARN poly-adénylés ont été utilisées pour séquencer des ARN messagers, issus aussi bien de l'expression de gènes que de celle d'ET.

J'ai mené de front les deux approches qui ont permis la rédaction des deux articles présentés précédemment. D'une part, une approche globale où l'expression de tous les gènes a été comparée à l'expression de tous les ET. Grâce à l'outil SQuIRE, j'ai pu étudier l'expression de chaque copie d'ET et la mettre en lien avec sa position dans le génome du médaka. Dans cette partie, j'ai mis en évidence une corrélation d'expression entre les gènes et les ET physiquement proches le long des chromosomes. Dans le but de comprendre s'il existe une organisation des gènes du sexe à l'échelle du génome, j'ai utilisé deux approches. La première, déjà décrite dans la littérature, consiste à rechercher des suites de gènes ou TE biaisés adjacents ; elle permet de tester l'hypothèse de la distribution non aléatoire des gènes selon leur expression. Cette approche m'a permis de montrer que les gènes sexe-biaisés ne sont pas distribués aléatoirement le long des chromosomes mais au contraire sont souvent adjacents. Par ailleurs, les ET sexe-biaisés forment des suites encore plus longues que les gènes. La nouvelle approche, que j'ai développée, m'a permis de détecter les loci particuliers présentant un biais moyen d'expression important lié au sexe ; ce biais peut-être mesuré à partir des niveaux d'expression des gènes ou des ET. J'ai eu l'opportunité de pouvoir appliquer cette méthode à d'autres jeux de données, notamment dans le cadre d'une collaboration avec William Toubiana (IGFL, équipe Khila) où nous avons pu mettre en évidence des clusters de gènes sexe-biaisés chez le gerris *Microvelia longipes* (TOUBIANA *et al.*, 2020). L'enrichissement des ET sexe-biaisés dans les clusters de gènes eux-mêmes sexe-biaisés indique que l'expression de certains gènes et ET n'est pas indépendante dans les gonades adultes d'*Oryzias latipes*. En outre, certains indices, comme la proximité phylogénétique des copies ayant des patrons d'expression similaires, soutiennent l'hypothèse que des ET régulent l'expression de certains gènes.

Pour tester cette hypothèse, l'objectif de la seconde partie a été de détecter des gènes dont l'expression pourrait être contrôlée en cis par des ET particulières. Dans ce but, l'approche présentée dans le second article consiste à rechercher des colocalisations entre des familles d'ET et des gènes sexe-biaisés. Les régions régulatrices des gènes pouvant être situées à des distances variables par rapport au site d'initiation de la transcription, il n'est pas évident de les identifier. C'est pourquoi dans un premier temps j'ai testé l'association entre des ET proches des gènes, ou insérés dans des régions facilement caractérisables comme les 5'UTR. Cette méthode a révélé plusieurs familles d'ET candidates dont une qui a retenu notre attention et que j'ai nommée *Finja*. En effet, ces

éléments contiennent des sites potentiels de fixation de facteurs de transcription; ces motifs sont particulièrement retenus dans les copies associées à un biais d'expression. De plus, la recherche de copies orthologues chez des espèces proches nous a permis d'associer la présence des copies au biais d'expression des gènes. Cette partie de la thèse montre qu'il semble exister des familles d'ET impliquées dans le contrôle de l'expression des gènes, à valider expérimentalement; qui restent difficile à détecter informatiquement.

Mes travaux s'inscrivent à la croisée de plusieurs domaines de la biologie de l'évolution. Tout d'abord ils apportent et testent de nouvelles hypothèses qui tentent d'expliquer la diversité observée chez les poissons téléostéens. Mais ils peuvent aussi bien se positionner dans une partie de la littérature scientifique qui s'intéresse à l'impact des ET sur l'évolution des génomes et des réseaux de régulation de gènes. Enfin, l'utilisation du sexe comme modèle d'étude permet aussi d'apporter de nouveaux éléments à la littérature traitant des réseaux de régulation liés au sexe chez les vertébrés. Les nouvelles méthodes et approches qui j'ai utilisées et développées pourront être réutilisées et améliorées par les différentes communautés. Les mêmes approches pourraient être appliquées pour tester l'impact des ET sur d'autre aspects de la biologie. Dans l'équipe, l'hypothèse de la duplication de génome spécifique des poissons téléostéens a été testée pour expliquer cette diversité de patron de pigmentation. Utiliser des données de séquençage de peau de poissons avec les mêmes approches permettrait en complément d'estimer le rôle des ET dans cette diversité de pigmentation. Il est aussi prévu d'appliquer cette approche pour comprendre le lien entre élément transposables et cancer en utilisant des données existantes de transcriptome de mélanome chez le médak, qui possède des mélanocytes au niveau de l'épiderme comme chez l'humain ce qui en fait un excellent modèle.

Mes travaux ont donc permis de montrer que l'expression des gènes n'est pas indépendante de l'expression des ET à l'échelle du génom, et que certaines familles d'ET sont associées à des gènes surexprimés dans les gonades mâles ou femelles. Cependant, plusieurs questions restent à élucider.

4.1.1 Identifier les familles d'éléments transposables particulièrement exprimées dans les gonades

En introduction, j'ai évoqué la balance qu'il existe entre expression et répression des ET dans les cellules germinales. Pour être transmise à la descendance, une nouvelle insertion d'ET doit avoir lieu dans une cellule germinale, qui sera potentiellement transmise à la nouvelle génération. Des insertions somatiques peuvent également avoir un impact sur la valeur sélective de l'hôte, mais ne seront pas transmises. C'est ce qui explique pourquoi les lignées germinales sont essentielles dans l'étude des ET. Les familles trop exprimées peuvent créer des mutations létales pour la descendance, et les familles peu exprimées ne formeront pas de nouvelles insertions. C'est cette fine balance qui a été discutée dans la revue « Sex and the TEs », publiée dans le journal Mobile DNA (Dechaud *et al.*, 2019). Nos données nous ont permis d'estimer l'expression des ET dans les gonades mâles et femelles et d'identifier des familles et des copies exprimées et différentiellement exprimées entre les sexes. Comme discuté dans les articles, puisqu'il n'est pas possible de séparer cellules germinales et cellules somatiques avant le séquençage, ces données contiennent les deux types de cellules. Il serait possible d'améliorer les données en séquençant des tissus purement somatiques comme du muscle, du cerveau ou du foie. La comparaison à nos données issues de gonades permettrait

d'identifier des familles d'ET surexprimées dans les gonades et donc de mieux comprendre la régulation de leur expression en lien avec la balance de régulation rappelée précédemment. Enfin, des données de séquençage en cellule unique permettraient de différencer les lignées somatiques et germinales pour précisément identifier des familles d'ET surexprimées dans les cellules germinales.

4.1.2 Comment avoir une vue d'ensemble du rôle des éléments transposables dans la fixation de facteurs de transcription ?

Les données de séquençage ChIP permettent d'identifier les régions du génome où se lie une protéine d'intérêt. Il est possible de générer de telles données pour des facteurs de transcription ou bien pour des marques d'histones. Des approches statistiques permettent ensuite de détecter des familles d'ET enrichies au niveau des sites de fixation de différents facteurs de transcription. La génération de ce type de données sur plusieurs espèces de poissons téléostéens pour les mêmes facteurs de transcription permettrait de découvrir des familles d'ET spécifiques à certaines espèces qui pourraient contribuer à la dispersion de sites de fixation à travers le génome. Cela permettrait d'analyser la manière dont des ET créent de la diversité, ou au contraire aboutissent à des convergences évolutives (dans le cas où des ET différents réguleraient de manière similaire des gènes orthologues dans différentes espèces). La fixation d'un facteur de transcription à une séquence n'étant pas systématiquement associée à une réelle fonction biologique, les données de séquençage ChIP ne dispensent pas de confirmer expérimentalement les résultats obtenus.

4.1.3 Confirmer l'organisation d'une partie des gènes en clusters

Dans le premier article traitant de l'organisation des gènes et des ET à l'échelle du génome, des régions particulièrement sexe-biaisées ont été mises en évidence. L'une des hypothèses avancées en introduction pour expliquer ces arrangements est que le regroupement de gènes ayant le même patron d'expression faciliterait la mise en place d'une co-régulation lors de l'introduction de séquences régulatrices ou via un contexte chromatinien particulier. Ces séquences réguleraient un ensemble de gènes, correspondant à un cluster. Utiliser des données de Hi-C générées dans les gonades de médaka permettrait de confirmer cette hypothèse. Le Hi-C consiste à séquencer des régions du génome proches dans l'espace, mais qui peuvent être éloignées sur les chromosomes. Cette méthode permet de détecter des interactions longue distance entre des parties du génome. Si notre hypothèse de clusters de gènes co-régulés se confirme, alors les régions que nous avons décrites pourraient correspondre à des régions qui interagissent observées avec les données Hi-C. Les ET retrouvés enrichis dans ces régions pourraient fixer des facteurs de transcription, ou recruter des marques d'histone ou d'autres marqueurs épigénétiques activateurs de la transcription. Ces familles seraient une fois de plus des candidates intéressantes qui devraient être testées expérimentalement pour confirmer leur rôle potentiel.

4.1.4 Le contrôle de l'expression des gènes n'est pas uniquement dû aux éléments transposables

Dans cette thèse j'ai beaucoup insisté sur le potentiel des ET à transporter des séquences régulatrices, mais il faut garder à l'esprit certaines limites de cette approche. L'expression des gènes

n'est pas déterminée que par la présence d'ET. Les marques d'histone, l'ouverture de la chromatine, les facteurs de transcription, les méthylations de l'ADN, les ARN non codants sont autant de facteurs qui influencent l'expression des gènes et qui sont souvent indépendants des ET. De plus ils peuvent rendre difficile l'identification des familles intéressantes en ajoutant des variations qui ne sont pas prises en compte dans nos approches.

4.1.5 Les autres mécanismes de contrôle de l'expression des gène par les élements transposables

4.1.5.1 ARNpi

Les ET peuvent contrôler l'expression des gènes par d'autres mécanismes que le transport de séquences régulatrices. Ils peuvent par exemple modifier l'expression des gènes proches de manière indirecte via leur propre contrôle par des ARNpi. Les ARNpi sont produits par l'hôte et sont complémentaires de l'ARN messager de certains ET. Leur fixation à l'ARN en cours de transcription recrute des protéines Piwi qui entraînent la méthylation des histones localement. L'expression de l'ET est donc inhibée, et avec elle celle des gènes localisés dans les environs. L'autre activité des ARNpi a lieu dans le cytoplasme. Les ET insérés dans l'intron d'un gène peuvent être complémentaires d'ARNpi qui activent le recrutement de protéines argonautes conduisant à la dégradation de l'ARNm. Dans ce manuscrit, la conséquence de la présence d'ARNpi n'a pas été étudiée. Étudier des données de séquençage d'ARNpi dans les gonades de médaka mâles et femelles permettrait d'identifier les familles les plus ciblées par ce type d'ARN. En croisant la localisation des copies ciblées avec la quantité d'ARNpi correspondants, il serait possible de vérifier que ces copies ont une expression diminuée. Ensuite, si ce constat est validé, analyser l'expression des gènes proches permettrait d'estimer l'impact des ARNpi sur l'expression des gènes liés au sexe et donc sur l'évolution des réseaux de régulation de ces gènes.

4.1.5.2 Épissage alternatif

Une autre forme de séquence régulatrice potentiellement encodée par les ET sont les sites d'épissages alternatifs. Cet aspect n'a pas été détaillé dans ce manuscrit. Les méthodes d'analyse déployées dans les articles précédents permettent de quantifier l'expression des gènes avec une précision au niveau du transcrit. Il serait donc possible dans le futur à partir de ces données de comparer les différents isoformes de gènes différentiellement exprimés entre conditions et associés à la présence d'éléments transposables.

Il existe ainsi de nombreuses façon d'estimer l'impact des ET sur les réseaux de régulation de gènes. La régulation en *cis* via la fixation de facteurs de transcription a été principalement étudiée, mais comme décrit ici, il peut exister des régulations en *trans* via des ARNpi, des méthylations d'histones, ou la création d'isoformes de transcrits. Les résultats présentés permettent d'avoir une idée de l'impact des ET mais ne traitent pas de tous les aspects sur lesquels ils peuvent potentiellement influer.

4.1.5.3 L'intérêt des données de lectures longues

Il serait difficile de ne pas mentionner les technologies de séquençage d'ARN en lectures longues pour estimer l'expression des ET. Ces méthodes permettraient de séquencer des transcrits entiers, qu'ils soient issus de gènes ou d'ET. Elles faciliteraient la détection des isoformes, mais aussi éviteraient les soucis d'alignements multiples rencontrés avec les ET. Les transcrits chimériques de gènes contenant des ET seraient facilement identifiés, ce qui faciliterait toutes les analyses faites précédemment et ouvrirait de nouvelles perspectives dans l'analyse d'expression des ET.

```
corentin@igfl:~$ sudo shutdown
```

A Annexes

Sommaire

A.1 Collaborations

A.1.1 Architecture génomique chez le gerris *Microvelia longipes*

Au cours de ma thèse, j'ai eu la chance de pouvoir collaborer avec William Toubiana alors en thèse dans l'équipe d'Abderaman Khila à l'IGFL. Ils s'intéressent aux bases génétiques de l'évolution, en se basant sur une espèce de gerris, *Microvelia longipes*, dont la taille des pattes varie selon le sexe. Ils cherchaient à tester si les gènes sexe-biaisés impliqués dans ce dimorphisme étaient localisés sur des chromosomes particuliers ou regroupés sous forme de clusters. Suite à nos discussions ils m'ont aidé à améliorer la méthode de détection des clusters de gènes le long des chromosomes présentée dans la première partie. De plus nous l'avons appliquée à leurs données, les résultats étant présentés dans un article soumis pour publication, disponible en ligne sur bioRχiv : https://www.biorxiv.org/content/10.1101/2020.01.10.901322v1

A.1.2 Le génome géant du dipneuste éclaire la conquête du milieu terrestre par les vertébrés

Pendant ma thèse, l'équipe a été contactée par un consortium scientifique international ayant pour but de séquencer le génome du dipneuste australien, *Neoceradotus forsteri*. Le dipneuste (lungfish en anglais) est l'organisme aquatique actuel le plus proche des tétrapodes (amphibiens, oiseaux, reptiles, mammifères), majoritairement terrestres. Il possède un poumon fonctionnel en plus de branchies et la capacité de sentir des odeurs dans l'air. Son génome de 43Gb est le plus grand génome animal jamais séquencé, et dépasse le record de 32Gb détenu par celui de l'axolotl. L'annotation des éléments transposables a montré que 90% du génome du dipneuste est constitué d'éléments répétés. J'ai été impliqué dans l'annotation, la classification ainsi que l'analyse de l'expression des éléments transposables chez cet organisme. Cela a aboutit à la publication d'un article dans *Nature* : https://www.nature.com/articles/s41586-021-03198-8.

A.1.3 L'évolution des génomes d'espèces proches de *Xiphophorus*

Dans le cadre d'une collaboration avec Manfred Schartl à l'université de Würzburg, les génomes de trois espèces de platy ont été séquencés : *Xiphophorus maculatus*, le platy commun, *Xiphophorus hellerii*, le porte-épée et *Xiphophorus couchianus*, le platy de Monterrey. Dans le but de comprendre l'évolution récente et les processus de spéciation ayant eu lieu dans ces espèces, j'ai été chargé d'annoter les éléments transposables des trois génomes. J'ai utilisé REPET (https://urgi.versailles.inra.fr/Tools/REPET), un pipeline permettant de générer une banque d'élément transposables *de novo* à partir d'un génome, ainsi que de l'annoter. Ces analyses ont permis d'identifier des familles d'éléments transposables spécifiques d'une espèces ainsi que des insertions polymorphes entre espèces. Un article présentant ces résultats est en préparation.

