



HAL
open science

Évolution des gènes et génomes après duplication complète du génome chez les poissons téléostéens

Elise Parey

► **To cite this version:**

Elise Parey. Évolution des gènes et génomes après duplication complète du génome chez les poissons téléostéens. Génomique, Transcriptomique et Protéomique [q-bio.GN]. Université Paris sciences et lettres, 2021. Français. NNT : 2021UPSLE008 . tel-03696774

HAL Id: tel-03696774

<https://theses.hal.science/tel-03696774>

Submitted on 16 Jun 2022

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

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



THÈSE DE DOCTORAT
DE L'UNIVERSITÉ PSL

Préparée à l'Institut de Biologie de l'ENS

**Évolution des gènes et génomes après duplication complète chez
les poissons téléostéens**

Soutenue par

Elise PAREY

Le 14 janvier 2021

École doctorale n°515

Complexité du vivant

Spécialité

Génomique



ENS

Composition du jury :

Hervé ISAMBERT Institut Curie, Paris.	<i>Président</i>
Céline SCORNAVACCA ISEM, Montpellier.	<i>Rapporteuse</i>
Christophe DESSIMOZ UNIL, Lausanne & UCL, Londres.	<i>Rapporteur</i>
Emmanuelle LERAT LBBE, Lyon.	<i>Examinatrice</i>
Yann GUIGUEN INRAE, Rennes.	<i>Examineur</i>
Hugues ROEST CROLLIUS Institut de Biologie de l'ENS, Paris.	<i>Directeur</i>
Camille BERTHELOT Institut de Biologie de l'ENS, Paris.	<i>Co-encadrante</i>

Remerciements

Avant tout, je remercie chaleureusement les membres de mon jury d'avoir participé avec enthousiasme à l'évaluation de mon travail. Merci Hervé, Céline, Christophe, Emmanuelle et Yann. Cela a été à la fois un plaisir et quelque chose d'assez effrayant d'être évaluée par des références comme vous.

A mes directeurs de thèse, Camille et Hugues : je réalise que j'ai eu une chance incroyable de pouvoir bénéficier de votre expertise scientifique et de votre soutien à tous les deux. Merci pour tout ce que vous m'avez appris, merci de m'avoir autant inspirée. Hugues, je ne pense pas que tu aies manqué une seule occasion de m'encourager ou de me féliciter. Sincèrement, merci ! Camille, j'ai beaucoup aimé travailler avec toi, merci pour tout le temps que tu m'as consacré. Et surtout, merci d'avoir toujours été là.

Ce travail n'aurait pas eu la même valeur sans l'ensemble des collaborateurs du consortium GenoFish. Merci à tous pour toutes les discussions que nous avons eues. Merci à Yann d'avoir rassemblé un groupe de travail aussi inspirant et de m'avoir ainsi permis de rencontrer quelques unes des rockstars citées dans mon chapitre d'introduction.

Je tiens à remercier également les membres de mon comité de thèse, Matthieu, Ingrid et Anton pour leur aide précieuse. Anton, une nouvelle fois, merci pour tes conseils, ils ont toujours touché en plein dans le mille. Je dois également remercier Pierre et l'ensemble des membres de la plateforme informatique de l'IBENS, qui nous permettent de toujours travailler dans les meilleures conditions. Merci en particulier à Bilel d'avoir réparé ma machine à plusieurs reprises... ! Enfin, j'ai également eu la chance de pouvoir découvrir l'enseignement : merci à Morgane et Pierre de m'avoir fait confiance pour le monitorat.

Je me dois de remercier l'ensemble des membres du labo. Alex, merci de t'être toujours occupée de moi, comme tu t'occupes toujours de tout le monde. Et merci d'avoir déniché LA formation qui a certainement permis à mon travail de prendre une plus grande dimension. Merci à tous les DYOGEN d'avoir participé à l'ambiance bienveillante et stimulante du labo : merci à Yves, François, Christine, Lambert, Guillaume, Nga, Franklin, Gosia, Axelle et Lada. Merci également à toute la team Canteen pour les bons moments passés ensemble. Merci Swann pour ton soutien à distance, et ton sixième sens pour toujours l'envoyer aux moments les plus critiques. Et merci à Enrique pour les pauses glace improvisées de quelques après-midis.

A ma famille, mes parents et mes soeurs Flore et Alice, merci pour votre soutien sans faille – quels que soient mes choix. A ma seconde famille, Larisa, Thierry et Jean-Daniel aka "JayDee" : vous êtes adorables.

J'en profite également pour décerner une mention spéciale à deux compagnons de rédaction d'un autre genre : merci aux margaritas de la lose et à Everything Not Saved Will Be Lost.

Enfin, naturellement : Спасибо огромное, мой Никитушка. Спасибо за терпение и поддержку. Я никто без тебя.

Table des matières

Table des figures	i
Liste des tableaux	iii
1 Introduction	1
1.1 La génomique comparative des espèces de vertébrés	1
1.1.1 Définitions et objectifs	1
1.1.2 L'annotation des gènes : premier pas vers la fonction des génomes . . .	4
1.1.3 Des génomes aux phénotypes	8
1.2 Les génomes des poissons téléostéens : entre diversité et complexité	15
1.2.1 Le poisson-zèbre, organisme modèle incontournable	15
1.2.2 Phylogénie des poissons téléostéens et projets de séquençage	18
1.2.3 Le défi de l'analyse des génomes de poissons	21
1.3 Les duplications complètes et leur impact sur les génomes de poissons	25
1.3.1 Événements de duplications complètes chez les Téléostéens	25
1.3.2 Conséquences sur l'organisation des gènes et des génomes	27
1.3.3 Origine et succès évolutifs des polyploïdes	29
1.4 Problématiques	37
2 La résolution des arbres de gènes à travers la synténie conservée clarifie l'impact fonctionnel des duplications complètes	39
2.1 Introduction	39
2.1.1 Contexte méthodologique	39
2.1.2 Reconstruction d'arbres de gènes	44
2.1.3 Indicateurs d'incertitude dans les arbres de gènes	46
2.1.4 Méthodes de correction des arbres de gènes	48
2.1.5 Principe de SCORPiOs	50
2.1.6 Apport aux connaissances sur l'évolution après duplication complète . .	52
2.2 Article	56
2.3 Résultats complémentaires	71
2.3.1 Extension du benchmark des arbres corrigés	71
2.3.2 Pertes réciproques de gènes	72
2.3.3 Accélération de SCORPiOs	74

3	Cartographie à haute résolution des régions anciennement tétraploïdes chez les poissons téléostéens	77
3.1	Introduction	77
3.1.1	Érosion des chromosomes dupliqués	77
3.1.2	Caractérisation des régions de synténie doublement conservée	79
3.1.3	Reconstruction de génomes ancestraux en présence de duplication complète	81
3.1.4	Établissement et impact de la carte des régions dupliquées	85
3.2	Article	88
4	Étude de la rediploïdisation suite à la duplication complète des poissons téléostéens	125
4.1	Introduction	125
4.1.1	Recombinaison méiotique et échanges entre homéologues	125
4.1.2	Les modèles de résolution des ohnologues : LORe et AORe	129
4.1.3	Mise en évidence de rediploïdisation lignée-spécifique chez les saumons	130
4.1.4	Résolution des ohnologues chez les poissons téléostéens	132
4.1.5	Potentiel de SCORPiOs pour l'étude des patrons de rediploïdisation . .	132
4.2	Résultats	134
4.2.1	SCORPiOs permet de retrouver et de préciser les patrons de rediploïdisation chez les Salmoninés	134
4.2.2	Caractérisation des patrons de rediploïdisation après la 3R	138
4.3	Matériel et Méthodes	142
4.3.1	Arbres de gènes	142
4.3.2	Identification des arbres incongruents entre séquence et synténie.	143
4.3.3	Clustering des arbres de gènes	143
4.3.4	Lissage par modèles de Markov cachés	144
4.3.5	Tests de vraisemblance	144
4.4	Discussion	145
5	Discussion	149
5.1	Résumé des résultats principaux	149
5.2	Validation de la pertinence biologique de SCORPiOs	149
5.3	Unification des résultats dans la nomenclature des gènes	151
5.4	Perspectives pour l'étude du génome non-codant	152
6	Perspectives	157
6.1	Contribution à la compréhension des génomes de Vertébrés	157
6.2	Pertinence de l'information de synténie pour la reconstruction de phylogénies d'espèces	160
6.3	Vers un enrichissement du modèle de l'arbre de gènes réconcilié	162

6.4	Méthodologies de la génomique comparative et quantité de données	165
6.4.1	Évolution du processus d'inférence d'arbres de gènes	165
6.4.2	Extension de SCORPiOs à la correction de duplications dans le cas général	165
6.5	Conclusion	167

Table des figures

1.1	Phylogénie des Vertébrés et espèces séquencées	4
1.2	Annotation des gènes codants.	6
1.3	Mécanismes de duplication de gènes	10
1.4	Modèles d'évolution de gène après duplication.	12
1.5	Construction d'un arbre de gène	13
1.6	Couverture de l'édition spéciale poisson-zèbre dans <i>Development</i>	16
1.7	Relations d'orthologie et de paralogie	17
1.8	Phylogénie des poissons téléostéens	20
1.9	Divergence de gènes dupliqués orthologues	22
1.10	Régions dupliquées dans les génomes de poissons téléostéens	23
1.11	Synténie conservée et orthologie	24
1.12	Événements de duplications complètes chez les poissons téléostéens	26
1.13	Organisation des génomes d'espèces dupliquées	30
1.14	Formation, établissement et rediploïdisation des polyploïdes	32
1.15	Gènes dupliqués hérités de manière tétrasomique.	34
2.1	Illustration des méthodes d'arbres et de graphes.	40
2.2	Score de confiance des nœuds de duplication	47
2.3	Contraintes d'orthologie et topologies d'arbres possibles.	49
2.4	Utilisation de la synténie conservée pour guider le problème de sélection d'arbre.	51
2.5	Représentation simplifiée de SCORPiOs.	52
2.6	Contribution de gènes dupliqués de différents âges à l'évolution des tissus chez les Mammifères.	55
2.7	Le bulbus arteriosus : une innovation évolutive dans le cœur des poissons téléostéens.	56
2.8	Comparaison des arbres de gènes inférés par SCORPiOs et Generax.	72
2.9	Perte réciproque de gène masquée par la topologie de l'arbre des espèces.	73
2.10	Exemple d'une perte différentielle de gène coïncidant avec la spéciation Clu- peocephala.	74
3.1	Fusions de micro-chromosomes avant la duplication 3R.	79
3.2	Perturbation des adjacences de gènes après duplication complète.	82

3.3	Analogie entre les modèles thématiques et le modèle macrosyntaxique développé par NAKATANI et MCLYSAGHT 2017.	84
3.4	Reconstruction du caryotype de l'ancêtre Teleostei pré-duplication.	86
4.1	Comportement méiotique de diploïdes et tétraploïdes	126
4.2	Mécanismes principaux de recombinaison méiotique	127
4.3	Modèles de résolution des ohnologues	129
4.4	Phylogénie des espèces de Salmonidés considérées dans l'étude de ROBERTSON et al. 2017	130
4.5	Régions chromosomiques à rediploïdisation tardive chez le saumon atlantique	131
4.6	Évolution des clusters de gènes <i>hox</i> chez les Osteoglossocephalai, selon MARTIN et HOLLAND 2014	133
4.7	Phylogénie des espèces Salmoninés utilisées.	135
4.8	Distribution des conflits séquences/synténie dans les génomes de Salmoninés.	136
4.9	Validation des régions AORE et LORe chez les Salmoninés.	137
4.10	Phylogénie des espèces Osteoglossocephalai utilisées.	138
4.11	Distribution des conflits séquences/synténie dans les génomes d'Osteoglossocephalai.	139
4.12	Topologies contraintes AORE et LORe.	140
4.13	Régions AORE et LORe sur le génome du medaka.	142
5.1	Annotation évolutive des gènes de poissons téléostéens.	152
6.1	Hypothèses de positionnement des duplications 1R-2R.	159
6.2	Phylogénie des poissons reconstruite à partir de données de synténie.	161
6.3	Réconciliation de topologie d'arbre de type LORe.	164
6.4	Identification d'erreurs dans les arbres à travers les graphes d'adjacences conservées.	166

Liste des tableaux

2.1	Principales méthodes de prédiction d'orthologie utilisées à grande échelle. . .	42
4.1	Résultats du clustering des arbres de gènes Salmoninés.	136

Chapitre 1

Introduction

1.1 La génomique comparative des espèces de vertébrés

1.1.1 Définitions et objectifs

De nombreuses métaphores se sont développées pour illustrer la motivation première de la génomique : « décrypter le code du vivant » ou encore « déchiffrer le livre de la vie ». Le génome désigne la séquence complète d'ADN disposée en chromosomes et contenue dans les noyaux de chaque cellule d'un organisme. Chez les Vertébrés, la taille de cette séquence peut aller de 350 millions à 130 milliards de bases (KAPUSTA, SUH et FESCHOTTE 2017). Les images intuitives du génome comme un livre ou un code mettent en avant la notion d'information, ici génétique, qu'il contient. Il s'agit de l'information nécessaire à la formation, survie et reproduction d'un individu. Cependant, le génome est une entité dynamique, soumis aux forces évolutives de mutations, dérive et sélection. Immédiatement, plusieurs questions théoriques peuvent se poser concernant le fonctionnement des génomes. A quel point sont-ils robustes, ou, pour user un peu plus des métaphores sus-citées, combien de modifications peuvent casser le programme de la vie ? Qu'est-ce qui différencie et/ou unit les génomes de différentes espèces ? Par exemple, combien de différences existe-t-il entre un génome humain, de souris, de poisson clown ? Ces premières questions en entraînent une multitude d'autres, représentant des enjeux majeurs pour les domaines de la médecine, de l'agronomie et des biotechnologies. Cette sous-partie vise à introduire l'émergence et les principes généraux de la génomique comparative moderne.

Émergence et principes de la génomique comparative

Disséquer le fonctionnement d'un génome commence par l'obtention de sa séquence. Jusqu'à récemment, le séquençage a demeuré le facteur limitant de la génomique comparative. Les premiers génomes complets séquencés furent ceux de la bactérie *Haemophilus influenzae* en 1995 (FLEISCHMANN et al. 1995), suivi du premier génome eukaryote, la levure, en 1996 (GOFFEAU et al. 1996). Le premier génome de vertébré, le poisson-globe (ou

Fugu), fut obtenu en 2002 (APARICIO et al. 2002). Tandis que ces génomes sont relativement compacts, l'assemblage du génome humain, 8 fois plus grand que celui du poisson Fugu, fut un véritable défi. Achevé en 2003, il nécessita l'implication de plus de 20 instituts, pour un coût estimé à 2,7 milliards de dollars (MEADOWS et LINDBLAD-TOH 2017).

Une fois passé l'enthousiasme suscité par cet accomplissement monumental, il est devenu évident que comprendre le génome humain nécessitait de le comparer à d'autres génomes d'espèces apparentées : c'est l'objet de la génomique comparative. L'espèce humaine appartient au grand groupe des Vertébrés, animaux caractérisés par un squelette osseux ou cartilagineux, et dont l'ancêtre est apparu il y a plus de 500 millions d'années. Les Vertébrés comprennent principalement : les Agnathes (poissons sans mâchoire), les Chondrichthyens (poissons cartilagineux), les Actinoptérygiens (poissons à nageoires rayonnées) et les Tétrapodes (« reptiles », Amphibiens, Oiseaux et Mammifères).

L'obtention d'autres génomes de vertébrés a significativement amélioré notre connaissance du génome humain : d'abord en affinant la prédiction de ses gènes (détaillé en 1.1.2), et, plus globalement, en définissant et disséquant sa fraction fonctionnelle (détaillé en 1.1.3). Classiquement, la notion de fonction en biologie est définie par le rôle expérimentalement démontré d'une séquence d'ADN. Dans le cadre de la génomique, elle repose sur des arguments évolutifs. Le paradigme est le suivant : si une séquence est conservée dans plusieurs espèces, elles-mêmes séparées par des millions d'années d'évolution, alors cette séquence doit posséder un rôle fonctionnel préservé par la sélection naturelle. Ce principe a permis de quantifier la part du génome humain sous contrainte sélective : elle occuperait 8% du génome (LINDBLAD-TOH et al. 2011 ; RANDS et al. 2014).