A.1.4 Annotation des éléménts transposables du génome de *Lucifuga dentata*

J'ai été impliqué dans l'annotation des éléments transposables du génome d'un poisson cavernicole, *Lucifuga dentata*, dans un projet traitant de la perte de gènes liés à la vision chez cette espèce. Cette collaboration a abouti à la publication d'un article dans *Molecular biology and evolution* : https://academic.oup.com/mbe/advance-article/doi/10.1093/molbev/msaa249/5912537

A.2 Conférences

A.2.1 Communications orales

Invité pour un séminaire de bioinformatique et biologie computationnelle, centre de bioinformatique de Bordeaux

- Mars 2021 - Conférence virtuelle
- https://www.cbib.u-bordeaux.fr/en/cbib
- Assessing the role of transposable elements in the control of sexual genes in teleost fish

Symposium EMBO|EMBL "The Molecular basis and evolution of sexual dimorphism"

- Septembre 2020 - Conférence virtuelle
- https://www.embo-embl-symposia.org/symposia/2020/EES20-09/programme/index.html
- Sex and the TEs : teleost fish gonad transcriptome analysis reveals clusters of sex-biased genes and transposable elements

Rencontres Alphy

- Février 2020 - Lyon
- https://lbbe.univ-lyon1.fr/alphy/spip.php?article77
- Sex and the TEs : role of transposable elements in the control of sexual genes in teleost fish

Réunion du groupement de recherche MobilET

- Décembre 2019 - Paris
- https://www.mobil-et.eu/fr/
- Role of transposable elements in the control of sexual genes in teleost fish

Réseau interdisciplinaire autour de la statistique : Journée statistique et génomique

- Septembre 2019 - Paris
- Détection de clusters de gènes différentiellement exprimés par une méthode de bootstraps

Congrès national sur les éléments transposables (CNET)

- Juillet 2019 - Lyon
- https://cnet2019.sciencesconf.org/
- Role of transposable elements in the control of sexual genes in teleost fish

Congrès national sur les éléments transposables (CNET)

- Juillet 2018 - Clermont-Ferrand

— https://cnet2018.sciencesconf.org/

— Role of transposable elements in the control of sexual genes in teleost fish

Réunion transversale : différenciation et fonction des gonades

— Mars 2018 - Rennes

— Role of transposable elements in the control of sexual genes in teleost fish

A.2.2 Posters

— GdR MobilET 2017 - Paris - Décembre 2017

— Crossroad between transposons and gene regulation - Londres - Mai 2019

Liste complète des références

ADOLFI, M. C., P. FISCHER, A. HERPIN, M. REGENSBURGER, M. KIKUCHI, M. TANAKA et M. SCHARTL. 2019, «Increase of cortisol levels after temperature stress activates dmrt1a causing female-to-male sex reversal and reduced germ cell number in medaka», *Molecular Reproduction and Development*, vol. 86, p. 1–13. 11, 59

AINSWORTH, C. 2015, «Sex redefined», *Nature News*, vol. 518, nº 7539, doi :10.1038/518288a, p. 288. URL http://www.nature.com/news/sex-redefined-1.16943. 4

ARBOLEDA, V. A., D. E. SANDBERG et E. VILAIN. 2014, «DSDs : genetics, underlying pathologies and psychosexual differentiation», *Nature reviews. Endocrinology*, vol. 10, nº 10, doi :10.1038/nrendo.2014.130, p. 603–615, ISSN 1759-5029. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4441533/. 4

ARKHIPOVA, I. R. 2005, «Mobile genetic elements and sexual reproduction», *Cytogenetic and Genome Research*, vol. 110, p. 372–382. 4

ARNOLD, A. P. et Y. ITOH. 2011, «Factors causing sex differences in birds», *Avian biology research*, vol. 4, nº 2, doi :10.3184/175815511X13070045977959, ISSN 1758-1559. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3864897/. 8

ASSIS, R., Q. ZHOU et D. BACHTROG. 2012, «Sex-biased transcriptome evolution in drosophila», *Genome Biology and Evolution*, vol. 4, p. 1189–1200. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3514954/. 15, 59, 114

BACHTROG, D., J. E. MANK, C. L. PEICHEL, M. KIRKPATRICK, S. P. OTTO, T.-L. ASHMAN, M. W. HAHN, J. KITANO, I. MAYROSE, R. MING, N. PERRIN, L. ROSS, N. VALENZUELA, J. C. VAMOSI et THE TREE OF SEX CONSORTIUM. 2014, «Sex determination : why so many ways of doing it?», *PLoS Biology*, vol. 12, p. e1001 899. URL http://dx.plos.org/10.1371/journal.pbio.1001899. 8, 59, 103

BAILEY, J. A. et E. E. EICHLER. 2006, «Primate segmental duplications : crucibles of evolution, diversity and disease», *Nature Reviews. Genetics*, vol. 7, nº 7, doi :10.1038/nrg1895, p. 552–564, ISSN 1471-0056. 22

BAILEY, J. A., G. LIU et E. E. EICHLER. 2003, «An Alu transposition model for the origin and expansion of human segmental duplications», *American Journal of Human Genetics*, vol. 73, nº 4, doi :10.1086/378594, p. 823–834, ISSN 0002-9297. 22

BANKEVICH, A., S. NURK, D. ANTIPOV, A. A. GUREVICH, M. DVORKIN, A. S. KULIKOV, V. M. LESIN, S. I. NIKOLENKO, S. PHAM, A. D. PRJIBELSKI, A. V. PYSHKIN, A. V. SIROTKIN, N. VYAHHI, G. TESLER, M. A. ALEKSEYEV et P. A. PEVZNER. 2012, «SPAdes : A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing», *Journal of Computational Biology*, vol. 19, nº 5, doi :10.1089/cmb.2012.0021, p. 455–477. URL https://www.liebertpub.com/doi/10.1089/cmb.2012.0021, publisher : Mary Ann Liebert, Inc., publishers. 48

BAO, Z. et S. R. EDDY. 2002, «Automated De Novo Identification of Repeat Sequence Families in Sequenced Genomes», *Genome Research*, vol. 12, p. 1269–1276. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC186642/. 50

BAR, I., S. CUMMINS et A. ELIZUR. 2016, «Transcriptome analysis reveals differentially expressed genes associated with germ cell and gonad development in the Southern bluefin tuna (Thunnus maccoyii)», *BMC Genomics*, vol. 17. 14, 70, 114

BARAU, J., A. TEISSANDIER, N. ZAMUDIO, S. ROY, V. NALESSO, Y. HÉRAULT, F. GUILLOU et D. BOURC'HIS. 2016, «The DNA methyltransferase DNMT3C protects male germ cells from transposon activity», *Science (New York, N.Y.)*, vol. 354, nº 6314, doi :10.1126/science.aah5143, p. 909–912, ISSN 1095-9203. 23

BAROILLER, J., H. D&RSQUO ;COTTA et E. SAILLANT. 2009, «Environmental Effects on Fish Sex Determination and Differentiation», *Sexual Development*, vol. 3, p. 118–135. URL https://www.karger.com/Article/FullText/223077. 8, 9, 12

BEAL, A. P., F. D. MARTIN et M. C. HALE. 2017, «Using RNA-seq to determine patterns of sex-bias in gene expression in the brain of the sex-role reversed Gulf Pipefish (Syngnathus scovelli)», *Marine Genomics*, vol. 37, p. 120–127. 14, 70, 114

BEJERANO, G., C. B. LOWE, N. AHITUV, B. KING, A. SIEPEL, S. R. SALAMA, E. M. RUBIN, W. J. KENT et D. HAUSSLER. 2006, «A distal enhancer and an ultraconserved exon are derived from a novel retroposon», *Nature*, vol. 441, nº 7089, doi :10.1038/nature04696, p. 87–90, ISSN 1476-4687. 23

BELTON, J.-M., R. P. MCCORD, J. H. GIBCUS, N. NAUMOVA, Y. ZHAN et J. DEKKER. 2012, «Hi-C : a comprehensive technique to capture the conformation of genomes», *Methods (San Diego, Calif.)*, vol. 58, nº 3, doi :10.1016/j.ymeth.2012.05.001, p. 268–276, ISSN 1095-9130. 48

BIÉMONT, C. et C. VIEIRA. 2006, «Junk DNA as an evolutionary force», *Nature*, vol. 443, p. 521–524. URL http://www.nature.com/nature/journal/v443/n7111/full/443521a.html. 22

BOBINNEC, Y., C. MARCAILLOU et A. DEBEC. 1999, «Microtubule polyglutamylation in Drosophila melanogaster brain

and testis», vol. 78, nº 9, p. 671–674, ISSN 01719335. URL https://linkinghub.elsevier.com/retrieve/pii/S0171933599800533. 65

BOGAERTS-MÁRQUEZ, M., M. G. BARRÓN, A.-S. FISTON-LAVIER, P. VENDRELL-MIR, R. CASTANERA, J. M. CASACUBERTA et J. GONZÁLEZ. 2019, «T-lex3 : an accurate tool to genotype and estimate population frequencies of transposable elements using the latest short-read whole genome sequencing data», *Bioinformatics (Oxford, England)*. 51

BOURQUE, G. 2009, «Transposable elements in gene regulation and in the evolution of vertebrate genomes», *Current Opinion in Genetics & Development*, vol. 19, nº 6, doi : 10.1016/j.gde.2009.10.013, p. 607–612, ISSN 0959-437X. URL http://www.sciencedirect.com/science/article/pii/S0959437X09001725. 22, 23

BOURQUE, G., K. H. BURNS, M. GEHRING, V. GORBUNOVA, A. SELUANOV, M. HAMMELL, M. IMBEAULT, Z. IZSVÁK, H. L. LEVIN, T. S. MACFARLAN, D. L. MAGER et C. FESCHOTTE. 2018, «Ten things you should know about transposable elements», *Genome Biology*, vol. 19. 60

BOURQUE, G., B. LEONG, V. B. VEGA, X. CHEN, Y. L. LEE, K. G. SRINIVASAN, J.-L. CHEW, Y. RUAN, C.-L. WEI, H. H. NG et E. T. LIU. 2008, «Evolution of the mammalian transcription factor binding repertoire via transposable elements», *Genome Research*, vol. 18, nº 11, doi :10.1101/gr.080663.108, p. 1752–1762, ISSN 1088-9051. URL http://genome.cshlp.org/cgi/doi/10.1101/gr.080663.108. 23

BOUTANAEV, A. M., A. I. KALMYKOVA, Y. Y. SHEVELYOV et D. I. NURMINSKY. 2002, «Large clusters of co-expressed genes in the Drosophila genome», *Nature*, vol. 420, p. 666. 59, 63, 66, 72, 77

BRANDT, J., A. M. VEITH et J.-N. VOLFF. 2005, «A family of neofunctionalized Ty3/gypsy retrotransposon genes in mammalian genomes», *Cytogenetic and Genome Research*, vol. 110, nº 1-4, doi :10.1159/000084963, p. 307–317, ISSN 1424-859X. 26

BRITTEN, R. J. et E. H. DAVIDSON. 1969, «Gene regulation for higher cells : a theory», *Science (New York, N.Y.)*, vol. 165, nº 3891, doi :10.1126/science.165.3891.349, p. 349–357, ISSN 0036-8075. 26

BULL, J. J. et R. C. VOGT. 1979, «Temperature-dependent sex determination in turtles», *Science*, vol. 206, p. 1186–1188. URL https://science.sciencemag.org/content/206/4423/1186. 8

BUNDO, M., M. TOYOSHIMA, Y. OKADA, W. AKAMATSU, J. UEDA, T. NEMOTO-MIYAUCHI, F. SUNAGA, M. TORITSUKA, D. IKAWA, A. KAKITA, M. KATO, K. KASAI, T. KISHIMOTO, H. NAWA, H. OKANO, T. YOSHIKAWA, T. KATO et K. IWAMOTO. 2014, «Increased l1 retrotransposition in the neuronal genome in schizophrenia», *Neuron*, vol. 81, nº 2, doi :10.1016/j.neuron.2013.10.053, p. 306–313, ISSN 1097-4199. 19

BÖHNE, A., C. SCHULTHEIS, D. GALIANA-ARNOUX, A. FROSCHAUER, Q. ZHOU, C. SCHMIDT, Y. SELZ, C. OZOUF-COSTAZ, A. DETTAI, B. SEGURENS, A. COULOUX, S. BERNARD-SAMAIN, V. BARBE, S. CHILMONCZYK, F. BRUNET, A. DARRAS, M. TOMASZKIEWICZ, M. SEMON, M. SCHARTL et J.-N. VOLFF. 2009, «Molecular analysis of the sex chromosomes of the platyfish Xiphophorus maculatus : Towards the identification of a new type of master sexual regulator in vertebrates», *Integrative Zoology*, vol. 4, nº 3, doi :https://doi.org/10.1111/j.1749-4877.2009.00166.x, p. 277–284, ISSN 1749-4877. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1749-4877.2009.00166.x. 9

BÖHNE, A., T. SENGSTAG et W. SALZBURGER. 2014, «Comparative transcriptomics in east african cichlids reveals sex- and species-specific expression and new candidates for sex differentiation in fishes», *Genome Biology and Evolution*, vol. 6, p. 2567–2585. 14, 70, 71, 114

C. ELEGANS SEQUENCING CONSORTIUM. 1998, «Genome sequence of the nematode C. elegans : a platform for investigating biology», *Science (New York, N.Y.)*, vol. 282, nº 5396, doi : 10.1126/science.282.5396.2012, p. 2012–2018, ISSN 0036-8075. 47

CAMATS, N., A. V. PANDEY, M. FERNÁNDEZ-CANCIO, P. ANDALUZ, M. JANNER, N. TORÁN, F. MORENO, A. BEREKET, T. AKCAY, E. GARCÍA-GARCÍA, M. T. MUÑOZ, R. GRACIA, M. NISTAL, L. CASTAÑO, P. E. MULLIS, A. CARRASCOSA, L. AUDÍ et C. E. FLÜCK. 2012, «Ten novel mutations in the NR5A1 gene cause disordered sex development in 46,XY and ovarian insufficiency in 46,XX individuals», *The Journal of Clinical Endocrinology and Metabolism*, vol. 97, p. E1294–1306. 3

CAPEL, B. 2017, «Vertebrate sex determination : evolutionary plasticity of a fundamental switch», *Nature Reviews Genetics*, vol. 18, p. 675–689. URL http://www.nature.com/doifinder/10.1038/nrg.2017.60. 8

CARDUCCI, F., M. BARUCCA, A. CANAPA, E. CAROTTI et M. A. BISCOTTI. 2020, «Mobile Elements in Ray-Finned Fish Genomes», *Life*, vol. 10, nº 10, doi :10.3390/life10100221. URL https://www.mdpi.com/2075-1729/10/10/221. 27

CARMONA, L. M. et D. G. SCHATZ. 2017, «New insights into the evolutionary origins of the RAG proteins and V(D)J recombination», *The FEBS journal*, vol. 284, p. 1590–1605. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5459667/. 26

CASAS, L., F. SABORIDO-REY, T. RYU, C. MICHELL, T. RAVASI et X. IRIGOIEN. 2016, «Sex Change in Clownfish : Molecular Insights from Transcriptome Analysis», *Scientific Reports*, vol. 6, p. 35 461. 6

CHAISSON, M. J. P., R. K. WILSON et E. E. EICHLER. 2015, «Genetic variation and the de novo assembly of human genomes», *Nature Reviews Genetics*, vol. 16, p. 627–640. URL http://www.nature.com/articles/nrg3933. 49

CHALOPIN, D., D. GALIANA et J.-N. VOLFF. 2012, «Genetic Innovation in Vertebrates : Gypsy Integrase Genes and Other Genes Derived from Transposable Elements», doi :https://doi.org/10.1155/2012/724519. URL https://www.hindawi.com/journals/ijeb/2012/724519/, iSSN : 2090-8032 Pages : e724519 Publisher : Hindawi Volume : 2012. 26

CHALOPIN, D., M. NAVILLE, F. PLARD, D. GALIANA et J.-N. VOLFF. 2015, «Comparative analysis of transposable elements highlights mobilome diversity and evolution in vertebrates», *Genome Biology and Evolution*, vol. 7, p. 567–580. 18, 27, 28, 29, 59, 61, 103

CHARLESWORTH, D., B. CHARLESWORTH et G. MARAIS. 2005, «Steps in the evolution of heteromorphic sex chromosomes», *Heredity*, vol. 95, p. 118–128. 73

CHENG, J. 2005, «Transcriptional Maps of 10 Human Chromosomes at 5-Nucleotide Resolution», *Science*, vol. 308, p. 1149–1154. 71

CHEUNG, S., S. MANHAS et V. MEASDAY. 2018, «Retrotransposon targeting to RNA polymerase III-transcribed genes», *Mobile DNA*, vol. 9. 73