Thématiques et questions de la génomique comparative

La génomique humaine occupe une grande place dans le paysage de la génomique comparative. De nombreuses études ont contribué à mieux comprendre les maladies humaines, ce qui, à terme, permet de proposer de nouveaux traitements (KARLSSON et LINDBLAD-TOH 2008 ; SCHARTL 2014 ; LI et AUWERX 2020). Elles reposent sur l'identification de gènes de maladie dans une espèce modèle, complétée par la validation de l'implication des gènes homologues humains dans les maladies ainsi modélisées. Par exemple, un résultat récent a permis de proposer un traitement contre l'épilepsie, en s'appuyant sur la conservation des réseaux de gènes impliqués chez l'humain et la souris : le traitement par blocage d'un récepteur clé de ce réseau, Csf1R, a été validé par un essai pré-clinique (SRIVASTAVA et al. 2018). Néanmoins, présenter une vision de la recherche en génomique centrée sur l'espèce humaine ne serait pas lui rendre justice. Les thématiques s'articulent également autour d'autres enjeux, comme la conservation des espèces et l'impact du changement climatique sur la biodiversité. Cela passe par l'établissement de l'arbre de la vie (détaillé avec l'exemple de l'établissement de la phylogénie des poissons en 1.2.2), la compréhension des méca-

nismes de diversification des espèces (STEPPAN et SCHENK 2017; SALZBURGER 2018) et la caractérisation des bases moléculaires, génomiques et évolutives d'adaptations (JONES et al. 2012b; LAMICHHANEY et al. 2017).

Cet élargissement de l'éventail des questions évolutives et biologiques examinées est grandement lié à la révolution des méthodes de séquençage. Aujourd'hui, un total d'environ 800 génomes vertébrés sur les ~ 70 000 espèces répertoriées ont été séquencés (Figure 1.1). La qualité de ces génomes reste très variable, allant d'assemblages chromosomiques, où la séquence entière des chromosomes est connue, à des génomes incomplets et fragmentés en des centaines voire milliers de scaffolds. Les scaffolds sont des fragments des séquences de chromosomes, assemblés, à la manière d'un puzzle, à partir des lectures de séquençage chevauchantes. Typiquement, les assemblages chromosomiques sont obtenus par des technologies de séquençage produisant de longues lectures, de l'ordre de plusieurs kilobases (PacBio et Nanopore, RHOADS et AU 2015; LU, GIORDANO et NING 2016), car plus faciles à combiner pour assembler des chromosomes; en opposition aux technologies aux lectures plus courtes, allant jusqu'à ~250 paires de base (Illumina, MINOCHE, DOHM et HIMMELBAUER 2011).

La quantité de génomes disponibles permet d'orienter les comparaisons à différentes échelles phylogénétiques. Comparer les génomes d'espèces ayant divergé proche de la base des Vertébrés, comme l'humain et les poissons, offre un signal fort pour identifier les éléments conservés après 430 millions d'années d'évolution indépendante (BOFFELLI, NOBREGA et RUBIN 2004). Inversement, comparer des espèces proches permet de mettre en évidence des séquences récemment évoluées. Séquencer de nouvelles espèces offre aussi l'opportunité d'étudier les bases de l'évolution convergente, c'est à dire l'évolution indépendantes des mêmes traits. Enfin, comparer des génomes à différents niveaux de l'arbre phylogénétique des vertébrés informe différents processus biologiques. Par exemple, on définit chez les Vertébrés trois grandes ères d'innovations (LOWE et al. 2011). Les réseaux de régulation qui contrôlent le développement étaient déjà bien établis dans l'ancêtre des vertébrés et peuvent donc être étudiés à des temps profonds de divergence. Les récepteurs extra-cellulaires et les voies de signalisations qu'ils contrôlent ont connu une vague d'innovations indépendantes chez les Actinoptérygiens et les Tétrapodes. Enfin, les réseaux de gènes impliqués dans les modifications de protéines ont évolué plus récemment, durant les 100 derniers millions d'années d'évolution des mammifères placentaires. En résumé, le choix des espèces d'une étude comparative dépend des questions biologiques et évolutives posées, ainsi que de la disponibilité et de la qualité de séquençage de leur génome.

L'essor des technologies de séquençage de nouvelle génération a permis d'entrer dans l'ère dite "post-génomique", où le séquençage n'est plus un obstacle à la génomique comparative. Preuve en est, le consortium G10K (Genome 10K) a lancé en 2017 le projet VGP

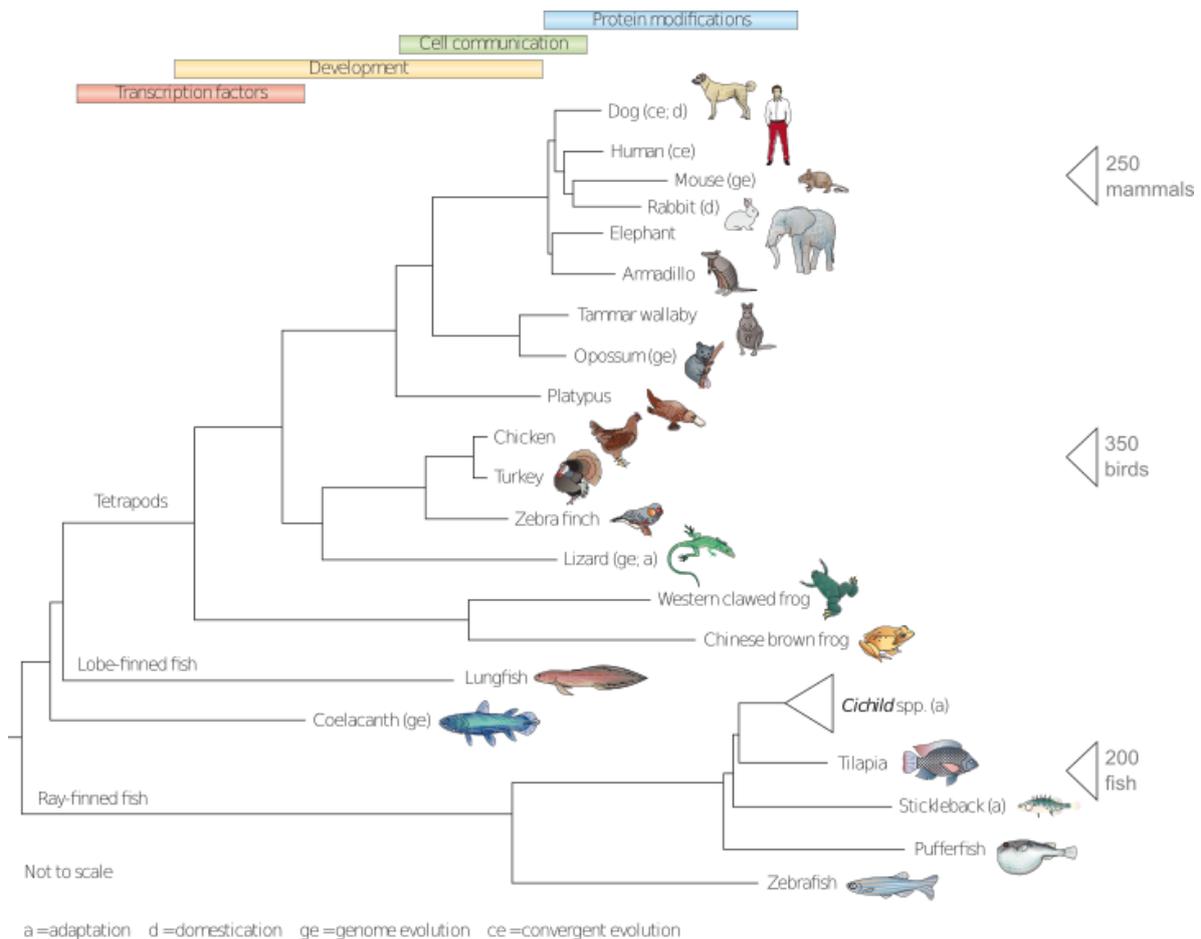


FIGURE 1.1 – Phylogénie des Vertébrés et espèces séquencées. Présentation des espèces séquencées chez les Vertébrés et aperçu des questions examinées par la génomique comparative. Figure tirée de MEADOWS et LINDBLAD-TOH 2017, avec autorisation. Les nombres d'espèces séquencées ont été mis à jour à partir de données de (FAN et al. 2020 ; FENG et al. 2020 ; GENEREUX et al. 2020).

(Vertebrate Genomes Project), visant, à terme, à séquencer la totalité des 70 000 espèces de vertébrés (RHIE et al. 2020). De fait, la génomique se trouve dans un moment à la fois critique et excitant, confrontée aux nombreux défis que représente le déluge de données qui se profile.

1.1.2 L'annotation des gènes : premier pas vers la fonction des génomes

Le gène est l'unité fondamentale fonctionnelle d'un génome. J'emploie dans cette sous-partie le terme de gène en entendant en réalité gène codant : un locus d'ADN transcrit en ARNm et traduit en protéine, molécule effectrice de sa fonction biologique. Déjà théorisé depuis les travaux de Mendel dans les années 1850, le terme de gène ne fut établi qu'en 1909, par Wilhelm Johannsen. A ce moment, un gène était défini comme une unité d'ADN à l'origine d'un trait observable. De premières estimations du nombre de gènes humains ont été proposées entre les années 1940 et 1970. Ces études, basées sur des arguments

mutationnels ou physiques, ont prédit entre 10 000 à 60 000 gènes humains (SPUHLER 1948 ; MULLER 1950 ; VOGEL 1964). Cette large fourchette était difficilement précisable : seule la séquence du génome humain pouvait trancher.

Compter les gènes humains... grâce à un poisson ?

Dans les années 90, avec le lancement du projet de séquençage du génome humain, le mystère du nombre de gènes allait bientôt être percé : la course aux paris était donc ouverte. Plusieurs études indépendantes se sont servies de données transcriptomiques, les marqueurs de séquences exprimées ("expressed sequence tags" ou EST), pour tenter d'estimer le nombre de gènes humains. Les ESTs sont des séquences partielles des transcrits des gènes, sous forme de banques d'ADNc. Dans tous les cas, les banques d'ADNc ne capturaient pas la totalité du génome humain et plusieurs étapes étaient nécessaires afin de dériver le nombre de gènes. Il s'agissait notamment d'éliminer les séquences contaminantes ou autre bruit de transcription, de regrouper les séquences provenant d'un même gène ainsi que d'extrapoler le nombre obtenu pour dériver un comptage sur le génome total. Divers choix méthodologiques pour répondre à ces différents problèmes sont à l'origine d'estimations très variables, allant de 64 000 jusqu'à 120 000 gènes humains prédits (FIELDS et al. 1994 ; ADAMS et al. 1995 ; LIANG et al. 2000).

Au même moment, une approche comparative a proposé d'utiliser le génome du poisson tétraodon, alors assemblé à son tiers, pour s'attaquer au problème (ROEST CROLLIUS et al. 2000). Le principe de la méthode consistait à identifier des séquences homologues (c'est à dire héritées depuis l'ancêtre Vertébré) particulièrement conservées entre les deux espèces. L'hypothèse était que les régions codantes, sous forte contrainte sélective, sont évolutivement plus conservées que les séquences non-codantes. L'algorithme a d'abord été calibré sur des jeux de gènes connus, afin de définir les paramètres d'alignement permettant de discriminer les séquences codantes des non-codantes. L'application de cette méthode aux séquences du génome humain alors disponibles a finalement permis de prédire entre 28 000 et 34 000 gènes humains. Ce nombre ne fut dans un premier temps pas beaucoup repris : il était jugé surprenamment bas. Le nombre de gènes d'*Arabidopsis thaliana* était déjà connu et estimé à 25 000 : comment concevoir que l'on possédait autant de gènes qu'une plante ? Pourtant, par la suite, la publication de la première version du génome humain donna raison à la génomique comparative en estimant environ 30 000 gènes codant des protéines (LANDER et al. 2001), plus tard affinés entre 20 000 et 25 000 (THE ENCODE PROJECT CONSORTIUM 2004). Ainsi, la génomique comparative s'est affirmée, dès l'ère pré-génomique, comme un outil puissant d'analyse des génomes.

Méthodologies actuelles pour l'annotation des gènes codants

Aujourd'hui, la dernière version de la base de données de référence, Ensembl (version 101, Août 2020, YATES et al. 2020), recense 20 440 gènes codants humains. Les génomes des ~240 vertébrés disponibles dans la base comptent entre 15 000 et 50 000 gènes. Le processus d'annotations de gènes s'effectue à travers des procédures sophistiquées, combinant à la fois données expérimentales issues de l'espèce à annoter (RNA-seq, ADNc, séquences protéiques) et jeux de protéines d'autres espèces. Par exemple, la procédure d'annotation d'Ensembl commence par l'établissement d'un premier jeu complet des séquences à potentiel codant, compilé en collectant les séquences prédites par chacune des méthodologies (AKEN et al. 2016). Cet ensemble est ensuite filtré par un système hiérarchique afin de ne garder dans le jeu final que les modèles de gènes les mieux supportés (Figure 1.2). De manière plus générale, la comparaison aux jeux de gènes d'autres espèces intervient également systématiquement à la fin du processus d'annotation : elle sert à évaluer la complétude et la qualité des génomes, en se basant sur des sous-jeux de gènes (~ 3000 chez les Vertébrés) attendus conservés entre espèces proches (méthode BUSCO, WATERHOUSE et al. 2018). Ainsi, les méthodologies d'identification de gènes sont aujourd'hui bien établies et permettent d'annoter le génome codant de manière fiable et systématique. Le défi actuel se situe plutôt au niveau des gènes non-codants, c'est à dire les gènes transcrits en ARNm mais non traduits, dont la séquence est beaucoup moins conservée.

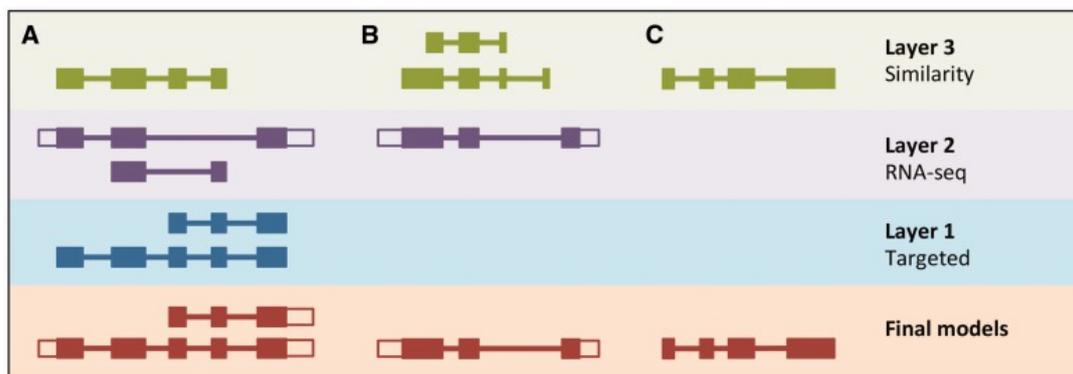


FIGURE 1.2 – Annotation des gènes codants. Dans le processus d'annotation d'Ensembl, différentes couches hiérarchiques permettent l'annotation des gènes (1 en bleu : données protéiques issues de l'espèce annotée, 2 en violet : données RNA-seq issues de l'espèce annotée, 3 en vert : données protéiques issues d'espèces proches). La prédiction finale des gènes (en rouge) s'effectue en utilisant en priorité les modèles des couches 1, puis la 2, puis la 3. Figure tirée de AKEN et al. 2016.

L'annotation systématique des gènes codants permet d'effectuer un premier pas vers la fonction des génomes. A travers l'identification de gènes homologues, les connaissances fonctionnelles peuvent être extrapolées à partir d'espèces modèles vers les espèces non-modèles, en faisant l'hypothèse que la fonction des gènes homologues est conservée. Chez les Vertébrés, environ 80% des gènes peuvent être fonctionnellement annotés par homolo-

gie (NAGY et al. 2020). Les principes du transfert des annotations fonctionnelles sont précisés dans la section consacrée aux espèces modèles de poissons (voir paragraphe 1.2.1).

Annotation des séquences fonctionnelles non-codantes

Comme introduit plus haut, le génome codant représente moins de 2% du génome humain. Le déploiement différentiel de ce génome codant dans des contextes spécifiques (développement, types cellulaires, stimuli environnementaux) est à l'origine de la complexité et plasticité des organismes. Ce sont les éléments régulateurs, constituants du génome fonctionnel non-codant, qui assurent cette expression contexte-spécifique des gènes. Ils comprennent les gènes non-codants, évoqués plus haut, avec notamment les longs ARN non-codants (lincRNA) ; mais aussi les séquences cis-régulatrices des gènes, les promoteurs (régions régulatrices proximales) et enhancers (régions régulatrices distales).