CHINWALLA, A. T., L. L. COOK, K. D. DELEHAUNTY, G. A. FEWELL, L. A. FULTON, R. S. FULTON, T. A. GRAVES, L. W. HILLIER, E. R. MARDIS, J. D. MCPHERSON, T. L. MINER, W. E. NASH, J. O. NELSON, M. N. NHAN, K. H. PEPIN, C. S. POHL, T. C. PONCE, B. SCHULTZ, J. THOMPSON, E. TREVASKIS, R. H. WATERSTON, M. C. WENDL, R. K. WILSON, S.-P. YANG, P. AN, E. BERRY, B. BIRREN, T. BLOOM, D. G. BROWN, J. BUTLER, M. DALY, R. DAVID, J. DERI, S. DODGE, K. FOLEY, D. GAGE, S. GNERRE, T. HOLZER, D. B. JAFFE, M. KAMAL, E. K. KARLSSON, C. KELLS, A. KIRBY, E. J. KULBOKAS, E. S. LANDER, T. LANDERS, J. P. LEGER, R. LEVINE, K. LINDBLAD-TOH, E. MAUCELI, J. H. MAYER, M. MCCARTHY, J. MELDRIM, J. MELDRIM, J. P. MESIROV, R. NICOL, C. NUSBAUM, S. SEAMAN, T. SHARPE, A. SHERIDAN, J. B. SINGER, R. SANTOS, B. SPENCER, N. STANGE-THOMANN, J. P. VINSON, C. M. WADE, J. WIERZBOWSKI, D. WYMAN, M. C. ZODY, E. BIRNEY, N. GOLDMAN, A. KASPRZYK, E. MONGIN, A. G. RUST, G. SLATER, A. STABENAU, A. URETA-VIDAL, S. WHELAN, R. AINSCOUGH, J. ATTWOOD, J. BAILEY, K. BARLOW, S. BECK, J. BURTON, M. CLAMP, C. CLEE, A. COULSON, J. CUFF, V. CURWEN, T. CUTTS, J. DAVIES, E. EYRAS, D. GRAFHAM, S. GREGORY, T. HUBBARD, A. HUNT, M. JONES, A. JOY, S. LEONARD, C. LLOYD, L. MATTHEWS, S. MCLAREN, K. MCLAY, B. MEREDITH, J. C. MULLIKIN, Z. NING, K. OLIVER, E. OVERTON-LARTY, R. PLUMB, S. POTTER, M. QUAIL, J. ROGERS, C. SCOTT, S. SEARLE, R. SHOWNKEEN, S. SIMS, M. WALL, A. P. WEST, D. WILLEY, S. WILLIAMS, J. F. ABRIL, R. GUIGÓ, G. PARRA, P. AGARWAL, R. AGARWALA, D. M. CHURCH, W. HLAVINA, D. R. MAGLOTT, V. SAPOJNIKOV, M. ALEXANDERSSON, L. PACHTER, S. E. ANTONARAKIS, E. T. DERMITZAKIS, A. REYMOND, C. UCLA, R. BAERTSCH, M. DIEKHANS, T. S. FUREY, A. HINRICHS, F. HSU, D. KAROLCHIK, W. J. KENT, K. M. ROSKIN, M. S. SCHWARTZ, C. SUGNET, R. J. WEBER, P. BORK, I. LETUNIC, M. SUYAMA, D. TORRENTS, E. M. ZDOBNOV, M. BOTCHERBY, S. D. BROWN, R. D. CAMPBELL, I. JACKSON, N. BRAY, O. COURONNE, I. DUBCHAK, A. POLIAKOV, E. M. RUBIN, M. R. BRENT, P. FLICEK, E. KEIBLER, I. KORF, S. BATALOV, C. BULT, W. N. FRANKEL, P. CARNINCI, Y. HAYASHIZAKI, J. KAWAI, Y. OKAZAKI, S. CAWLEY, D. KULP, R. WHEELER, F. CHIAROMONTE, F. S. COLLINS, A. FELSENFELD, M. GUYER, J. PETERSON, K. WETTERSTRAND, R. R. COPLEY, R. MOTT, C. DEWEY, N. J. DICKENS, R. D. EMES, L. GOODSTADT, C. P. PONTING, E. WINTER, D. M. DUNN, A. C. VON NIEDERHAUSERN, R. B. WEISS, S. R. EDDY, L. S. JOHNSON, T. A. JONES, L. ELNITSKI, D. L. KOLBE, P. ESWARA, W. MILLER, M. J. O'CONNOR, S. SCHWARTZ, R. A. GIBBS, D. M. MUZNY, G. GLUSMAN, A. SMIT, E. D. GREEN, R. C. HARDISON, S. YANG, D. HAUSSLER, A. HUA, B. A. ROE, R. S. KUCHERLAPATI, K. T. MONTGOMERY, J. LI, M. LI, S. LUCAS, B. MA, W. R. MCCOMBIE, M. MORGAN, P. PEVZNER, G. TESLER, J. SCHULTZ, D. R. SMITH, J. TROMP, K. C. WORLEY, E. S. LANDER, J. F. ABRIL, P. AGARWAL, M. ALEXANDERSSON, S. E. ANTONARAKIS, R. BAERTSCH, E. BERRY, E. BIRNEY, P. BORK, N. BRAY, M. R. BRENT, D. G. BROWN, J. BUTLER, C. BULT, F. CHIAROMONTE, A. T. CHINWALLA, D. M. CHURCH, M. CLAMP, F. S. COLLINS, R. R. COPLEY, O. COURONNE, S. CAWLEY, J. CUFF, V. CURWEN, T. CUTTS, M. DALY, E. T. DERMITZAKIS, C. DEWEY, MOUSE GENOME SEQUENCING CONSORTIUM, GENOME SEQUENCING CENTER :, WHITEHEAD INSTITUTE/MIT CENTER FOR GENOME RESEARCH :, EUROPEAN BIOINFORMATICS INSTITUTE :, WELLCOME TRUST SANGER INSTITUTE, RESEARCH GROUP IN BIOMEDICAL INFORMATICS, BIOINFORMATICS, NATIONAL CENTER FOR BIOTECHNOLOGY INFORMATION, DEPARTMENT OF MATHEMATICS, DIVISION OF MEDICAL GENETICS, CENTER FOR BIOMOLECULAR SCIENCE AND ENGINEERING, EMBL, UK MRC MOUSE SEQUENCING CONSORTIUM, LAWRENCE BERKELEY NATIONAL LABORATORY, DEPARTMENT OF COMPUTER SCIENCE, SCHOOL OF COMPUTER SCIENCE, THE JACKSON LABORATORY, LABORATORY FOR GENOME EXPLORATION, AFFYMETRIX INC., DEPARTMENTS OF STATISTICS AND HEALTH EVALUATION SCIENCES, NATIONAL HUMAN GENOME RESEARCH INSTITUTE, WELLCOME TRUST CENTRE FOR HUMAN GENETICS, DEPARTMENT OF ELECTRICAL ENGINEERING, DEPARTMENT OF HUMAN ANATOMY AND GENETICS, DEPARTMENT OF HUMAN GENETICS, HOWARD HUGHES MEDICAL INSTITUTE AND DEPARTMENT OF GENETICS, DEPARTMENTS OF BIOCHEMISTRY AND MOLECULAR BIOLOGY AND COMPUTER SCIENCE AND ENGINEERING, DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING, BAYLOR COLLEGE OF MEDICINE, THE INSTITUTE FOR SYSTEMS BIOLOGY, DEPARTMENT OF BIOCHEMISTRY AND MOLECULAR BIOLOGY, HOWARD HUGHES MEDICAL INSTITUTE, DEPARTMENT OF CHEMISTRY AND BIOCHEMISTRY, DEPARTMENTS OF GENETICS AND MEDICINE AND HARVARD-PARTNERS CENTER FOR GENETICS AND GENOMICS, DEPARTMENT OF STATISTICS, US DOE JOINT GENOME INSTITUTE, COLD SPRING HARBOR LABORATORY, WELLCOME TRUST, MAX PLANCK INSTITUTE FOR MOLECULAR GENETICS, GENOME THERAPEUTICS CORPORATION, BIOINFORMATICS SOLUTIONS INC., DEPARTMENT OF MOLECULAR AND HUMAN GENETICS, DEPARTMENT OF BIOLOGY et MEMBERS OF THE MOUSE GENOME ANALYSIS GROUP. 2002, «Initial sequencing and comparative analysis of the mouse genome», *Nature*, vol. 420, nº 6915, doi :10.1038/nature01262, p. 520–562, ISSN 1476-4687. URL https://www.nature.com/articles/nature01262, number : 6915 Publisher : Nature Publishing Group. 47

CHUONG, E. B. 2013, «Retroviruses facilitate the rapid evolution of the mammalian placenta», *BioEssays : News and Reviews in Molecular, Cellular and Developmental Biology*, vol. 35, p. 853–861. 74

CHUONG, E. B., N. C. ELDE et C. FESCHOTTE. 2016, «Regulatory activities of transposable elements : from conflicts to benefits», *Nature Reviews Genetics*, vol. 18, p. 71–86. 23, 60

CHÉNAIS, B., A. CARUSO, S. HIARD et N. CASSE. 2012, «The impact of transposable elements on eukaryotic genomes : From genome size increase to genetic adaptation to stressful environments», *Gene*, vol. 509, p. 7–15. URL http://www.sciencedirect.com/science/article/pii/S0378111912008931. 18

COBB, M. 2017, «60 years ago, Francis Crick changed the

logic of biology», *PLOS Biology*, vol. 15, p. e2003 243. URL https://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.2003243. 17

COFFIN, J. M. 1992, «Structure and Classification of Retroviruses», dans *The Retroviridae*, édité par J. A. Levy, The Viruses, Springer US, Boston, MA, ISBN 978-1-4615-3372-6, p. 19–49, doi :10.1007/978-1-4615-3372-6_2. URL https://doi.org/10.1007/978-1-4615-3372-6_2. 20

CORNELIS, G., M. FUNK, C. VERNOCHET, F. LEAL, O. A. TARAZONA, G. MEURICE, O. HEIDMANN, A. DUPRESSOIR, A. MIRALLES, M. P. RAMIREZ-PINILLA et T. HEIDMANN. 2017, «An endogenous retroviral envelope syncytin and its cognate receptor identified in the viviparous placental Mabuya lizard», *Proceedings of the National Academy of Sciences of the United States of America*, vol. 114, p. E10 991–E11 000. 26

CRICK, F. H. 1958, *On protein synthesis*, vol. 12, 8 p.. 17

CUTTER, A. D. et S. WARD. 2005, «Sexual and temporal dynamics of molecular evolution in C. elegans development», *Molecular Biology and Evolution*, vol. 22, p. 178–188. 15, 59, 114

DAVIDSON, E. H. et R. J. BRITTEN. 1979, «Regulation of gene expression : possible role of repetitive sequences», *Science (New York, N.Y.)*, vol. 204, nº 4397, doi :10.1126/science.451548, p. 1052–1059, ISSN 0036-8075. 26

DECHAUD, C., J.-N. VOLFF, M. SCHARTL et M. NAVILLE. 2019, «Sex and the TEs : transposable elements in sexual development and function in animals», *Mobile DNA*, vol. 10. 31, 60, 71, 73, 74, 104, 115, 123

DEININGER, P. 2011, «Alu elements : know the SINEs», *Genome Biology*, vol. 12, p. 236. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3334610/. 17, 21

DEININGER, P. L. et M. A. BATZER. 2002, «Mammalian retroelements», *Genome Research*, vol. 12, p. 1455–1465. 52, 71

DEWANNIEUX, M., C. ESNAULT et T. HEIDMANN. 2003, «LINE-mediated retrotransposition of marked Alu sequences», *Nature Genetics*, vol. 35, nº 1, doi :10.1038/ng1223, p. 41–48, ISSN 1061-4036. 20

DORUS, S., S. A. BUSBY, U. GERIKE, J. SHABANOWITZ, D. F. HUNT et T. L. KARR. 2006, «Genomic and functional evolution of the Drosophila melanogaster sperm proteome», *Nature Genetics*, vol. 38, p. 1440–1445. 59, 72

DU, J., A. LEUNG, C. TRAC, M. LEE, B. W. PARKS, A. J. LUSIS, R. NATARAJAN et D. E. SCHONES. 2016, «Chromatin variation associated with liver metabolism is mediated by transposable elements», *Epigenetics & Chromatin*, vol. 9, doi :10.1186/s13072-016-0078-0, p. 28, ISSN 1756-8935. 23

DUPRESSOIR, A., G. MARCEAU, C. VERNOCHET, L. BÉNIT, C. KANELLOPOULOS, V. SAPIN et T. HEIDMANN. 2005, «Syncytin-A and syncytin-B, two fusogenic placenta-specific murine envelope genes of retroviral origin conserved in Muridae», *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, p. 725–730. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC545540/. 26

DUPRESSOIR, A., C. VERNOCHET, F. HARPER, G. PIERRON, J. GUÉGAN, P. DESSEN et T. HEIDMANN. 2011, «Contribution of captured retroviral envelope genes, the "syncytins" to the formation of the mouse placenta», *Retrovirology*, vol. 8, p. O24. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3236869/. 26

EDGAR, R. C. et E. W. MYERS. 2005, «PILER : identification and classification of genomic repeats», *Bioinformatics (Oxford, England)*, vol. 21 Suppl 1, p. i152–158. 50

EGGERS, S., T. OHNESORG et A. SINCLAIR. 2014, «Genetic regulation of mammalian gonad development», *Nature Reviews Endocrinology*, vol. 10, p. 673–683. 12, 13

EICKBUSH, T. H. et H. S. MALIK. 2002, *Mobile DNA II*, ASM Press, Washington, D.C, ISBN 978-1-55581-209-6. 21

ELISAPHENKO, E. A., N. N. KOLESNIKOV, A. I. SHEVCHENKO, I. B. ROGOZIN, T. B. NESTEROVA, N. BROCKDORFF et S. M. ZAKIAN. 2008, «A dual origin of the Xist gene from a protein-coding gene and a set of transposable elements», *PloS One*, vol. 3, nº 6, doi :10.1371/journal.pone.0002521, p. e2521, ISSN 1932-6203. 27

ELLEGREN, H. et J. PARSCH. 2007, «The evolution of sex-biased genes and sex-biased gene expression», *Nature Reviews Genetics*, vol. 8, p. 689–698. URL http://www.nature.com/articles/nrg2167. 14, 15

ELLINGHAUS, D., S. KURTZ et U. WILLHOEFT. 2008, «LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons», *BMC Bioinformatics*, vol. 9. 75

ELLISON, C. E. et D. BACHTROG. 2013, «Dosage compensation via transposable element mediated rewiring of a regulatory network», *Science*, vol. 342, p. 846–850. 23, 60, 66, 69

ELLISON, C. E. et D. BACHTROG. 2015, «Non-allelic gene conversion enables rapid evolutionary change at multiple regulatory sites encoded by transposable elements», *Elife*, vol. 4. 60, 69

EMANUEL, B. S. et T. H. SHAIKH. 2001, «Segmental duplications : an 'expanding' role in genomic instability and disease», *Nature Reviews. Genetics*, vol. 2, nº 10, doi :10.1038/35093500, p. 791–800, ISSN 1471-0056. 22

ENGELSTÄDTER, J. 2017, «Asexual but Not Clonal : Evolutionary Processes in Automictic Populations», *Genetics*, vol. 206, p. 993–1009. URL http://www.genetics.org/lookup/doi/10.1534/genetics.116.196873. 5

ETCHEGARAY, E., M. NAVILLE, J.-N. VOLFF et Z. HAFTEK-TERREAU. 2021, «Transposable element-derived sequences in vertebrate development», *Mobile DNA*, vol. 12, nº 1, doi :10.1186/s13100-020-00229-5, p. 1, ISSN 1759-8753. 26, 27

FERRIER, D. E. K. et P. W. H. HOLLAND. 2001, «Ancient origin of the Hox gene cluster», *Nature Reviews Genetics*, vol. 2, p. 33–38. 72

FESCHOTTE, C. 2008, «Transposable elements and the evolution of regulatory networks», *Nature Reviews Genetics*, vol. 9, p. 397–405. 60, 74

FISHER, R. 1930, «The Genetical Theory of Natural Selection», *Genetics*, vol. 154, p. 1419–1426. URL https://www.genetics.org/content/154/4/1419. 5

FLUTRE, T., E. DUPRAT, C. FEUILLET et H. QUESNEVILLE. 2011, «Considering Transposable Element Diversification in De Novo Annotation Approaches», *PLOS ONE*, vol. 6, p. e16526. URL https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0016526. 50

FLYNN, J. M., R. HUBLEY, C. GOUBERT, J. ROSEN, A. G. CLARK, C. FESCHOTTE et A. F. SMIT. 2020, «RepeatModeler2 for automated genomic discovery of transposable element families», *Proceedings of the National Academy of Sciences*, vol. 117, p. 9451–9457. URL https://www.pnas.org/content/117/17/9451. 50

FRAZEE, A. C., G. PERTEA, A. E. JAFFE, B. LANGMEAD, S. L. SALZBERG et J. T. LEEK. 2015, «Ballgown bridges the gap between transcriptome assembly and expression analysis», *Nature Biotechnology*, vol. 33, p. 243–246. 80

GIFFORD, R. J., J. BLOMBERG, J. M. COFFIN, H. FAN, T. HEIDMANN, J. MAYER, J. STOYE, M. TRISTEM et W. E. JOHNSON. 2018, «Nomenclature for endogenous retrovirus (ERV) loci», *Retrovirology*, vol. 15, nº 1, doi :10.1186/s12977-018-0442-1, p. 59, ISSN 1742-4690. 20

GNAD, F. et J. PARSCH. 2006, «Sebida : a database for the functional and evolutionary analysis of genes with sex-biased expression», *Bioinformatics (Oxford, England)*, vol. 22, p. 2577–2579. 15

GONEN, N., C. R. FUTTNER, S. WOOD, S. A. GARCIA-MORENO, I. M. SALAMONE, S. C. SAMSON, R. SEKIDO, F. POULAT, D. M. MAATOUK et R. LOVELL-BADGE. 2018, «Sex reversal following deletion of a single distal enhancer of Sox9», *Science*, vol. 360, p. 1469–1473. URL http://science.sciencemag.org/content/360/6396/1469. 3

GOUBERT, C., L. MODOLO, C. VIEIRA, C. VALIENTEMORO, P. MAVINGUI et M. BOULESTEIX. 2015, «De Novo Assembly and Annotation of the Asian Tiger Mosquito (Aedes albopictus) Repeatome with dnaPipeTE from Raw Genomic Reads and Comparative Analysis with the Yellow Fever Mosquito (Aedes aegypti)», *Genome Biology and Evolution*, vol. 7, p. 1192–1205. URL https://academic.oup.com/gbe/article/7/4/1192/533768. 49

GRAHAM, P., J. K. M. PENN et P. SCHEDL. 2003, «Masters change, slaves remain», *BioEssays*, vol. 25, p. 1–4. 12

GRATH, S. et J. PARSCH. 2016, «Sex-Biased Gene Expression», *Annual Review of Genetics*, vol. 50, p. 29–44. 59

GREGORY, T. R. 2001, «Coincidence, coevolution, or causation? DNA content, cell size, and the C-value enigma», *Biological Reviews of the Cambridge Philosophical Society*, vol. 76, p. 65–101. 18

GUO, Y., P. K. SINGH et H. L. LEVIN. 2015, «A long terminal repeat retrotransposon of Schizosaccharomyces japonicus integrates upstream of RNA pol III transcribed genes», *Mobile DNA*, vol. 6, p. 19. 73

HAAS, B. J., A. PAPANICOLAOU, M. YASSOUR, M. GRABHERR, P. D. BLOOD, J. BOWDEN, M. B. COUGER, D. ECCLES, B. LI, M. LIEBER, M. D. MACMANES, M. OTT, J. ORVIS, N. POCHET, F. STROZZI, N. WEEKS, R. WESTERMAN, T. WILLIAM, C. N. DEWEY, R. HENSCHEL, R. D. LEDUC, N. FRIEDMAN et A. REGEV. 2013, «De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis», *Nature Protocols*, vol. 8, nº 8, doi :10.1038/nprot.2013.084, p. 1494–1512, ISSN 1750-2799. URL https://www.nature.com/articles/nprot.2013.084, number : 8 Publisher : Nature Publishing Group. 49

HAMILTON, N. E. et M. FERRY. 2018, «ggtern : Ternary Diagrams Using ggplot2», *Journal of Statistical Software*, vol. 87, nº 1, doi :10.18637/jss.v087.c03, p. 1–17, ISSN 1548-7660. URL https://www.jstatsoft.org/index.php/jss/article/view/v087c03, number : 1. 118

HAN, J. S. et J. D. BOEKE. 2005, «LINE-1 retrotransposons : modulators of quantity and quality of mammalian gene expression?», *BioEssays : News and Reviews in Molecular, Cellular and Developmental Biology*, vol. 27, nº 8, doi :10.1002/bies.20257, p. 775–784, ISSN 0265-9247. 20

HAN, K., J. LEE, T. J. MEYER, P. REMEDIOS, L. GOODWIN et M. A. BATZER. 2008, «L1 recombination-associated deletions generate human genomic variation», *Proceedings of the National Academy of Sciences of the United States of America*, vol. 105, nº 49, doi :10.1073/pnas.0807866105, p. 19366–19371, ISSN 1091-6490. 22

HAN, Y. et S. R. WESSLER. 2010, «MITE-Hunter : a program for discovering miniature inverted-repeat transposable elements from genomic sequences», *Nucleic Acids Research*, vol. 38, p. e199. 76

HAYES, T. B. 1998, «Sex determination and primary sex differentiation in amphibians : Genetic and developmental mechanisms», *The Journal of Experimental Zoology*, vol. 281, p. 373–399. URL https://doi.org/10.1002/(SICI)1097-010X(19980801)281:5%3C373::AID-JEZ4%3E3.0.CO;2-L. 6

HECQUET, R. 2018, «La vie secrète du poisson-clown», URL https://lejournal.cnrs.fr/articles/la-vie-secrete-du-poisson-clown. 6

HENCH, K., W. O. MCMILLAN, R. BETANCUR-R et O. PUEBLA. 2017, «Temporal changes in hamlet communities (*Hypoplectrus* spp., Serranidae) over 17 years : TEMPORAL CHANGES IN HAMLET COMMUNITIES», *Journal of Fish Biology*, vol. 91, p. 1475–1490. URL http://doi.wiley.com/10.1111/jfb.13481. 6

HERPIN, A., M. C. ADOLFI, B. NICOL, M. HINZMANN, C. SCHMIDT, J. KLUGHAMMER, M. ENGEL, M. TANAKA, Y. GUIGUEN et M. SCHARTL. 2013, «Divergent expression regulation of gonad development genes in medaka shows incomplete conservation of the downstream regulatory network of vertebrate sex determination», *Molecular Biology and Evolution*, vol. 30, p. 2328–2346. 61, 84

HERPIN, A., I. BRAASCH, M. KRAEUSSLING, C. SCHMIDT, E. C. THOMA, S. NAKAMURA, M. TANAKA et M. SCHARTL. 2010, «Transcriptional rewiring of the sex determining dmrt1 gene

duplicate by transposable elements», *PLoS Genetics*, vol. 6, p. e1000 844. 11, 60, 103

HERPIN, A. et M. SCHARTL. 2015, «Plasticity of gene-regulatory networks controlling sex determination : of masters, slaves, usual suspects, newcomers, and usurpators», *EMBO reports*, vol. 16, p. 1260–1274. 12, 13, 59, 61

HERPIN, A., C. SCHMIDT, S. KNEITZ, C. GOBÉ, M. REGENSBURGER, A. LE CAM, J. MONTFORT, M. C. ADOLFI, C. LILLESAAR, D. WILHELM, M. KRAEUSSLING, B. MOUROT, B. PORCON, M. PANNETIER, E. PAILHOUX, L. ETTWILLER, D. DOLLE, Y. GUIGUEN et M. SCHARTL. 2019, «A novel evolutionary conserved mechanism of RNA stability regulates synexpression of primordial germ cell-specific genes prior to the sex-determination stage in medaka», *PLoS biology*, vol. 17, p. e3000 185. 73

HICKEY, D. A. 1982, «Selfish DNA : a sexually-transmitted nuclear parasite», *Genetics*, vol. 101, p. 519–531. 19

HOF, A. E. V., P. CAMPAGNE, D. J. RIGDEN, C. J. YUNG, J. LINGLEY, M. A. QUAIL, N. HALL, A. C. DARBY et I. J. SACCHERI. 2016, «The industrial melanism mutation in British peppered moths is a transposable element», *Nature*, vol. 534, p. 102–105. URL https://www.nature.com/articles/nature17951. 24

HORIE, Y., T. MYOSHO, T. SATO, M. SAKAIZUMI, S. HAMAGUCHI et T. KOBAYASHI. 2016, «Androgen induces gonadal soma-derived factor, Gsdf, in XX gonads correlated to sex-reversal but not Dmrt1 directly, in the teleost fish, northern medaka (Oryzias sakaizumii)», *Molecular and Cellular Endocrinology*, vol. 436, p. 141–149. 61, 84

HOUWING, S., L. M. KAMMINGA, E. BEREZIKOV, D. CRONEMBOLD, A. GIRARD, H. VAN DEN ELST, D. V. FILIPPOV, H. BLASER, E. RAZ, C. B. MOENS, R. H. A. PLASTERK, G. J. HANNON, B. W. DRAPER et R. F. KETTING. 2007, «A Role for Piwi and piRNAs in Germ Cell Maintenance and Transposon Silencing in Zebrafish», *Cell*, vol. 129, p. 69–82. URL http://www.sciencedirect.com/science/article/pii/S0092867407003923. 72

HUANG, S., L. YE et H. CHEN. 2017, «Sex determination and maintenance : the role of DMRT1 and FOXL2», *Asian Journal of Andrology*, vol. 19, nº 6, doi :10.4103/1008-682X.194420, p. 619–624, ISSN 1745-7262. 13

HUTTER, S., S. S. SAMINADIN-PETER, W. STEPHAN et J. PARSCH. 2008, «Gene expression variation in African and European populations of Drosophila melanogaster», *Genome Biology*, vol. 9, nº 1, doi :10.1186/gb-2008-9-1-r12, p. R12, ISSN 1465-6906. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2395247/. 15

HUYLMANS, A. K., A. MACON et B. VICOSO. 2017, «Global Dosage Compensation Is Ubiquitous in Lepidoptera, but Counteracted by the Masculinization of the Z Chromosome», *Molecular Biology and Evolution*, vol. 34, nº 10, doi :10.1093/molbev/msx190, p. 2637–2649, ISSN 0737-4038. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5850747/. 15

ILLUMINA. 2010, «Illumina Sequencing Technology», , p. 5URL https://www.illumina.com/documents/products/techspotlights/techspotlight_sequencing.pdf. 47

JACQUES, P.-E., J. JEYAKANI et G. BOURQUE. 2013, «The Majority of Primate-Specific Regulatory Sequences Are Derived from Transposable Elements», *PLOS Genetics*, vol. 9, nº 5, doi : 10.1371/journal.pgen.1003504, p. e1003 504, ISSN 1553-7404. URL https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1003504, publisher : Public Library of Science. 23

JANSZ, N. 2019, «DNA methylation dynamics at transposable elements in mammals», *Essays in Biochemistry*, vol. 63, nº 6, doi : 10.1042/EBC20190039, p. 677–689, ISSN 1744-1358. 23

JIN, Y. et M. HAMMELL. 2018, «Analysis of RNA-Seq data using TEtranscripts», *Transcriptome Data Analysis*, p. 153–167. 53

JIN, Y., O. H. TAM, E. PANIAGUA et M. HAMMELL. 2015, «TEtranscripts : a package for including transposable elements in differential expression analysis of RNA-seq datasets», *Bioinformatics*, vol. 31, p. 3593–3599. 49, 53

JORDAN, I. K., I. B. ROGOZIN, G. V. GLAZKO et E. V. KOONIN. 2003, «Origin of a substantial fraction of human regulatory sequences from transposable elements», *Trends in genetics : TIG*, vol. 19, nº 2, doi :10.1016/s0168-9525(02)00006-9, p. 68–72, ISSN 0168-9525. 23

JORDAN ROWLEY, M. et V. G. CORCES. 2018, «Organizational Principles of 3D Genome Architecture», *Nature reviews. Genetics*, vol. 19, p. 789–800. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6312108/. 72

JURKA, J., V. V. KAPITONOV, A. PAVLICEK, P. KLONOWSKI, O. KOHANY et J. WALICHIEWICZ. 2005, «Repbase Update, a database of eukaryotic repetitive elements», *Cytogenetic and Genome Research*, vol. 110, nº 1-4, doi :10.1159/000084979, p. 462–467, ISSN 1424-859X. 20

KAJIKAWA, M. et N. OKADA. 2002, «LINEs mobilize SINEs in the eel through a shared 3' sequence», *Cell*, vol. 111, nº 3, doi : 10.1016/s0092-8674(02)01041-3, p. 433–444, ISSN 0092-8674. 20

KAPITONOV, V. V. et J. JURKA. 2005, «RAG1 core and V(D)J recombination signal sequences were derived from Transib transposons», *PLoS Biology*, vol. 3. 26

KAPITONOV, V. V. et J. JURKA. 2008, «A universal classification of eukaryotic transposable elements implemented in Repbase», *Nature Reviews Genetics*, vol. 9, p. 411–412. URL https://www.nature.com/articles/nrg2165-c1. 19

KAPRANOV, P., G. ST LAURENT, T. RAZ, F. OZSOLAK, C. P. REYNOLDS, P. H. SORENSEN, G. REAMAN, P. MILOS, R. J. ARCECI, J. F. THOMPSON et T. J. TRICHE. 2010, «The majority of total nuclear-encoded non-ribosomal RNA in a human cell is 'dark matter' un-annotated RNA», *BMC Biology*, vol. 8. 71

KAPUSTA, A., Z. KRONENBERG, V. J. LYNCH, X. ZHUO, L. RAMSAY, G. BOURQUE, M. YANDELL et C. FESCHOTTE. 2013, «Transposable elements are major contributors to the origin, diversification, and regulation of vertebrate long noncoding RNAs», *PLoS genetics*, vol. 9, p. e1003 470. 27, 71, 115, 116

KARAKÜLAH, G. et A. SUNER. 2017, «PlanTEnrichment : A tool for enrichment analysis of transposable elements in plants», *Genomics*, vol. 109, p. 336–340. 79

KATOH, K. 2002, «MAFFT : a novel method for rapid multiple sequence alignment based on fast Fourier transform», *Nucleic Acids Research*, vol. 30, p. 3059–3066. URL https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkf436. 77, 118

KATOH, K. et D. M. STANDLEY. 2013, «MAFFT multiple sequence alignment software version 7 : improvements in performance and usability», *Molecular Biology and Evolution*, vol. 30, p. 772–780. URL https://academic.oup.com/mbe/article/30/4/772/1073398. 77, 118

KHAITOVICH, P., I. HELLMANN, W. ENARD, K. NOWICK, M. LEINWEBER, H. FRANZ, G. WEISS, M. LACHMANN et S. PÄÄBO. 2005, «Parallel patterns of evolution in the genomes and transcriptomes of humans and chimpanzees», *Science (New York, N.Y.)*, vol. 309, p. 1850–1854. 15, 59, 114

KIDWELL, M. G. 2002, «Transposable elements and the evolution of genome size in eukaryotes», *Genetica*, vol. 115, p. 49–63. 18

KIKUCHI, K. et S. HAMAGUCHI. 2013, «Novel sex-determining genes in fish and sex chromosome evolution : Novel sex-determining genes in fish», *Developmental Dynamics*, vol. 242, p. 339–353. URL http://doi.wiley.com/10.1002/dvdy.23927. 8, 10, 59, 103, 114

KIKUTA, H., M. LAPLANTE, P. NAVRATILOVA, A. Z. KOMISARCZUK, P. G. ENGSTRÖM, D. FREDMAN, A. AKALIN, M. CACCAMO, I. SEALY, K. HOWE, J. GHISLAIN, G. PEZERON, P. MOURRAIN, S. ELLINGSEN, A. C. OATES, C. THISSE, B. THISSE, I. FOUCHER, B. ADOLF, A. GELING, B. LENHARD et T. S. BECKER. 2007, «Genomic regulatory blocks encompass multiple neighboring genes and maintain conserved synteny in vertebrates», *Genome Research*, vol. 17, p. 545–555. 72