L'annotation du génome fonctionnel non-codant représente un défi plus complexe que celui de l'annotation des gènes codants. Le génome non-codant est dynamique et contexte-spécifique, ses séquences sont moins contraintes et ses mécanismes de fonctionnement moins bien compris. De fait, l'annotation des séquences régulatrices s'effectue à travers la combinaison d'approches de génomique comparative, pour identifier les éléments conservés non-codant (« CNE ») par analyse de séquences ; et de génomique fonctionnelle qui caractérisent empiriquement les propriétés biochimiques de l'ADN (KELLIS et al. 2014). En effet, dans les noyaux, l'ADN est organisé, empaqueté : les régions actives sont marquées par des modifications particulières et sont dans un état de conformation ouverte, accessible à la machinerie transcriptionnelle. L'identification de régions de chromatine ouverte (ATAC-seq) et la combinaison des marquages d'histones (ChiP-Seq) permet de segmenter le génome en différents états fonctionnels (méthodes ChromHMM and Segway, HOFFMAN et al. 2012 ; ERNST et KELLIS 2017). Ces états décrivent, entre autres, les séquences promotrices, enhancers et les lincRNA. Cependant, la génération de ces données fonctionnelles est un processus coûteux. La majorité des efforts actuels sont concentrés sur la caractérisation du génome non-codant humain et de souris (Projet ENCODE, intégrant 164 types cellulaires humains et 66 murins, LIBBRECHT et al. 2019 ; MOORE et al. 2020). De plus, ces analyses définissent uniquement des séquences avec un potentiel régulateur : des validations expérimentales complémentaires sont nécessaires afin de les valider et de confirmer leur(s) gène(s) cible(s).

Contrairement à l'annotation des gènes codants, l'annotation des séquences fonctionnelles non-codantes ne peut pas reposer sur l'extrapolation de données expérimentales d'autres espèces, et ces ressources ne sont disponibles que pour un petit jeu d'espèces. Les analyses comparatives reposent de fait sur l'identification et la comparaison des éléments conservés non-codants. La majorité des CNEs correspondent effectivement à des séquences régulatrices identifiées expérimentalement : il s'agit principalement d'enhancers impliqués

dans la régulation du développement embryonnaire (POLYCHRONOPOULOS et al. 2017). Cependant, les CNEs ne représentent qu'un sous-jeu de l'ensemble des séquences régulatrices qui restent donc encore difficiles à annoter de manière systématique dans les génomes.

1.1.3 Des génomes aux phénotypes

Cette sous-partie introduit les différentes approches visant à lier génome et phénotype, avec un intérêt particulier sur le rôle de l'évolution du contenu en gènes dans l'acquisition de nouveautés évolutives. Le but ultime de la génomique comparative est d'être capable de prédire le lien entre les génomes et les phénotypes. Quelles modifications de l'ADN sont à l'origine de variations observables entre individus, responsables de maladie ou encore à l'origine d'innovations évolutives? Le premier point est plutôt l'objet de la génomique des populations, où il existe des méthodes bien établies pour lier des variants génétiques à des phénotypes. Les études d'association pangénomique (Genome-Wide Association Study, « GWAS », MCCARTHY et al. 2008) ou encore les analyses QTL (Quantitative Trait Loci, COMPLEX TRAIT CONSORTIUM 2003) permettent d'identifier les variants génétiques les plus fortement corrélés à un trait étudié. Toutefois, ces approches ne sont pas directement applicables aux comparaisons entre espèces, car elles ne prennent pas en compte leurs relations évolutives : un signal de similarité génomique peut être due à la proximité entre espèces plus qu'à un rôle fonctionnel.

En partant du gène comme unité fonctionnelle des génomes, il semble intuitif que comparer le contenu en gènes de différentes espèces peut permettre de lier l'absence, la présence ou le nombre de copies de certains gènes à des traits particuliers. Dans la pratique, cela implique de comparer le contenu en gènes d'espèces qui possèdent un trait d'intérêt à d'autres pour lesquelles le trait est absent. Plusieurs étapes critiques sont alors requises : établir correctement les relations d'homologie entre gènes de différentes espèces, détecter des associations significatives en prenant en compte les relations phylogénétiques, puis démontrer la pertinence biologique des gènes mis en évidence. L'interprétation biologique s'appuie généralement soit sur des connaissances *a priori*, soit sur une validation expérimentale. La question du lien génome-phénotype est à considérer sous deux angles principaux : l'angle évolutif ("quels mécanismes évolutifs et moléculaires sont à l'origine d'innovations génomiques et phénotypiques?") et l'angle fonctionnel ("quels gènes sont responsables de quels traits?").

J'ai axé cette sous-partie autour de l'évolution des familles de gènes, avec une attention particulière sur les événements de duplications de gènes. Bien que reconnues comme une source majeure de nouveauté génomique, les duplications de gènes sont loin d'être les seuls mécanismes à leur origine. Ainsi, je ne discute pas des modifications génomiques altérant le génome fonctionnel non-codant. Pourtant, il est important de rappeler que, chez l'humain,

si les gènes occupent environ 1.5 % du génome, la part de génome sous contrainte évolutive est estimée à 8% (LINDBLAD-TOH et al. 2011 ; RANDS et al. 2014). Une partie de ces régions conservées est capable de réguler l'expression des gènes, c'est à dire de moduler la quantité de transcrits produits. Les modifications de l'activité de ces régions régulatrices (insertion d'élément transposable, réarrangement génomique, substitutions) sont largement impliquées dans l'acquisition d'innovations évolutives (KING et WILSON 1975 ; NECSULEA et KAESSMANN 2014).

Évolution du contenu en gènes

Le contenu en gènes évolue via trois grands mécanismes : les créations de gènes *de novo*, les pertes, et les duplications de gènes. Les gènes *de novo* se créent à partir de séquences non-géniques. Ils sont encore mal caractérisés, même si leur contribution à l'évolution du contenu en gènes est réelle (VAKIRLIS, CARVUNIS et MCLYSAGHT 2020). Les pertes de gènes, quant à elles, peuvent se produire par accumulation de mutations rendant la séquence codante non-fonctionnelle (pseudogénéisation), par insertion d'éléments transposables ou par réarrangement génomique. Chez les Vertébrés, l'importance des pertes de gènes dans l'acquisition d'adaptations morphologiques ou physiologiques a été démontrée, même si elles sont généralement estimées neutres (ALBALAT et CAÑESTRO 2016 ; SHARMA et al. 2018). Par exemple, il a été estimé que la perte 85 gènes aurait facilité le retour au milieu aquatique des espèces de cétacés (HUELSMANN et al. 2019).

Depuis les années 1970 et les travaux fondateurs de Susumu Ohno, les duplications de gènes sont considérées comme le moteur majeur des innovations évolutives, parce qu'elles fournissent un substrat supplémentaire à l'évolution moléculaire (OHNO 1970). Ces nouvelles fonctions apparaissent à travers des modifications dans la séquence d'une copie du gène ou dans ses régions régulatrices. On estime qu'au moins 40% des gènes existent en au moins deux copies dans le génome humain (ZHANG 2003 ; MAKINO et MCLYSAGHT 2010 ; SACERDOT et al. 2018). Suivant les mécanismes mis en jeu, les duplications de gènes sont classifiées en : duplications complètes de génome, duplications en tandem, duplications par transposition répllicative, duplications segmentales et duplication par rétro-transposition (Figure 1.3).

Devenir des gènes dupliqués

Dû à la redondance fonctionnelle que les duplications génèrent et quel que soit le mécanisme mis en jeu, le destin le plus fréquent des gènes dupliqués est la pseudogénéisation d'une des deux copies (Figure 1.4, LYNCH 2000). Cette non-fonctionnalisation s'opère à travers la relaxation des contraintes sélectives dans une copie, tandis que l'autre est maintenue par la sélection et assure la fonction ancestrale. L'accumulation de mutations dans les séquences régulatrices et codantes accompagne la pseudogénéisation, en abaissant le niveau