KIM, B. Y., J. WANG, D. E. MILLER, O. BARMINA, E. K. DELANEY, A. THOMPSON, A. A. COMEAULT, D. PEEDE, E. R. D'AGOSTINO, J. PELAEZ, J. M. AGUILAR, D. HAJI, T. MATSUNAGA, E. E. ARMSTRONG, M. ZYCH, Y. OGAWA, M. STAMENKOVIĆ-RADAK, M. JELIĆ, M. S. VESELINOVIĆ, M. TANASKOVIĆ, P. ERIĆ, J.-J. GAO, T. K. KATOH, M. J. TODA, H. WATABE, M. WATADA, J. S. DAVIS, L. C. MOYLE, G. MANOLI, E. BERTOLINI, V. KOŠŤÁL, R. S. HAWLEY, A. TAKAHASHI, C. D. JONES, D. K. PRICE, N. K. WHITEMAN, A. KOPP, D. R. MATUTE et D. A. PETROV. 2020, «Highly contiguous assemblies of 101 drosophilid genomes», doi :10.1101/2020.12.14.422775. URL http://biorxiv.org/lookup/doi/10.1101/2020.12.14.422775. 18

KIM, D., J. M. PAGGI, C. PARK, C. BENNETT et S. L. SALZBERG. 2019, «Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype», *Nature Biotechnology*, vol. 37, p. 907–915. URL https://www.nature.com/articles/s41587-019-0201-4. 49, 51, 76, 79

KIM, P. M., H. Y. K. LAM, A. E. URBAN, J. O. KORBEL, J. AFFOURTIT, F. GRUBERT, X. CHEN, S. WEISSMAN, M. SNYDER et M. B. GERSTEIN. 2008, «Analysis of copy number variants and segmental duplications in the human genome : Evidence for a change in the process of formation in recent evolutionary history», *Genome Research*, vol. 18, nº 12, doi :10.1101/gr.081422.108, p. 1865–1874, ISSN 1088-9051. 22

KISTLER, W. S., D. BAAS, S. LEMEILLE, M. PASCHAKI, Q. SEGUIN-ESTEVEZ, E. BARRAS, W. MA, J.-L. DUTEYRAT, L. MORLÉ, B. DURAND et W. REITH. 2015, «RFX2 Is a Major Transcriptional Regulator of Spermiogenesis», *PLoS genetics*, vol. 11, nº 7, doi :10.1371/journal.pgen.1005368, p. e1005 368, ISSN 1553-7404. 104

KISTLER, W. S., G. C. HORVATH, A. DASGUPTA et M. K. KISTLER. 2009, «Differential expression of Rfx1-4 during mouse spermatogenesis», *Gene Expression Patterns*, vol. 9, p. 515–519. URL http://linkinghub.elsevier.com/retrieve/pii/S1567133X09000763. 116

KNEITZ, S., R. R. MISHRA, D. CHALOPIN, J. POSTLETHWAIT, W. C. WARREN, R. B. WALTER et M. SCHARTL. 2016, «Germ cell and tumor associated piRNAs in the medaka and Xiphophorus melanoma models», *BMC Genomics*, vol. 17. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4869193/. 72

KOBAYASHI, T., T. MYOSHO, J. YAMAMOTO, T. OKAMURA, Y. ONISHI, M. SAKAIZUMI, S. HAMAGUCHI, T. IGUCHI, Y. HORIE, I. T et H. Y. 2017, «Estrogen alters gonadal soma-derived factor (Gsdf)/Foxl2 expression levels in the testes associated with testis-ova differentiation in adult medaka, Oryzias latipes», *Aquat. Toxicol.*, vol. 191, p. 209–218. 61, 84

KOBAYASHI, Y., Y. NAGAHAMA et M. NAKAMURA. 2013, «Diversity and plasticity of sex determination and differentiation in fishes», *Sexual Development*, vol. 7, p. 115–125. URL https://www.karger.com/Article/FullText/342009. 59, 103

KOFLER, R., D. GOMEZ-SANCHEZ et C. SCHLOETTERER. 2016, «PoPoolationTE2 : comparative population genomics of transposable elements using Pool-Seq», *bioRxiv*, p. 038 745. URL http://biorxiv.org/content/early/2016/02/03/038745. 51

KONDO, M. 2006, «Genomic organization of the sex-determining and adjacent regions of the sex chromosomes of medaka», *Genome Research*, vol. 16, p. 815–826. URL http://www.genome.org/cgi/doi/10.1101/gr.5016106. 73

KRAMEROV, D. A. et N. S. VASSETZKY. 2005, «Short retroposons in eukaryotic genomes», *International Review of Cytology*, vol. 247, doi :10.1016/S0074-7696(05)47004-7, p. 165–221, ISSN 0074-7696. 20, 21

KROON, F. J., P. L. MUNDAY et N. W. PANKHURST. 2003, «Steroid hormone levels and bi-directional sex change in Gobiodon histrio», *Journal of Fish Biology*, vol. 62, p. 153–167. URL http://doi.wiley.com/10.1046/j.1095-8649.2003.00017.x. 6

KUMAR, A. et J. L. BENNETZEN. 1999, «Plant retrotransposons», *Annual Review of Genetics*, vol. 33, doi :10.1146/annurev.genet.33.1.479, p. 479–532, ISSN 0066-4197. 20

KUROKAWA, H., D. SAITO, S. NAKAMURA, Y. KATOH-FUKUI, K. OHTA, T. BABA, K.-I. MOROHASHI et M. TANAKA. 2007, «Germ cells are essential for sexual dimorphism in the medaka gonad», *Proceedings of the National Academy of Sciences*, vol. 104, p. 16 958–16 963. URL https://www.pnas.org/content/104/43/16958. 61, 104, 114

LANCIANO, S. et G. CRISTOFARI. 2020, «Measuring and interpreting transposable element expression», *Nature Reviews. Genetics*. 52, 53, 71

LANDER, E., J. SULSTON, WATERSTON et F. R. COLLINS. 2001, «Initial sequencing and analysis of the human genome», *Nature*, vol. 409, nº 6822, doi :10.1038/35057062, p. 860–921, ISSN 1476-4687. URL https://www.nature.com/articles/35057062. 47

LANG, J. W. et H. V. ANDREWS. 1994, «Temperature-dependent sex determination in crocodilians», *Journal of Experimental Zoology*, vol. 270, p. 28–44. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/jez.1402700105. 8

LANGMEAD, B. et S. L. SALZBERG. 2012, «Fast gapped-read alignment with Bowtie 2», *Nature Methods*, vol. 9, p. 357–359. URL http://www.nature.com/nmeth/journal/v9/n4/full/nmeth.1923.html. 49, 51

LAVIALLE, C., G. CORNELIS, A. DUPRESSOIR, C. ESNAULT, O. HEIDMANN, C. VERNOCHET et T. HEIDMANN. 2013, «Paleovirology of 'syncytins', retroviral env genes exapted for a role in placentation», *Philosophical Transactions of the Royal Society B : Biological Sciences*, vol. 368. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3758191/. 26

LEE, J., K. HAN, T. J. MEYER, H.-S. KIM et M. A. BATZER. 2008, «Chromosomal inversions between human and chimpanzee lineages caused by retrotransposons», *PloS One*, vol. 3, nº 12, doi :10.1371/journal.pone.0004047, p. e4047, ISSN 1932-6203. 22

LERAT, E. et P. CAPY. 1999, «Retrotransposons and retroviruses : analysis of the envelope gene», *Molecular Biology and Evolution*, vol. 16, p. 1198–1207. URL https://academic.oup.com/mbe/article-lookup/doi/10.1093/oxfordjournals.molbev.a026210. 20

LERAT, E., M. FABLET, L. MODOLO, H. LOPEZ-MAESTRE et C. VIEIRA. 2016, «TEtools facilitates big data expression analysis of transposable elements and reveals an antagonism between their activity and that of piRNA genes», *Nucleic Acids Research*, vol. 45, p. e17. URL https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkw953. 49, 52, 53, 71

LERCHER, M. J., A. O. URRUTIA et L. D. HURST. 2002, «Clustering of housekeeping genes provides a unified model of gene order in the human genome», *Nature Genetics*, vol. 31, p. 180–183. URL https://www.nature.com/articles/ng887z. 72

LI, Q., B. T. LEE et L. ZHANG. 2005, «Genome-scale analysis of positional clustering of mouse testis-specific genes», *BMC Genomics*, vol. 6. 59, 72

LI, X., W. A. SCARINGE, K. A. HILL, S. ROBERTS, A. MENGOS, D. CARERI, M. T. PINTO, C. K. KASPER et S. S. SOMMER. 2001, «Frequency of recent retrotransposition events in the human factor IX gene», *Human Mutation*, vol. 17, p. 511–519. 19

LISCH, D. 2009, «Epigenetic regulation of transposable elements in plants», *Annual Review of Plant Biology*, vol. 60, doi :10.1146/annurev.arplant.59.032607.092744, p. 43–66, ISSN 1545-2123. 23

LIU, H., M. S. LAMM, K. RUTHERFORD, M. A. BLACK, J. R. GODWIN et N. J. GEMMELL. 2015, «Large-scale transcriptome sequencing reveals novel expression patterns for key sex-related genes in a sex-changing fish», *Biology of Sex Differences*, vol. 6. 14, 70, 71, 114

LONG, M., N. W. VANKUREN, S. CHEN et M. D. VIBRANOVSKI. 2013, «New Gene Evolution : Little Did We Know», *Annual review of genetics*, vol. 47, p. 307–333. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4281893/. 15

LOVE, M. I., W. HUBER et S. ANDERS. 2014, «Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2», *Genome Biology*, vol. 15, p. 550. URL https://doi.org/10.1186/s13059-014-0550-8. 51, 76

LYNCH, V. J., R. D. LECLERC, G. MAY et G. P. WAGNER. 2011, «Transposon-mediated rewiring of gene regulatory networks contributed to the evolution of pregnancy in mammals», *Nature Genetics*, vol. 43, p. 1154–1159. URL http://www.nature.com/doifinder/10.1038/ng.917. 26, 60, 103

LYNCH, V. J., M. C. NNAMANI, A. KAPUSTA, K. BRAYER, S. L. PLAZA, E. C. MAZUR, D. EMERA, S. Z. SHEIKH, F. GRÜTZNER, S. BAUERSACHS, A. GRAF, S. L. YOUNG, J. D. LIEB, F. J. DEMAYO, C. FESCHOTTE et G. P. WAGNER. 2015, «Ancient Transposable Elements Transformed the Uterine Regulatory Landscape and Transcriptome during the Evolution of Mammalian Pregnancy», *Cell Reports*, vol. 10, p. 551–561. URL https://www.cell.com/cell-reports/abstract/S2211-1247(14)01105-X. 26, 103

MANK, J. 2009, «Sex Chromosomes and the Evolution of Sexual Dimorphism : Lessons from the Genome», *The American Naturalist*, vol. 173, p. 141–150. URL https://www.journals.uchicago.edu/doi/10.1086/595754. 13

MANK, J. E. et H. ELLEGREN. 2009, «Sex-linkage of sexually antagonistic genes is predicted by female, but not male, effects in birds», *Evolution ; International Journal of Organic Evolution*, vol. 63, p. 1464–1472. 14

MARNETTO, D., F. MANTICA, I. MOLINERIS, E. GRASSI, I. PESANDO et P. PROVERO. 2018, «Evolutionary Rewiring of Human Regulatory Networks by Waves of Genome Expansion», *The American Journal of Human Genetics*, vol. 102, nº 2, doi :10.1016/j.ajhg.2017.12.014, p. 207–218, ISSN 0002-9297. URL http://www.sciencedirect.com/science/article/pii/S0002929717305037. 23

MATSUDA, M. 2005, «Sex determination in the teleost medaka, Oryzias latipes», *Annu. Rev. Genet.*, vol. 39, p. 293–307. 11

MCCLINTOCK, B. 1950, «The origin and behavior of mutable loci in maize», *Proceedings of the National Academy of Sciences*, vol. 36, p. 344–355. URL http://www.pnas.org/cgi/doi/10.1073/pnas.36.6.344. 17, 22, 23

MCCLURE, M. A. 1991, «Evolution of retroposons by acquisition or deletion of retrovirus-like genes», *Molecular Biology and Evolution*, vol. 8, nº 6, doi :10.1093/oxfordjournals.molbev.a040686, p. 835–856, ISSN 0737-4038. 20

MCLEAY, R. C. et T. L. BAILEY. 2010, «Motif Enrichment Analysis : a unified framework and an evaluation on ChIP data», *BMC Bioinformatics*, vol. 11, doi :10.1186/1471-2105-11-165, p. 165, ISSN 1471-2105. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2868005/. 118

MEDSTRAND, P., L. N. VAN DE LAGEMAAT et D. L. MAGER. 2002, «Retroelement distributions in the human genome : variations associated with age and proximity to genes», *Genome Research*, vol. 12, p. 1483–1495. 74

MEISEL, R. P., J. H. MALONE et A. G. CLARK. 2012, «Disentangling the relationship between sex-biased gene expression and X-linkage», *Genome Research*, vol. 22, p. 1255–1265. 15

MEISEL, R. P., P. U. OLAFSON, K. ADHIKARI, F. D. GUERRERO, K. KONGANTI et J. B. BENOIT. 2020, «Sex Chromosome Evolution in Muscid Flies», *G3 (Bethesda, Md.)*. 59

MITCHESON, Y. S. D. et M. LIU. 2008, «Functional hermaphroditism in teleosts», *Fish and Fisheries*, vol. 9, p. 1–43. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-2979.2007.00266.x. 6

MOLARO, A. et H. S. MALIK. 2016, «Hide and seek : how chromatin-based pathways silence retroelements in the mammalian germline», *Current Opinion in Genetics & Development*, vol. 37, p. 51–58. URL http://www.sciencedirect.com/science/article/pii/S0959437X15001422. 72, 74

MOORE, E. C. et R. B. ROBERTS. 2013, «Polygenic sex determination», *Current Biology*, vol. 23, nº 12, doi : 10.1016/j.cub.2013.04.004, p. R510–R512, ISSN 0960-9822. URL https://www.cell.com/current-biology/abstract/S0960-9822(13)00412-0, publisher : Elsevier. 9

MORENO-HAGELSIEB, G. et K. LATIMER. 2008, «Choosing BLAST options for better detection of orthologs as reciprocal best hits», *Bioinformatics (Oxford, England)*, vol. 24, nº 3, doi : 10.1093/bioinformatics/btm585, p. 319–324, ISSN 1367-4811. 117

MORGAN, T. 1913, *Heredity and Sex*. URL https://www.questia.com/library/70329749/heredity-and-sex. 5

MULLER, H. J. 1932, «Some Genetic Aspects of Sex», *The American Naturalist*, vol. 66, p. 118–138. URL https://www.journals.uchicago.edu/doi/abs/10.1086/280418. 5

MUNDAY, P. 2002, «Bi-directional sex change : testing the growth-rate advantage model», *Behavioral Ecology and Sociobiology*, vol. 52, p. 247–254. URL http://link.springer.com/10.1007/s00265-002-0517-8. 6

MYOSHO, T., H. OTAKE, H. MASUYAMA, M. MATSUDA, Y. KUROKI, A. FUJIYAMA, K. NARUSE, S. HAMAGUCHI et M. SAKAIZUMI. 2012, «Tracing the Emergence of a Novel Sex-Determining Gene in Medaka, Oryzias luzonensis», *Genetics*, vol. 191, nº 1, doi :10.1534/genetics.111.137497, p. 163–170, ISSN 0016-6731. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3338257/. 114

MÉREL, V., M. BOULESTEIX, M. FABLET et C. VIEIRA. 2020, «Transposable elements in Drosophila», *Mobile DNA*, vol. 11, nº 1, doi :10.1186/s13100-020-00213-z, p. 23, ISSN 1759-8753. URL https://doi.org/10.1186/s13100-020-00213-z. 20, 21

MÜLLER, L., S. GRATH, K. VON HECKEL et J. PARSCH. 2012, «Inter- and Intraspecific Variation in Drosophila Genes with Sex-Biased Expression», *International Journal of Evolutionary Biology*, vol. 2012, doi :10.1155/2012/963976, ISSN 2090-8032. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3270394/. 15

NAKAMOTO, M., M. MATSUDA, D.-S. WANG, Y. NAGAHAMA et N. SHIBATA. 2006, «Molecular cloning and analysis of gonadal expression of Foxl2 in the medaka, Oryzias latipes», *Biochemical and Biophysical Research Communications*, vol. 344, p. 353–361. URL https://linkinghub.elsevier.com/retrieve/pii/S0006291X06006206. 61, 84

NAKAMURA, R., Y. MOTAI, M. KUMAGAI, H. NISHIYAMA, N. C. DURAND, K. KONDO, T. KONDO, T. TSUKAHARA, A. SHIMADA, E. L. AIDEN, S. MORISHITA et H. TAKEDA. 2018, «CTCF looping is established during gastrulation in medaka embryos», *bioRxiv*. 73

NANDA, I., M. KONDO, U. HORNUNG, S. ASAKAWA, C. WINKLER, A. SHIMIZU, Z. SHAN, T. HAAF, N. SHIMIZU, A. SHIMA, M. SCHMID et M. SCHARTL. 2002, «A duplicated copy of DMRT1 in the sex-determining region of the Y chromosome of the medaka, Oryzias latipes», *Proceedings of the National Academy of Sciences*, vol. 99, p. 11 778–11 783. URL http://www.pnas.org/cgi/doi/10.1073/pnas.182314699. 11

NAVILLE, M., S. HENRIET, I. WARREN, S. SUMIC, M. REEVE, J.-N. VOLFF et D. CHOURROUT. 2019, «Massive Changes of Genome Size Driven by Expansions of Non-autonomous Transposable Elements», *Current Biology*, vol. 29, p. 1161–1168.e6. URL https://linkinghub.elsevier.com/retrieve/pii/S0960982219301393. 18

NECSULEA, A., M. SOUMILLON, M. WARNEFORS, A. LIECHTI, T. DAISH, U. ZELLER, J. C. BAKER, F. GRÜTZNER et H. KAESSMANN. 2014, «The evolution of lncRNA repertoires and expression patterns in tetrapods», *Nature*, vol. 505, p. 635–640. URL https://www.nature.com/articles/nature12943. 71

NELSON, J. S., T. C. GRANDE et M. V. H. WILSON. 2006, *Fishes of the World, 5th Edition*. 59, 103

NIEHRS, C. et N. POLLET. 1999, «Synexpression groups in eukaryotes», *Nature*, vol. 402, p. 483–487. 73

OHNO, S. 1972, «So much 'junk' DNA in our genome», *Evolution of Genetic Systems, Brookhaven Symp.Biol.*, p. 366–370. URL https://ci.nii.ac.jp/naid/10005720377/. 17

OMS. «Gender and genetics», URL https://www.who.int/genomics/gender/en/index1.html. 4

ORTEGA-RECALDE, O., A. GOIKOETXEA, T. A. HORE, E. V. TODD et N. J. GEMMELL. 2019, «The Genetics and Epigenetics of Sex Change in Fish», *Annual Review of Animal Biosciences*. 12, 13

OTTO, S. P. et T. LENORMAND. 2002, «Resolving the paradox of sex and recombination», *Nature Reviews Genetics*, vol. 3, p. 252–261. URL http://www.nature.com/articles/nrg761. 3, 5

OU, S. et N. JIANG. 2018, «LTR_retriever : A Highly Accurate and Sensitive Program for Identification of Long Terminal Repeat Retrotransposons1», *Plant Physiology*, vol. 176, p. 1410–1422. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5813529/. 50

PACHTER, L. 2011, «Models for transcript quantification from RNA-Seq», *arXiv :1104.3889 [q-bio.GN]*. 80

PAPA, F., N. WINDBICHLER, R. M. WATERHOUSE, A. CAGNETTI, R. D'AMATO, T. PERSAMPIERI, M. K. N. LAWNICZAK, T. NOLAN et P. A. PAPATHANOS. 2017, «Rapid evolution of female-biased genes among four species of Anopheles malaria mosquitoes», *Genome Research*, vol. 27, nº 9, doi :10.1101/gr.217216.116, p. 1536–1548, ISSN 1549-5469. 15

PARK, P. J. 2009, «ChIP–seq : advantages and challenges of a maturing technology», *Nature Reviews Genetics*, vol. 10, nº 10, doi : 10.1038/nrg2641, p. 669–680, ISSN 1471-0064. URL https://www.nature.com/articles/nrg2641. 48

PARK, S., S. HANNENHALLI et S. CHOI. 2014, «Conservation in first introns is positively associated with the number of exons within genes and the presence of regulatory epigenetic signals», *BMC Genomics*, vol. 15, nº 1, doi :10.1186/1471-2164-15-526, p. 526, ISSN 1471-2164. URL http://bmcgenomics.biomedcentral.com/articles/10.1186/1471-2164-15-526. 107

PASCUAL-ANAYA, J., S. D'ANIELLO, S. KURATANI et J. GARCIA-FERNÀNDEZ. 2013, «Evolution of Hox gene clusters in deuterostomes», *BMC Developmental Biology*, vol. 13, p. 26. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3707753/. 72

PASTUZYN, E. D., C. E. DAY, R. B. KEARNS, M. KYRKE-SMITH, A. V. TAIBI, J. MCCORMICK, N. YODER, D. M. BELNAP, S. ERLENDSSON, D. R. MORADO, J. A. BRIGGS, C. FESCHOTTE et J. D. SHEPHERD. 2018, «The Neuronal Gene Arc Encodes a Repurposed Retrotransposon Gag Protein that Mediates Intercellular RNA Transfer», *Cell*, vol. 172, p. 275–288.e18. URL https://linkinghub.elsevier.com/retrieve/pii/S0092867417315040. 27

PAVLICEV, M., K. HIRATSUKA, K. A. SWAGGART, C. DUNN et L. MUGLIA. 2015, «Detecting endogenous retrovirus-driven tissue-specific gene transcription», *Genome Biology and Evolution*, vol. 7, p. 1082–1097. 23

PERTEA, M., D. KIM, G. M. PERTEA, J. T. LEEK et S. L. SALZBERG. 2016, «Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown», *Nature Protocols*, vol. 11, p. 1650–1667. URL http://www.nature.com/doifinder/10.1038/nprot.2016.095. 60, 76, 79, 80

PERTEA, M., G. M. PERTEA, C. M. ANTONESCU, T.-C. CHANG, J. T. MENDELL et S. L. SALZBERG. 2015, «StringTie enables improved reconstruction of a transcriptome from RNA-seq reads», *Nature Biotechnology*, vol. 33, p. 290–295. URL https://www.nature.com/articles/nbt.3122. 76, 80

PLATT, R. N., L. BLANCO-BERDUGO et D. A. RAY. 2016, «Accurate Transposable Element Annotation Is Vital When Analyzing New Genome Assemblies», *Genome Biology and Evolution*, vol. 8, p. 403–410. 50

POULTER, R. T. M. et T. J. D. GOODWIN. 2005, «DIRS-1 and the other tyrosine recombinase retrotransposons», *Cytogenetic and Genome Research*, vol. 110, nº 1-4, doi :10.1159/000084991, p. 575–588, ISSN 1424-8581, 1424-859X. URL https://www.karger.com/Article/FullText/84991, publisher : Karger Publishers. 21

PRICE, A. L., N. C. JONES et P. A. PEVZNER. 2005, «De novo identification of repeat families in large genomes», *Bioinformatics*, vol. 21, p. i351–i358. URL https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/bti1018. 50

PRICE, M. N., P. S. DEHAL et A. P. ARKIN. 2010, «FastTree 2 – Approximately maximum-likelihood trees for large alignments», *PLOS ONE*, vol. 5, p. e9490. URL https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0009490. 77

QUESNEVILLE, H., C. M. BERGMAN, O. ANDRIEU, D. AUTARD, D. NOUAUD, M. ASHBURNER et D. ANXOLABEHERE. 2005, «Combined Evidence Annotation of Transposable Elements in Genome Sequences», *PLOS Computational Biology*, vol. 1, p. e22. URL https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.0010022. 50

QUESNEVILLE, H., D. NOUAUD et D. ANXOLABÉHÈRE. 2003, «Detection of new transposable element families in Drosophila melanogaster and Anopheles gambiae genomes», *Journal of Molecular Evolution*, vol. 57 Suppl 1, p. S50–59. 50

RAMIALISON, M., R. REINHARDT, T. HENRICH, B. WITTBRODT, T. KELLNER, C. M. LOWY et J. WITTBRODT. 2012, «Cis-regulatory properties of medaka synexpression groups», *Development (Cambridge, England)*, vol. 139, p. 917–928. 73

RAYAN, N. A., R. C. H. DEL ROSARIO et S. PRABHAKAR. 2016, «Massive contribution of transposable elements to mammalian regulatory sequences», *Seminars in Cell & Developmental Biology*, vol. 57, doi :10.1016/j.semcdb.2016.05.004, p. 51–56, ISSN 1096-3634. 23

REBOLLO, R., K. MICELI-ROYER, Y. ZHANG, S. FARIVAR, L. GAGNIER et D. L. MAGER. 2012a, «Epigenetic interplay between mouse endogenous retroviruses and host genes», *Genome Biology*, vol. 13, nº 10, doi :10.1186/gb-2012-13-10-r89, p. R89, ISSN 1474-760X. 23

REBOLLO, R., M. T. ROMANISH et D. L. MAGER. 2012b, «Transposable Elements : An abundant and natural source of regulatory sequences for host genes», *Annual Review of Genetics*, vol. 46, p. 21–42. URL http://dx.doi.org/10.1146/annurev-genet-110711-155621. 23, 60, 74, 115

ROBLEDO, D., L. RIBAS, R. CAL, L. SÁNCHEZ, F. PIFERRER, P. MARTÍNEZ et A. VIÑAS. 2015, «Gene expression analysis at the onset of sex differentiation in turbot (Scophthalmus maximus)», *BMC Genomics*, vol. 16. 14, 70, 114

ROSS, R. M. 1990, «The evolution of sex-change mechanisms in fishes», *Environmental Biology of Fishes*, vol. 29, p. 81–93. URL http://link.springer.com/10.1007/BF00005025. 6

ROY, P. J., J. M. STUART, J. LUND et S. K. KIM. 2002, «Chromosomal clustering of muscle-expressed genes in Caenorhabditis elegans», *Nature*, vol. 418, p. 975–979. URL https://www.nature.com/articles/nature01012. 72

ROZE, D. 2012, «Disentangling the Benefits of Sex», *PLoS Biology*, vol. 10, p. e1001321. URL https://dx.plos.org/10.1371/journal.pbio.1001321. 5, 6

SAARISTO, M., B. B. M. WONG, L. MINCARELLI, A. CRAIG, C. P. JOHNSTONE, M. ALLINSON, K. LINDSTROM et J. A. CRAFT. 2017, «Characterisation of the transcriptome of male and female wild-type guppy brains with RNA-Seq and consequences of exposure to the pharmaceutical pollutant, 17 alpha-ethinyl estradiol», *Aquatic Toxicology*, vol. 186, p. 28–39. 14, 71, 114

SABOT, F. et A. H. SCHULMAN. 2006, «Parasitism and the retrotransposon life cycle in plants : a hitchhiker's guide to the genome», *Heredity*, vol. 97, nº 6, doi :10.1038/sj.hdy.6800903, p. 381–388, ISSN 0018-067X. 20

SALIS, P., T. LORIN, V. LEWIS, C. REY, A. MARCIONETTI, M.-L. ESCANDE, N. ROUX, L. BESSEAU, N. SALAMIN, M. SÉMON, D. PARICHY, J.-N. VOLFF et V. LAUDET. 2019, «Developmental and comparative transcriptomic identification of iridophore contribution to white barring in clownfish», *Pigment Cell & Melanoma Research*, vol. 32, p. 391–402. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/pcmr.12766. 25

SANMIGUEL, P., A. TIKHONOV, Y. K. JIN, N. MOTCHOULSKAIA, D. ZAKHAROV, A. MELAKE-BERHAN, P. S. SPRINGER, K. J. EDWARDS, M. LEE, Z. AVRAMOVA et J. L. BENNETZEN. 1996, «Nested retrotransposons in the intergenic regions of the maize genome», *Science (New York, N.Y.)*, vol. 274, nº 5288, doi : 10.1126/science.274.5288.765, p. 765–768, ISSN 0036-8075. 20

SANTOS, M. E., I. BRAASCH, N. BOILEAU, B. S. MEYER, L. SAUTEUR, A. BÖHNE, H.-G. BELTING, M. AFFOLTER et W. SALZBURGER. 2014, «The evolution of cichlid fish egg-spots is linked with a *cis*-regulatory change», *Nature Communications*, vol. 5, p. 5149. URL https://www.nature.com/articles/ncomms6149. 24, 25

SCHARTL, M., S. SCHORIES, Y. WAKAMATSU, Y. NAGAO, H. HASHIMOTO, C. BERTIN, B. MOUROT, C. SCHMIDT, D. WILHELM, L. CENTANIN, Y. GUIGUEN et A. HERPIN. 2018, «Sox5 is involved in germ-cell regulation and sex determination in medaka following co-option of nested transposable elements», *BMC Biology*, vol. 16. 13

SCHLUPP, I., R. RIESCH et M. TOBLER. 2007, «Amazon mollies», *Current Biology*, vol. 17, p. R536–R537. URL http://www.sciencedirect.com/science/article/pii/S0960982207013929. 7

SCHMIDT, D., P. C. SCHWALIE, M. D. WILSON, B. BALLESTER, A. GONÇALVES, C. KUTTER, G. D. BROWN, A. MARSHALL, P. FLICEK et D. T. ODOM. 2012, «Waves of retrotransposon expansion remodel genome organization and CTCF binding in multiple mammalian lineages», *Cell*, vol. 148, nº 1-2, doi :10.1016/j.cell.2011.11.058, p. 335–348, ISSN 1097-4172. 26

SCHONES, D. E., K. CUI, S. CUDDAPAH, T.-Y. ROH, A. BARSKI, Z. WANG, G. WEI et K. ZHAO. 2008, «Dynamic regulation of nucleosome positioning in the human genome», *Cell*, vol. 132, nº 5, doi :10.1016/j.cell.2008.02.022, p. 887–898, ISSN 1097-4172. 48

SCHULTHEIS, C., A. BOUMLHNE, M. SCHARTL, J. VOLFF et D. GALIANA-ARNOUX. 2009, «Sex Determination Diversity and Sex Chromosome Evolution in Poeciliid Fish», *Sexual Development*, vol. 3, nº 2-3, doi :10.1159/000223072, p. 68–77, ISSN 1661-5433, 1661-5425. URL https://www.karger.com/Article/FullText/223072. 114

SEBERG, O. et G. PETERSEN. 2009, «A unified classification system for eukaryotic transposable elements should reflect their phylogeny», *Nature Reviews Genetics*, vol. 10, p. 276. URL https://www.nature.com/articles/nrg2165-c3. 19

SESSEGOLO, C., N. BURLET et A. HAUDRY. 2016, «Strong phylogenetic inertia on genome size and transposable element content among 26 species of flies», *Biology Letters*, vol. 12. 18

SHAO, H. et Z. TU. 2001, «Expanding the diversity of the IS630-Tc1-mariner superfamily : discovery of a unique DD37E transposon and reclassification of the DD37D and DD39D transposons», *Genetics*, vol. 159, nº 3, p. 1103–1115, ISSN 0016-6731. 22

SHEN, S., L. LIN, J. J. CAI, P. JIANG, E. J. KENKEL, M. R. STROIK, S. SATO, B. L. DAVIDSON et Y. XING. 2011, «Widespread establishment and regulatory impact of Alu exons in human genes», *Proceedings of the National Academy of Sciences of the United States of America*, vol. 108, nº 7, doi :10.1073/pnas.1012834108, p. 2837–2842, ISSN 1091-6490. 23

SHEN, X., H. YAN, L. ZHANG, Z. YUAN, W. LIU, Y. WU, Q. LIU, X. LUO et Y. LIU. 2020, «Transcriptomic analyses reveal novel genes with sexually dimorphic expression in Takifugu rubripes brain during gonadal sex differentiation», *Genes & Genomics*, vol. 42, p. 425–439. 14, 71, 114

SHIBA, T. et K. SAIGO. 1983, «Retrovirus-like particles containing RNA homologous to the transposable element copia in Drosophila melanogaster», *Nature*, vol. 302, nº 5904, doi :10.1038/302119a0, p. 119–124, ISSN 0028-0836. 20

SIMONTI, C. N., M. PAVLIČEV et J. A. CAPRA. 2017, «Transposable element exaptation into regulatory regions is rare, influenced by evolutionary age, and subject to pleiotropic constraints», *Molecular Biology and Evolution*, vol. 34, p. 2856–2869. URL https://academic.oup.com/mbe/article/34/11/2856/4082767. 74, 115

SINGER, G. A. C., A. T. LLOYD, L. B. HUMINIECKI et K. H. WOLFE. 2005, «Clusters of co-expressed genes in mammalian genomes are conserved by natural selection», *Molecular Biology and Evolution*, vol. 22, p. 767–775. URL https://academic.oup.com/mbe/article/22/3/767/1076036. 72

SORRELLS, T. R. et A. D. JOHNSON. 2015, «Making sense of transcription networks», *Cell*, vol. 161, p. 714–723. 60

SOUMILLON, M., A. NECSULEA, M. WEIER, D. BRAWAND, X. ZHANG, H. GU, P. BARTHÈS, M. KOKKINAKI, S. NEF, A. GNIRKE, M. DYM, B. DE MASSY, T. MIKKELSEN et H. KAESSMANN. 2013, «Cellular source and mechanisms of high transcriptome complexity in the mammalian testis», *Cell Reports*, vol. 3, p. 2179–2190. URL http://www.sciencedirect.com/science/article/pii/S2211124713002489. 74

SPALLER, T., E. KLING, G. GLÖCKNER, F. HILLMANN et T. WINCKLER. 2016, «Convergent evolution of tRNA gene targeting preferences in compact genomes», *Mobile DNA*, vol. 7, p. 17. 73

STEPHENS, M. 2017, «False discovery rates : a new deal», *Biostatistics*, vol. 18, p. 275–294. URL https://academic.oup.com/biostatistics/article/18/2/275/2557030.