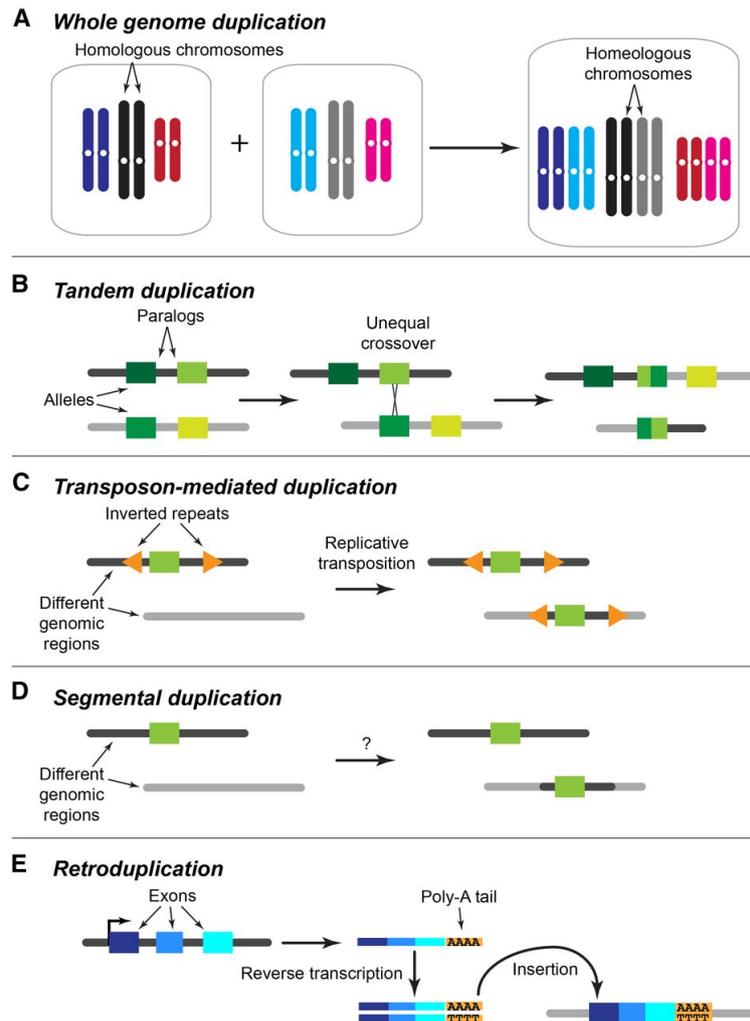


FIGURE 1.3 – Mécanismes de duplication de gènes. A. Duplication complète de génome : par fusion de gamètes non-réduits, la totalité du génome est copié. B. Duplication en tandem : par crossing-over inégal, deux allèles se retrouvent à la suite sur le même chromosome. C. Duplication par transposition : un gène est dupliqué par la mobilisation d'un élément transposable pendant la réplication. D. Duplication segmentale : une large région d'ADN est copiée à un autre endroit du génome à la suite d'un réarrangement génomique. E. Duplication par rétrotransposition : un ARNm est rétro-transcrit en ADNc et réinséré dans le génome. A noter que les régions régulatrices ne sont pas copiées par ce procédé. Figure tirée de PANCHY, LEHTI-SHIU et SHIU 2016.

d'expression et en menant à la production d'une protéine non fonctionnelle.

Alternativement, la fonction d'une ou des deux copies du gène peut être altérée, ce qui brise la redondance fonctionnelle et favorise la rétention des deux copies. Par subfonctionnalisation, les deux copies se partagent la fonction ancestrale : par exemple, pour un gène ancestralement exprimé dans deux tissus, chacun se spécialisera dans l'un des deux (Figure 1.4, FORCE et al. 1999 ; POSTLETHWAIT et al. 2004). Dans ce cas, la subfonctionnalisation est

dite « régulatrice » et peut également passer par un partage du niveau d'expression total. De nombreux exemples de gènes ont suivi cette trajectoire : c'est le cas, par exemple, des gènes *scn4aa* et *scn4ab*, codant des sous-unités de canaux de sodium chez le poisson-zèbre, et présentant chacun un patron d'expression spatio-temporel spécifique dans le muscle au cours du développement (NOVAK et al. 2006). La subfonctionnalisation peut également concerner la séquence codante, si des mutations délétères distinctes sont fixées dans chacune des copies, inactivant différentes sous-fonctions. Initialement, la subfonctionnalisation a été proposée dans le modèle DDC (Duplication Dégénération Complémentation), qui prédit que la subfonctionnalisation est un processus neutre (FORCE et al. 1999). Dans ce modèle, la perte d'une sous-fonction dans l'une des copies entraîne son maintien par sélection négative dans la seconde copie. La propension d'une paire de gènes dupliquée à être subfonctionnalisée corrèle avec sa pléiotropie : plus le paysage régulateur est complexe plus le potentiel de subfonctionnalisation est grand.

Enfin, par néofonctionnalisation, une copie maintient la fonction ancestrale tandis que l'autre en acquiert une nouvelle (Figure 1.4, OHNO 1970). Proposée pour la première fois par Ohno, la néofonctionnalisation représente le mécanisme principal à l'origine de nouveautés génomiques. Comme pour la subfonctionnalisation, la néofonctionnalisation peut être régulatrice, si elle est médiée par l'acquisition de nouveaux patrons d'expression, et/ou codante. Dans ce modèle, la relaxation de la pression sélective sur une des deux copies mène à l'acquisition d'une nouvelle fonction, à travers la fixation de mutations bénéfiques par sélection positive. Chez les poissons électriques, la paire de gènes dupliquées *scn4aa* et *scn4ab* a évolué, de manière convergente chez les Mormyridae et Gymnotiformes, selon ce modèle : une sélection forte sur la séquence *scn4ab* a maintenu son expression dans le muscle tandis que la relaxation des contraintes sur *scn4aa* a entraîné sa divergence et l'acquisition d'une nouvelle expression dans l'organe électrique, accompagné d'une diversification de sa séquence codante par sélection positive (ZAKON et al. 2006).

Ces modèles classiques d'évolution des gènes dupliqués expliquent leur rétention, de manière générale, et sont complétés par d'autres modèles qui s'expriment en fonction du mécanisme à l'origine des duplications. Les modes de rétention dépendent notamment de la duplication ou non des éléments régulateurs associés et/ou de la duplication des gènes en interaction. Dans le cas remarquable des duplications complètes de génomes, la totalité des séquences et gènes sont dupliqués, ce qui implique également la duplication et la complexification de la structure complète de réseaux d'expression et d'interaction de gènes (DE SMET et VAN DE PEER 2012 ; MARLÉTAZ et al. 2018). De ce fait, les propriétés intrinsèques (fonctions biologiques et propriétés génomiques) des gènes amplifiés par duplication complète, ou gènes ohnologues, sont distinctes de celles résultant de duplications à petite échelle (SSD, "Small-scale duplication") (ACHARYA et GHOSH 2016). De plus, les gènes issus de duplications complète ont été montré réfractaires aux duplications SSD (MAKINO et

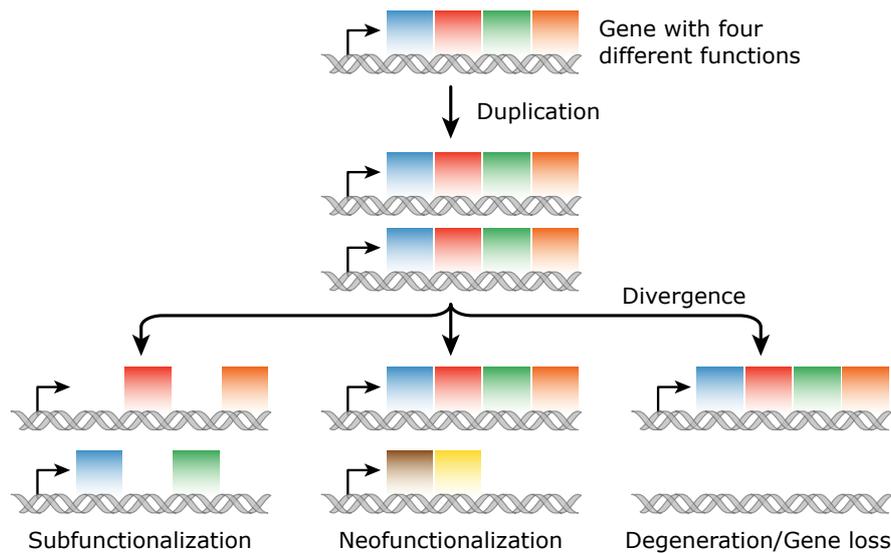


FIGURE 1.4 – Modèles d'évolution de gène après duplication. Exemple d'un gène dupliqué, ancestralement exprimé dans 4 tissus différents (symbolisés par des carrés de couleurs). Par sous-fonctionnalisation, les gènes dupliqués se partagent les tissus d'expression ancestraux. Par néofonctionnalisation, un gène retient l'expression ancestrale et un second acquiert deux nouveaux tissus d'expression. Par non-fonctionnalisation des mutations s'accumulent et entraînent la perte d'une copie.