STOREY, J. D., A. J. BASS, A. DABNEY et D. ROBINSON. 2019, *qvalue : Q-value estimation for false discovery rate control*. URL http://github.com/jdstorey/qvalue. 78

SULTANA, T., A. ZAMBORLINI, G. CRISTOFARI et P. LESAGE. 2017, «Integration site selection by retroviruses and transposable elements in eukaryotes», *Nature Reviews. Genetics*, vol. 18, p. 292–308. 73

SUN, Q., Q. HAO et K. V. PRASANTH. 2018, «Nuclear long noncoding RNAs : key regulators of gene expression», *Trends in genetics : TIG*, vol. 34, p. 142–157. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6002860/. 71

SUNDARAM, V., Y. CHENG, Z. MA, D. LI, X. XING, P. EDGE, M. P. SNYDER et T. WANG. 2014, «Widespread contribution of transposable elements to the innovation of gene regulatory networks», *Genome Research*, vol. 24, p. 1963–1976. 23, 74, 115

SUNDARAM, V., M. N. K. CHOUDHARY, E. PEHRSSON, X. XING, C. FIORE, M. PANDEY, B. MARICQUE, M. UDAWATTA, D. NGO, Y. CHEN, A. PAGUNTALAN, T. RAY, A. HUGHES, B. A. COHEN et T. WANG. 2017, «Functional cis-regulatory modules encoded by mouse-specific endogenous retrovirus», *Nature Communications*, vol. 8. 103

SUNDARAM, V. et T. WANG. 2017, «Transposable Element Mediated Innovation in Gene Regulatory Landscapes of Cells : Re-Visiting the "Gene-Battery" Model», *BioEssays*, p. 1700 155. 26

SUNDARAM, V. et J. WYSOCKA. 2020, «Transposable elements as a potent source of diverse cis-regulatory sequences in mammalian genomes», *Philosophical Transactions of the Royal Society B : Biological Sciences*, vol. 375, nº 1795, doi :10.1098/rstb.2019.0347, p. 20190 347. URL https://royalsocietypublishing.org/doi/10.1098/rstb.2019.0347. 23

TAKEHANA, Y., S. HAMAGUCHI et M. SAKAIZUMI. 2008, «Different origins of ZZ/ZW sex chromosomes in closely related medaka fishes, *Oryzias javanicus* and *O. hubbsi*», *Chromosome Research*, vol. 16, nº 5, doi :10.1007/s10577-008-1227-5, p. 801–811, ISSN 0967-3849, 1573-6849. URL http://link.springer.com/10.1007/s10577-008-1227-5. 114

TAKEHANA, Y., M. MATSUDA, T. MYOSHO, M. L. SUSTER, K. KAWAKAMI, T. SHIN, Y. KOHARA, Y. KUROKI, A. TOYODA, A. FUJIYAMA, S. HAMAGUCHI, M. SAKAIZUMI et K. NARUSE. 2014, «Co-option of Sox3 as the male-determining factor on the Y chromosome in the fish Oryzias dancena», *Nature Communications*, vol. 5. 59

TAKEHANA, Y., M. ZAHM, C. CABAU, C. KLOPP, C. ROQUES, O. BOUCHEZ, C. DONNADIEU, C. BARRACHINA, L. JOURNOT, M. KAWAGUCHI, S. YASUMASU, S. ANSAI, K. NARUSE, K. INOUE, C. SHINZATO, M. SCHARTL, Y. GUIGUEN et A. HERPIN. 2020, «Genome Sequence of the Euryhaline Javafish Medaka, Oryzias javanicus : A Small Aquarium Fish Model for Studies on Adaptation To Salinity», *G3 (Bethesda, Md.)*. 117

TANG, M., C. CECCONI, H. KIM, C. BUSTAMANTE et D. C. RIO. 2005, «Guanosine triphosphate acts as a cofactor to promote assembly of initial P-element transposase-DNA synaptic complexes», *Genes & Development*, vol. 19, nº 12, doi :10.1101/gad.1317605, p. 1422–1425, ISSN 0890-9369. 22

TAO, W., J. CHEN, D. TAN, J. YANG, L. SUN, J. WEI, M. A. CONTE, T. D. KOCHER et D. WANG. 2018, «Transcriptome display during tilapia sex determination and differentiation as revealed by RNA-Seq analysis», *BMC Genomics*, vol. 19. 14, 70, 114

TENG, L., B. HE, P. GAO, L. GAO et K. TAN. 2014, «Discover context-specific combinatorial transcription factor interactions by integrating diverse ChIP-Seq data sets», *Nucleic Acids Research*, vol. 42, p. e24. 74

THE ARABIDOPSIS GENOME INITIATIVE. 2000, «Analysis of the genome sequence of the flowering plant Arabidopsis thaliana», *Nature*, vol. 408, nº 6814, doi :10.1038/35048692, p. 796–815, ISSN 1476-4687. URL https://www.nature.com/articles/35048692, number : 6814 Publisher : Nature Publishing Group. 47

THOMPSON, P., T. MACFARLAN et M. LORINCZ. 2016, «Long Terminal Repeats : From parasitic elements to building blocks of the transcriptional regulatory repertoire», *Molecular Cell*, vol. 62, p. 766–776. URL http://www.sciencedirect.com/science/article/pii/S1097276516300120. 74

TODD, E. V., H. LIU, S. MUNCASTER et N. J. GEMMELL. 2016, «Bending Genders : The Biology of Natural Sex Change in Fish», *Sexual Development*, vol. 10, p. 223–241. URL https://www.karger.com/Article/FullText/449297. 6

TOUBIANA, W., D. ARMISÉN, C. DECHAUD, R. ARBORE et A. KHILA. 2020, «Impact of trait exaggeration on sex-biased gene expression and genome architecture in a water strider», *bioRxiv*, p. 2020.01.10.901 322. URL https://www.biorxiv.org/content/10.1101/2020.01.10.901322v1. 65, 78, 122

TRIZZINO, M., A. KAPUSTA et C. BROWN. 2018, «Transposable elements generate regulatory novelty in a tissue-specific fashion», *BMC Genomics*. URL https://bmcgenomics.biomedcentral.com/articles/10.1186/s12864-018-4850-3. 23

TRIZZINO, M., Y. PARK, M. HOLSBACH-BELTRAME, K. ARACENA, K. MIKA, M. CALISKAN, G. H. PERRY, V. J. LYNCH et C. D. BROWN. 2017, «Transposable elements are the primary source of novelty in primate gene regulation», *Genome Research*, vol. 27, p. 1623–1633. URL http://genome.cshlp.org/lookup/doi/10.1101/gr.218149.116. 23

TSAKOGIANNIS, A., T. MANOUSAKI, J. LAGNEL, N. PAPANIKOLAOU, N. PAPANDROULAKIS, C. C. MYLONAS et C. S. TSIGENOPOULOS. 2018a, «The Gene Toolkit Implicated in Functional Sex in Sparidae Hermaphrodites : Inferences From Comparative Transcriptomics», *Frontiers in Genetics*, vol. 9, p. 749. 14, 114

TSAKOGIANNIS, A., T. MANOUSAKI, J. LAGNEL, A. STERIOTI, M. PAVLIDIS, N. PAPANDROULAKIS, C. C. MYLONAS et C. S. TSIGENOPOULOS. 2018b, «The transcriptomic signature of different sexes in two protogynous hermaphrodites : Insights into the molecular network underlying sex phenotype in fish», *Scientific Reports*, vol. 8. 70, 71

TURATSINZE, J.-V., M. THOMAS-CHOLLIER, M. DEFRANCE et J. VAN HELDEN. 2008, «Using RSAT to scan genome sequences for transcription factor binding sites and cis-regulatory modules», *Nature Protocols*, vol. 3, nº 10, doi :10.1038/nprot.2008.97, p. 1578–1588, ISSN 1750-2799. 118

VALDEBENITO-MATURANA, B. et G. RIADI. 2018, «TEcandidates : prediction of genomic origin of expressed transposable elements using RNA-seq data», *Bioinformatics*, vol. 34, p. 3915–3916. URL https://academic.oup.com/bioinformatics/article/34/22/3915/5026658. 54

VALENZUELA, N. 2008, «Sexual development and the evolution of sex determination», *Sexual Development*, vol. 2, p. 64–72. 4

VALOUEV, A., S. M. JOHNSON, S. D. BOYD, C. L. SMITH, A. Z. FIRE et A. SIDOW. 2011, «Determinants of nucleosome organization in primary human cells», *Nature*, vol. 474, nº 7352, doi :10.1038/nature10002, p. 516–520, ISSN 1476-4687. 48

VANKUREN, N. W. et M. LONG. 2018, «Gene duplicates resolving sexual conflict rapidly evolved essential gametogenesis functions», *Nature Ecology & Evolution*. URL http://www.nature.com/articles/s41559-018-0471-0. 14, 15

VENTER, J. C., M. D. ADAMS, E. W. MYERS, P. W. LI, R. J. MURAL, G. G. SUTTON, H. O. SMITH, M. YANDELL, C. A. EVANS, R. A. HOLT, J. D. GOCAYNE, P. AMANATIDES, R. M. BALLEW, D. H. HUSON, J. R. WORTMAN, Q. ZHANG, C. D. KODIRA, X. H. ZHENG, L. CHEN, M. SKUPSKI, G. SUBRAMANIAN, P. D. THOMAS, J. ZHANG, G. L. G. MIKLOS, C. NELSON, S. BRODER, A. G. CLARK, J. NADEAU, V. A. MCKUSICK, N. ZINDER, A. J. LEVINE, R. J. ROBERTS, M. SIMON, C. SLAYMAN, M. HUNKAPILLER, R. BOLANOS, A. DELCHER, I. DEW, D. FASULO, M. FLANIGAN, L. FLOREA, A. HALPERN, S. HANNENHALLI, S. KRAVITZ, S. LEVY, C. MOBARRY, K. REINERT, K. REMINGTON, J. ABU-THREIDEH, E. BEASLEY, K. BIDDICK, V. BONAZZI, R. BRANDON, M. CARGILL, I. CHANDRAMOULISWARAN, R. CHARLAB, K. CHATURVEDI, Z. DENG, V. D. FRANCESCO, P. DUNN, K. EILBECK, C. EVANGELISTA, A. E. GABRIELIAN, W. GAN, W. GE, F. GONG, Z. GU, P. GUAN, T. J. HEIMAN, M. E. HIGGINS, R.-R. JI, Z. KE, K. A. KETCHUM, Z. LAI, Y. LEI, Z. LI, J. LI, Y. LIANG, X. LIN, F. LU, G. V. MERKULOV, N. MILSHINA, H. M. MOORE, A. K. NAIK, V. A. NARAYAN, B. NEELAM, D. NUSSKERN, D. B. RUSCH, S. SALZBERG, W. SHAO, B. SHUE, J. SUN, Z. Y. WANG, A. WANG, X. WANG, J. WANG, M.-H. WEI, R. WIDES, C. XIAO, C. YAN, A. YAO, J. YE, M. ZHAN, W. ZHANG, H. ZHANG, Q. ZHAO, L. ZHENG, F. ZHONG, W. ZHONG, S. C. ZHU, S. ZHAO, D. GILBERT, S. BAUMHUETER, G. SPIER, C. CARTER, A. CRAVCHIK, T. WOODAGE, F. ALI, H. AN, A. AWE, D. BALDWIN, H. BADEN, M. BARNSTEAD, I. BARROW, K. BEESON, D. BUSAM, A. CARVER, A. CENTER, M. L. CHENG, L. CURRY, S. DANAHER, L. DAVENPORT, R. DESILETS, S. DIETZ, K. DODSON, L. DOUP, S. FERRIERA, N. GARG, A. GLUECKSMANN, B. HART, J. HAYNES, C. HAYNES, C. HEINER, S. HLADUN, D. HOSTIN, J. HOUCK, T. HOWLAND, C. IBEGWAM, J. JOHNSON, F. KALUSH, L. KLINE, S. KODURU, A. LOVE, F. MANN, D. MAY, S. MCCAWLEY, T. MCINTOSH, I. MCMULLEN, M. MOY, L. MOY, B. MURPHY, K. NELSON, C. PFANNKOCH, E. PRATTS, V. PURI, H. QURESHI, M. REARDON, R. RODRIGUEZ, Y.-H. ROGERS, D. ROMBLAD, B. RUHFEL, R. SCOTT, C. SITTER, M. SMALLWOOD, E. STEWART, R. STRONG, E. SUH, R. THOMAS, N. N. TINT, S. TSE, C. VECH, G. WANG, J. WETTER, S. WILLIAMS, M. WILLIAMS, S. WINDSOR, E. WINN-DEEN, K. WOLFE, J. ZAVERI, K. ZAVERI, J. F. ABRIL, R. GUIGÓ, M. J. CAMPBELL, K. V. SJOLANDER, B. KARLAK, A. KEJARIWAL, H. MI, B. LAZAREVA, T. HATTON, A. NARECHANIA, K. DIEMER, A. MURUGANUJAN, N. GUO, S. SATO, V. BAFNA, S. ISTRAIL, R. LIPPERT, R. SCHWARTZ, B. WALENZ, S. YOOSEPH, D. ALLEN, A. BASU, J. BAXENDALE, L. BLICK, M. CAMINHA, J. CARNES-STINE, P. CAULK, Y.-H. CHIANG, M. COYNE, C. DAHLKE, A. D. MAYS, M. DOMBROSKI, M. DONNELLY, D. ELY, S. ESPARHAM, C. FOSLER, H. GIRE, S. GLANOWSKI, K. GLASSER, A. GLODEK, M. GOROKHOV, K. GRAHAM, B. GROPMAN, M. HARRIS, J. HEIL, S. HENDERSON, J. HOOVER, D. JENNINGS, C. JORDAN, J. JORDAN, J. KASHA, L. KAGAN, C. KRAFT, A. LEVITSKY, M. LEWIS, X. LIU, J. LOPEZ, D. MA, W. MAJOROS, J. MCDANIEL, S. MURPHY, M. NEWMAN, T. NGUYEN, N. NGUYEN, M. NODELL, S. PAN, J. PECK, M. PETERSON, W. ROWE, R. SANDERS, J. SCOTT, M. SIMPSON, T. SMITH, A. SPRAGUE, T. STOCKWELL, R. TURNER, E. VENTER, M. WANG, M. WEN, D. WU, M. WU, A. XIA, A. ZANDIEH et X. ZHU. 2001, «The Sequence of the Human Genome», *Science*, vol. 291, nº 5507, doi :10.1126/science.1058040, p. 1304–1351, ISSN 0036-8075, 1095-9203. URL https://science.sciencemag.org/content/291/5507/1304, publisher : American Association for the Advancement of Science Section : Special Reviews. 47