MCLYSAGHT 2010). Les particularités des modèles expliquant la rétention de gènes ohnologues seront détaillées dans la partie dédiée aux duplications complètes (modèle d'équilibre de dosage, rétention de gènes dominants négatifs, voir le paragraphe 1.3.2).

Duplications de gènes et innovations évolutives

Les innovations évolutives désignent l'acquisition et la fixation de nouvelles caractéristiques morphologiques ou physiologiques dans une population, dont la valeur adaptative peut participer à son succès évolutif. Chez les Vertébrés, l'augmentation du nombre de copie de gènes a été précédemment associée à des innovations évolutives. J'ai évoqué plus haut le rôle de la néofonctionnalisation d'une paire de gènes dupliqués dans l'acquisition de l'organe électrique des poissons électriques. Un autre exemple concerne la duplication de gènes d'opsine chez les poissons Percomorphes, corrélée à leur radiation évolutive (CORTESI et al. 2015). Les gènes d'opsine codent des protéines photo-réceptrices de la rétine. La duplication, suivie de la néofonctionnalisation d'une copie du gène *SWS2A*, a permis aux Percomorphes de sophistication leur vision, facilitant la colonisation de nouveaux milieux. D'autres exemples significatifs comme les duplications des gènes *hox*, à l'origine du plan d'organisation morphologique des Vertébrés, ont démontré l'importance des duplications de gènes dans l'acquisition de nouveaux traits (WAGNER, AMEMIYA et RUDDLE 2003).

Méthodologies de mise en évidence d'associations duplication-innovation

Les événements de duplications de gènes sont mis en évidence à travers la reconstruction de leur histoire évolutive. Cette histoire évolutive est représentée sous forme d'arbre et inférée à partir des séquences de gènes homologues. J'introduis ici brièvement le principe de la reconstruction d'arbres de gènes, qui sera détaillé au chapitre 2 (voir le paragraphe 2.1.2). La méthode la plus couramment employée pour inférer les arbres de gènes à partir d'un alignement de séquences est le maximum de vraisemblance. Un modèle d'évolution décrit comment les séquences évoluent le long des branches de l'arbre, permettant de calculer la probabilité d'observer les séquences. Trouver l'arbre de maximum de vraisemblance (modèle qui maximise la probabilité d'observer les séquences), consiste à explorer l'espace des topologies d'arbre, longueur de branche et paramètres du modèle d'évolution, calculer leur vraisemblance, et retenir la meilleure solution. Une fois cet arbre obtenu il est superposé à l'arbre des espèces pour expliquer les discordances par des événements de duplications et de pertes de gènes (Figure 1.5). Ainsi, la reconstruction des arbres de gènes est au cœur de la caractérisation des patrons de duplication.

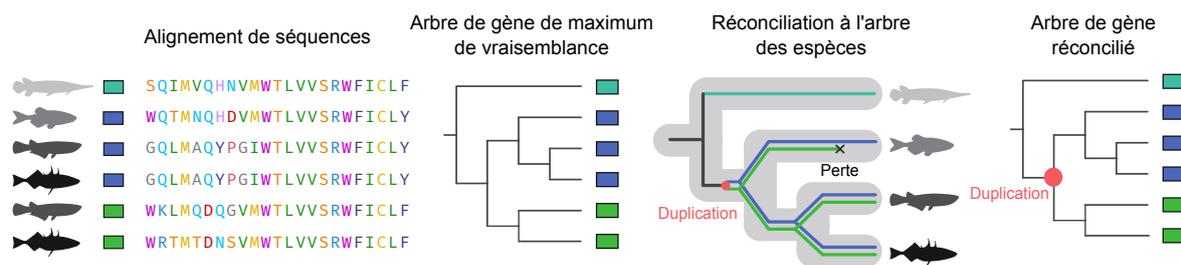


FIGURE 1.5 – Construction d'un arbre de gène. Inférence de l'histoire évolutive d'un gène présent dans 4 espèces de poissons. L'arbre de maximum de vraisemblance est réconcilié à l'arbre des espèces : les discordances sont expliquées par des événements de duplication(s) et perte(s) de gènes. La réconciliation est effectuée par maximum de parcimonie : la solution retenue est celle qui infère le moins d'événements de duplications et pertes (ici une duplication et une perte).

Différentes approches ont été entreprises afin de lier les duplications de gènes à l'acquisition d'innovations. L'approche la plus simple, et peut-être la plus communément utilisée, consiste de partir d'une hypothèse *a priori*. Étant donnée la fonction connue d'un gène et sa pertinence pour le phénotype en question, on s'attache dans un premier temps à reconstruire l'arbre du gène d'intérêt. Ensuite, on confronte ses patrons de duplications aux patrons d'acquisition du trait. Cette stratégie tend à biaiser les études vers ce que NAGY et al. 2020 appellent les 'knowns unknowns' ou 'inconnus connus' : les mêmes familles de gènes sont constamment ré-étudiées. Une seconde approche, agnostique, consiste à parcourir l'ensemble complet des gènes pour en extraire une liste d'intérêt : ceux dont la duplication coïncide avec l'acquisition du trait étudié. Cette liste de gènes peut ensuite être analysée pour valider sa pertinence biologique. Enfin, les méthodes de phylogénie compa-

rative (« Phylogenetic Comparative Methods », PCM) permettent de tester rigoureusement des corrélations dans un contexte phylogénétique. Preuve que les méthodes PCM prennent de l'importance dans le domaine de la génomique comparative, deux nouvelles approches (COMPARE et PAM) ont été développées récemment pour associer duplication de gènes et innovation évolutive (NAGY et al. 2017 ; KIEFER et al. 2019). Ces méthodes permettent de tester statistiquement la corrélation entre nombre de copies de gènes et traits étudiés, tout en prenant en compte la structure de covariance impliquée par la phylogénie des espèces.

Une raison pour laquelle la génomique comparative ne tire pas encore pleinement avantage des PCM, en tout cas dans le contexte des duplications de gènes, est sans doute la difficulté d'identifier les gènes dupliqués correctement. Les méthodes actuelles de reconstruction d'arbre de gènes sont confondues en présence de nombreux événements de duplication. En effet, en présence de plusieurs copies de séquences, l'espace de solution des arbres possibles augmente et les heuristiques qui explorent cet espace peuvent « rater » l'arbre de maximum de vraisemblance. De plus, s'il n'y a pas suffisamment de signal phylogénétique dans les substitutions pour favoriser une topologie d'arbre parmi plusieurs, l'arbre de maximum de vraisemblance reste incertain. Ces problèmes biaisent notamment l'inférence des duplications vers les temps anciens (HAHN 2007). De part la quantité massive de gènes dupliqués qu'elles produisent, les duplications complètes de génomes ont un potentiel significatif d'être à l'origine d'innovation évolutive. Une première étape cruciale avant de pouvoir lier rigoureusement duplication et innovation évolutive est donc d'améliorer les méthodes de phylogénie de gènes, en présence d'événements de duplications.

1.2 Les génomes des poissons téléostéens : entre diversité et complexité

1.2.1 Le poisson-zèbre, organisme modèle incontournable

L'utilisation d'espèces modèles a eu un impact incommensurable sur nos connaissances de la physiologie humaine, rendant possible l'étude de phénomènes biologiques complexes dans un système doté d'outils de manipulation génétique puissants. Pour être informatif vis-à-vis de la biologie humaine, une espèce modèle doit présenter une conservation suffisante des gènes, voies moléculaires, cellulaires et métaboliques sous-jacentes au processus biologique étudié. La souris de laboratoire *Mus musculus* est probablement l'organisme modèle le plus populaire, dû à sa proximité évolutive avec l'espèce humaine. Les poissons, et en particulier le poisson-zèbre, ont également émergé comme des modèles d'intérêt. Le poisson-zèbre, *Danio rerio*, représente un bon compromis entre facilité d'étude et proximité avec le génome humain. Bien que significativement plus distant de l'humain que ne l'est la souris (divergence datée à 430 contre 90 millions d'années), le poisson-zèbre, véritable couteau-suisse des maladies humaines, permet de modéliser à la fois les cancers, les maladies développementales, les défauts immunitaires ou encore les troubles du comportement (voir LIESCHKE et CURRIE 2007, pour une revue détaillée). Il est estimé que 70% des gènes humains impliqués dans des maladies ont un gène homologue chez le poisson-zèbre (BRADFORD et al. 2017). Lieschke et Currie désignent le poisson-zèbre comme un modèle génétiquement aussi puissant qu'un modèle invertébré, mais possédant l'avantage de pouvoir examiner les questions du développement des vertébrés (LIESCHKE et CURRIE 2007).