VINCKENBOSCH, N., I. DUPANLOUP et H. KAESSMANN. 2006, «Evolutionary fate of retroposed gene copies in the human genome», *Proceedings of the National Academy of Sciences*, vol. 103, nº 9, doi :10.1073/pnas.0511307103, p. 3220–3225, ISSN 0027-8424, 1091-6490. URL https://www.pnas.org/content/103/9/3220. 15

VITTE, C. et O. PANAUD. 2005, «LTR retrotransposons and flowering plant genome size : emergence of the increase/decrease model», *Cytogenetic and Genome Research*, vol. 110, nº 1-4, doi :10.1159/000084941, p. 91–107, ISSN 1424-859X. 18

VOLFF, J.-N. 2005, «Genome evolution and biodiversity in teleost fish», *Heredity*, vol. 94, p. 280–294. URL https://www.nature.com/articles/6800635. 27

VOLFF, J.-N. 2006, «Turning junk into gold : domestication of transposable elements and the creation of new genes in eukaryotes», *BioEssays : News and Reviews in Molecular, Cellular and Developmental Biology*, vol. 28, p. 913–922. 26

VOLFF, J.-N., L. BOUNEAU, C. OZOUF-COSTAZ et C. FISCHER. 2003, «Diversity of retrotransposable elements in compact pufferfish genomes», *Trends in genetics : TIG*, vol. 19, nº 12, doi :10.1016/j.tig.2003.10.006, p. 674–678, ISSN 0168-9525. 27

VOLFF, J.-N., C. KÖRTING et M. SCHARTL. 2001, «Ty3/Gypsy Retrotransposon Fossils in Mammalian Genomes : Did They Evolve into New Cellular Functions?», *Molecular Biology and Evolution*, vol. 18, nº 2, doi :10.1093/oxfordjournals.molbev.a003801, p. 266–270, ISSN 0737-4038. URL https://doi.org/10.1093/oxfordjournals.molbev.a003801. 26

VOLFF, J.-N., I. NANDA, M. SCHMID et M. SCHARTL. 2007, «Governing sex determination in fish : regulatory putsches and ephemeral dictators», *Sexual Development*, vol. 1, p. 85–99. URL https://www.karger.com/Article/FullText/100030. 59, 103

VOYTAS, D. F. et J. D. BOEKE. 1992, «Yeast retrotransposon revealed», *Nature*, vol. 358, nº 6389, doi :10.1038/358717a0, p. 717, ISSN 0028-0836. 20

VRIJENHOEK, R., R. DAWLEY, C. COLE et J. BOGART. 1989, *Evolution and Cytology of Unisexual Vertebrates*. 5

WAGNER, G. P., K. KIN et V. J. LYNCH. 2012, «Measurement of mRNA abundance using RNA-seq data : RPKM measure is inconsistent among samples», *Theory in Biosciences = Theorie in Den Biowissenschaften*, vol. 131, p. 281–285. 80

WANG, T., J. ZENG, C. B. LOWE, R. G. SELLERS, S. R. SALAMA, M. YANG, S. M. BURGESS, R. K. BRACHMANN et D. HAUSSLER. 2007, «Species-specific endogenous retroviruses shape the transcriptional network of the human tumor suppressor protein p53», *Proceedings of the National Academy of Sciences of the United States of America*, vol. 104, nº 47, doi :10.1073/pnas.0703637104, p. 18 613–18 618, ISSN 1091-6490. 26

WANG, Z., X. QIU, D. KONG, X. ZHOU, Z. GUO, C. GAO, S. MA, W. HAO, Z. JIANG, S. LIU, T. ZHANG, X. MENG et X. WANG. 2017, «Comparative RNA-Seq analysis of differentially expressed genes in the testis and ovary of Takifugu rubripes», *Comp. Biochem. Physiol.*, vol. 22, p. 50–57. URL http://linkinghub.elsevier.com/retrieve/pii/S1744117X17300187. 14, 70, 71, 114

WARNER, J. R. 1999, «The economics of ribosome biosynthesis in yeast», *Trends in Biochemical Sciences*, vol. 24, p. 437–440. URL https://linkinghub.elsevier.com/retrieve/pii/S0968000499014607. 48

WARNER, R. R. et S. E. SWEARER. 1991, «Social Control of Sex Change in the Bluehead Wrasse, Thalassoma bifasciatum (Pisces : Labridae)», *The Biological Bulletin*, vol. 181, p. 199–204. URL https://www.journals.uchicago.edu/doi/abs/10.2307/1542090. 6

WARREN, I. A., M. NAVILLE, D. CHALOPIN, P. LEVIN, C. S. BERGER, D. GALIANA et J.-N. VOLFF. 2015, «Evolutionary impact of transposable elements on genomic diversity and lineage-specific innovation in vertebrates», *Chromosome Research*, vol. 23, p. 505–531. URL https://link.springer.com/article/10.1007/s10577-015-9493-5. 26

WATERHOUSE, A. M., J. B. PROCTER, D. M. A. MARTIN, M. CLAMP et G. J. BARTON. 2009, «Jalview Version 2 : a multiple sequence alignment editor and analysis workbench», *Bioinformatics*, vol. 25, nº 9, doi :10.1093/bioinformatics/btp033, p. 1189–1191, ISSN 1367-4803. URL https://academic.oup.com/bioinformatics/article/25/9/1189/203460, publisher : Oxford Academic.

WATERS, P. D., M. C. WALLIS et J. A. M. GRAVES. 2007, «Mammalian sex—Origin and evolution of the Y chromosome and SRY», *Seminars in Cell & Developmental Biology*, vol. 18, p. 389–400. URL http://www.sciencedirect.com/science/article/pii/S1084952107000468. 59, 73, 103

WICKER, T., F. SABOT, A. HUA-VAN, J. L. BENNETZEN, P. CAPY, B. CHALHOUB, A. FLAVELL, P. LEROY, M. MORGANTE, O. PANAUD et OTHERS. 2007, «A unified classification system for eukaryotic transposable elements», *Nature Reviews Genetics*, vol. 8, p. 973–982. URL http://www.nature.com/nrg/journal/v8/n12/abs/nrg2165.html. 19, 20, 21, 22, 103

WILLIAMS, G. C. et J. B. MITTON. 1973, «Why reproduce sexually?», *Journal of Theoretical Biology*, vol. 39, p. 545–554. 5

DE WIT, E. 2019, «TADs as the Caller Calls Them», *Journal of Molecular Biology*. 72

WOLFENBARGER, L. L. et G. S. WILKINSON. 2001, «Sex-linked expression of a sexually selected trait in the stalk-eyed fly, Cyrtodiopsis dalmanni», *Evolution; International Journal of Organic Evolution*, vol. 55, p. 103–110. 15

WU, J.-J., Y.-L. ZHOU, Z.-W. WANG, G.-H. LI, F.-P. JIN, L.-L. CUI, H.-T. GAO, X.-P. LI, L. ZHOU et J.-F. GUI. 2019, «Comparative transcriptome analysis reveals differentially expressed genes and signaling pathways between male and female red-tail catfish (Mystus wyckioides)», *Marine Biotechnology (New York, N.Y.)*, vol. 21, p. 463–474. 14, 71, 114

WU, Y., X. HU, Z. LI, M. WANG, S. LI, X. WANG, X. LIN, S. LIAO, Z. ZHANG, X. FENG, S. WANG, X. CUI, Y. WANG, F. GAO, R. A. HESS et C. HAN. 2016, «Transcription Factor RFX2 Is a Key Regulator of Mouse Spermiogenesis», *Scientific Reports*, vol. 6, p. 20 435. URL https://www.nature.com/articles/srep20435. 116

XIE, M., C. HONG, B. ZHANG, R. F. LOWDON, X. XING, D. LI, X. ZHOU, H. J. LEE, C. L. MAIRE, K. L. LIGON, P. GASCARD, M. SIGAROUDINIA, T. D. TLSTY, T. KADLECEK, A. WEISS, H. O'GEEN, P. J. FARNHAM, P. A. F. MADDEN, A. J. MUNGALL, A. TAM, B. KAMOH, S. CHO, R. MOORE, M. HIRST, M. A. MARRA, J. F. COSTELLO et T. WANG. 2013, «DNA hypomethylation within specific transposable element families associates with tissue-specific enhancer landscape», *Nature Genetics*, vol. 45, p. 836–841. URL https://www.nature.com/articles/ng.2649. 23

XIONG, W., L. HE, J. LAI, H. K. DOONER et C. DU. 2014, «HelitronScanner uncovers a large overlooked cache of Helitron transposons in many plant genomes», *Proceedings of the National Academy of Sciences of the United States of America*, vol. 111, p. 10 263–10 268. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4104883/. 76

XU, Z. et H. WANG. 2007, «LTR_finder : an efficient tool for the prediction of full-length LTR retrotransposons», *Nucleic Acids Research*, vol. 35, p. W265–W268. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1933203/. 50

YANG, L., Z. ZHANG et S. HE. 2016, «Both male-biased and female-biased genes evolve faster in fish genomes», *Genome Biology and Evolution*, vol. 8, p. 3433–3445. URL https://academic.oup.com/gbe/article-lookup/doi/10.1093/gbe/evw239. 15, 59, 114

YANG, W. R., D. ARDELJAN, C. N. PACYNA, L. M. PAYER et K. H. BURNS. 2019, «SQuIRE reveals locus-specific regulation of interspersed repeat expression», *Nucleic Acids Research*, vol. 47, p. e27. 49, 54, 61, 76

YANO, A., R. GUYOMARD, B. NICOL, E. JOUANNO, E. QUILLET, C. KLOPP, C. CABAU, O. BOUCHEZ, A. FOSTIER et Y. GUIGUEN. 2012, «An Immune-Related Gene Evolved into the Master Sex-Determining Gene in Rainbow Trout, Oncorhynchus mykiss», *Current Biology*, vol. 22, nº 15, doi :10.1016/j.cub.2012.05.045, p. 1423–1428, ISSN 0960-9822. URL https://www.cell.com/current-biology/abstract/S0960-9822(12)00632-X, publisher : Elsevier. 10

YOKOI, H., T. KOBAYASHI, M. TANAKA, Y. NAGAHAMA, Y. WAKAMATSU, H. TAKEDA, K. ARAKI, K.-I. MOROHASHI et K. OZATO. 2002, «Sox9 in a teleost fish, medaka (Oryzias latipes) : evidence for diversified function of Sox9 in gonad differentiation», *Molecular Reproduction and Development*, vol. 63, nº 1, doi :10.1002/mrd.10169, p. 5–16, ISSN 1040-452X. 13

ZENG, Q., S. LIU, J. YAO, Y. ZHANG, Z. YUAN, C. JIANG, A. CHEN, Q. FU, B. SU, R. DUNHAM et Z. LIU. 2016, «Transcriptome display during testicular differentiation of channel catfish (Ictalurus punctatus) as revealed by RNA-seq analysis», *Biology of Reproduction*, vol. 95. 14, 70, 114

ZHANG, L., A. DAWSON et D. J. FINNEGAN. 2001, «DNA-binding activity and subunit interaction of the mariner transposase», *Nucleic Acids Research*, vol. 29, nº 17, doi :10.1093/nar/29.17.3566, p. 3566–3575, ISSN 1362-4962. 22

ZHANG, X., G. GUAN, M. LI, F. ZHU, Q. LIU, K. NARUSE, A. HERPIN, Y. NAGAHAMA, J. LI et Y. HONG. 2016, «Autosomal gsdf acts as a male sex initiator in the fish medaka», *Scientific Reports*, vol. 6. URL http://www.nature.com/articles/srep19738. 114

ZHANG, Y., F. LI, D. SUN, J. LIU, N. LIU et Q. YU. 2011, «Molecular analysis shows differential expression of R-spondin1 in zebrafish (Danio rerio) gonads», *Molecular Biology Reports*, vol. 38, nº 1, doi :10.1007/s11033-010-0105-3, p. 275–282, ISSN 1573-4978. URL https://doi.org/10.1007/s11033-010-0105-3. 13

ZHAO, L., J. WIT, N. SVETEC et D. J. BEGUN. 2015, «Parallel Gene Expression Differences between Low and High Latitude Populations of Drosophila melanogaster and D. simulans», *PLoS Genetics*, vol. 11, nº 5, doi :10.1371/journal.pgen.1005184, ISSN 1553-7390. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4423912/. 15

ZHAO, S., Y. ZHANG, R. GAMINI, B. ZHANG et D. VON SCHACK. 2018, «Evaluation of two main RNA-seq approaches for gene quantification in clinical RNA sequencing : polyA+ selection versus rRNA depletion», *Scientific Reports*, vol. 8, nº 1, doi :10.1038/s41598-018-23226-4, p. 4781, ISSN 2045-2322. URL https://www.nature.com/articles/s41598-018-23226-4, number : 1 Publisher : Nature Publishing Group. 48

ZHOU, Q., A. FROSCHAUER, C. SCHULTHEIS, C. SCHMIDT, G. P. BIENERT, M. WENNING, A. DETTAI et J.-N. VOLFF. 2006, «Helitron transposons on the sex chromosomes of the platyfish *Xiphophorus maculatus* and their evolution in animal genomes», *Zebrafish*, vol. 3, p. 39–52. URL http://www.liebertpub.com/doi/10.1089/zeb.2006.3.39. 69

ZHU, A., J. G. IBRAHIM et M. I. LOVE. 2019, «Heavy-tailed prior distributions for sequence count data : removing the noise and preserving large differences», *Bioinformatics*, vol. 35, p. 2084–2092. URL https://academic.oup.com/bioinformatics/article/35/12/2084/5159452. 76

ÁVILA, V., J. L. CAMPOS et B. CHARLESWORTH. 2015, «The effects of sex-biased gene expression and X-linkage on rates of adaptive protein sequence evolution in Drosophila», *Biology Letters*, vol. 11, nº 4, doi :10.1098/rsbl.2015.0117, ISSN 1744-9561. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4424624/. 15

RÉSUMÉ DE LA THÈSE

Chez les poissons téléostéens, les modes de reproduction sexuée et les réseaux de régulation des gènes liés au sexe sont très variables. La détermination du sexe des espèces gonochoriques, par exemple, peut être aussi bien génétique qu'environnementale et peut impliquer des gènes différents selon les espèces. Les régulations du développement et du maintien du sexe apparaissent également variables dans ce groupe.

Pour essayer de comprendre l'origine de cette diversité, je me suis intéressé à l'impact possible des éléments transposables sur la régulation de gènes liés au sexe chez ces poissons. Les éléments transposables sont des séquences d'ADN endogènes capables de se déplacer ou de se copier dans les génomes. Bien que souvent neutres, et parfois délétères pour leur hôte, les éléments transposables peuvent aussi transporter des séquences régulatrices, comme des sites de fixation de facteurs de transcription, et les disséminer à travers les génomes. Leur forte diversité dans les génomes de poissons constitue un réservoir de séquences régulatrices disponibles.

Pour tester cette hypothèse, j'ai utilisé des données de séquençage d'ARN issues de gonades mâles et femelles d'*Oryzias latipes*, le médaka japonais. Dans un premier temps, j'ai analysé l'expression des gènes et des éléments transposables et mis en évidence des régions du génome enrichies en gènes et éléments transposables différentiellement exprimés entre les gonades mâles et femelles. De plus, les gènes et les éléments transposables proches le long des chromosomes ont tendance à présenter des biais d'expression similaires. Deux hypothèses, non mutuellement exclusives, peuvent rendre compte de cette observation : d'une part, les éléments transposables pourraient modifier l'expression des gènes voisins, et d'autre part, l'environnement génomique du site d'insertion pourrait influencer l'expression des éléments transposables. Les travaux réalisés ne permettent pas de trancher définitivement entre ces deux hypothèses, mais plusieurs observations sont en faveur d'un rôle régulateur de certains éléments transposables. Dans un deuxième axe et de manière complémentaire, j'ai mis en évidence des familles d'éléments transposables physiquement enrichies dans l'environnement des gènes sexe-biaisés. Une famille candidate a été étudiée plus en détail, et j'ai pu mettre en évidence dans ces éléments des sites de fixation de facteurs de transcription connus pour être impliqués dans la fonction sexuelle.

Ces travaux montrent ainsi le rôle potentiel des éléments transposables dans l'évolution rapide de certains réseaux de régulation de gènes et serviront de socle pour de futures études fonctionnelles.