Établissement du modèle poisson-zèbre

Dès les années 60, le potentiel du poisson-zèbre comme organisme modèle était déjà largement reconnu. Il est facile à élever et reproduire en aquarium : la femelle peut produire aux alentours de 300 œufs par semaine, et ils incubent seulement quelques jours avant éclosion. Une seconde caractéristique remarquable du poisson-zèbre est la transparence de ses œufs et larves, largement exploitée par les pionniers de la génétique du poisson-zèbre pour étudier les mécanismes de son développement embryonnaire et larvaire. A la fin des années 70, la biologie des organes et tissus du poisson-zèbre était déjà bien caractérisée (LAALE 1977).

A partir des années 90, la mise en place progressive d'un arsenal d'outils de manipulation génétique a permis de faire passer le modèle du poisson-zèbre au niveau supérieur. La mise au point de méthodes de clonage, mutagenèse et transgénèse ont mené à la conception de stratégies génétiques directes (partant d'un phénotype d'intérêt pour disséquer ses bases

génétiques) et indirectes (partant de modifications génétiques pour étudier les phénotypes induits). Les résultats du premier essai mutationnel à grande échelle chez le poisson-zèbre furent publiés en 1996, établissant la première grande collection de mutants (~ 1500 mutations affectant le développement embryonnaire, dans 400 gènes, CURRIE 1996 ; DRIEVER et al. 1996 ; HAFFTER et al. 1996). Cet effort fut le fruit du travail des groupes de Nüsslein-Volhard au Max-Planck Institut et du groupe de Wolfgang Driever au Massachusetts General Hospital de Boston. Il culmina avec la publication de 37 papiers dans l'édition spéciale poisson-zèbre de la revue *Development* de décembre 1996 (Figure 1.6). Christiane Nüsslein-Volhard venait alors d'obtenir le prix Nobel de médecine pour sa contribution à l'établissement du modèle invertébré dominant, la *Drosophile*. Wolfgang Driever, quant à lui, était l'ancien étudiant en thèse de Nüsslein-Volhard, largement impliqué dans le succès du modèle *Drosophile*. Après avoir chacun croisé le chemin de Monte Westerfield, lui même inspiré par le pionnier de la génétique du poisson-zèbre George Streisinger, Nüsslein-Volhard et Driever avaient tous les deux été convaincus du potentiel incroyable du poisson-zèbre comme organisme modèle (NÜSSLEIN-VOLHARD 2012).



FIGURE 1.6 – Couverture de l'édition spéciale poisson-zèbre dans *Development* (décembre 1996). Phénotypes mutants de poissons-zèbre présentant différents patrons de coloration des naevoires.

Peu après, en 1997, la première base de données génétiques dédiée au poisson-zèbre a vu le jour, sous le nom de ZFIN (ZebraFish Information Network). Maintenue et étendue durant les années suivantes, ZFIN représente aujourd'hui la base de référence de la génomique des poissons et est largement intégrée aux autres ressources pour la génomique comparative des vertébrés. Enfin, l'établissement de cartes génétiques puis l'assemblage du

génomique du poisson-zèbre ont grandement contribué à mieux caractériser les mutations d'intérêt (POSTLETHWAIT et al. 2000 ; HOWE et al. 2013). En résumé, la mise en place rapide de ressources génétiques, génomiques et fonctionnelles ont permis d'établir le poisson-zèbre comme un organisme modèle incontournable, permettant de caractériser les bases génétiques de nombreuses maladies humaines et menant au développement de nouveaux traitements (TAN et ZON 2011 ; TAMPLIN et al. 2012).

Le poisson-zèbre comme moteur de la génomique comparative des poissons

Les gènes expérimentalement caractérisés chez le poisson-zèbre (répertoriés dans la base ZFIN) servent de connaissance *a priori* pour annoter la fonction de gènes de la majorité des espèces de poissons, sous l'hypothèse de la conservation de fonction des gènes d'origine commune (gènes homologues). Deux classes principales de gènes homologues sont définies. Les orthologues sont des homologues descendant d'une même séquence ancestrale, séparés au moment la spéciation des espèces. Les paralogues sont des homologues séparés par un événement de duplication, ils descendent donc de deux copies différentes d'un gène ancestral (Figure 1.7).

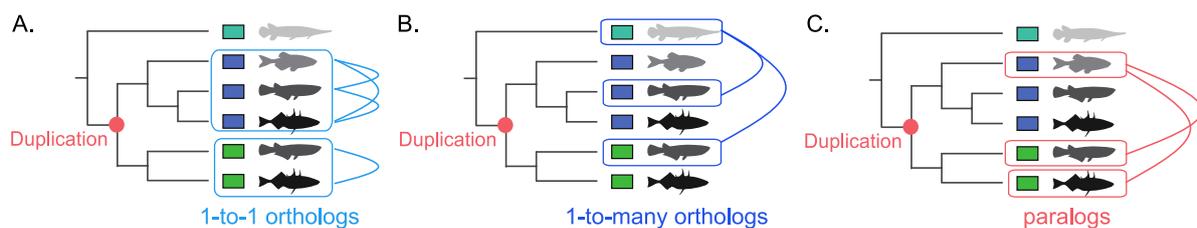


FIGURE 1.7 – Relations d'orthologie et de paralogie. Les gènes paralogues sont séparés par un événement de duplication (nœud rouge), les orthologues par un événement de spéciation (nœuds non colorés). A. Orthologues 1-à-1 : les gènes d'une espèce ont un unique orthologue dans une autre espèce. B. Orthologues 1-à-plusieurs, le gène d'une espèce (ici le spotted gar, en haut) a plusieurs orthologues (ici 2) dans une seconde espèce. C. Paralogues : gènes séparés par un événement de duplication. Les orthologues plusieurs-à-plusieurs ne sont pas représentés, ils concernent le cas où plusieurs gènes de deux espèces sont orthologues entre eux.

La "conjecture orthologue" propose l'idée que la fonction est plus souvent conservée entre gènes orthologues qu'entre paralogues. L'idée sous-jacente est que les paralogues sont moins contraints évolutivement et plus divergents en terme de fonction (voir les modèles de neo- et sub-fonctionnalisation introduits en 1.1.3). En pratique, cette hypothèse a longtemps été remise en question (NEHRT et al. 2011), notamment parce que de nombreux facteurs confondants la rendent difficile à tester. Il semblerait néanmoins que la conservation de fonction soit significativement plus élevée entre orthologues 1-à-1 (aucune duplication) qu'entre les gènes paralogues (ALTENHOFF et al. 2012). Les autres types d'orthologies (1-à-plusieurs, plusieurs-à-plusieurs) présentent une conservation de fonction intermédiaire. En conséquence, la propagation des annotations fonctionnelles aux espèces non-modèles

s'effectue systématiquement par orthologie avec les espèces modèles.

Plus récemment, de nouveaux modèles biomédicaux ont émergé, dans l'ombre du poisson-zèbre. Le medaka détient le rôle de second couteau de la génomique des poissons. Il sert de modèle complémentaire, permettant de confirmer la nature généralisable des résultats observés chez le poisson-zèbre. Le troisième modèle biomédical est le platy, informatif en particulier à l'étude des cancers. Le poisson-zèbre, le medaka et le platy sont à l'heure actuelle les seuls poissons à posséder des centres de ressources génétiques qui leur sont dédiés, assurant le maintien de collections de lignées et les rendant disponibles à la communauté scientifique. Plus à la marge, d'autres modèles poissons sont utilisés, pour étudier des conditions précises (ALBERTSON et al. 2009, voir SCHARTL 2014 pour une revue complète). Par exemple, les Cichlidés servent de modèles pour comprendre les anomalies cranio-faciales, le tétra mexicain sert à comprendre les affections dégénérative de la rétine et les troubles du sommeil, enfin, les kilis sont des modèles dans le contexte du vieillissement.

Le poisson-zèbre a amorcé l'intégration de la génomique des poissons dans le paysage de la génomique comparative des vertébrés. Les enjeux futurs visent à intégrer les nouveaux modèles poissons pour épauler et mieux comprendre le génome du poisson-zèbre. Le second enjeu sera de continuer de séquencer et intégrer les espèces non-modèles aux bases de données d'annotations de gènes de poissons, afin de maximiser le potentiel de chaque espèce.

1.2.2 Phylogénie des poissons téléostéens et projets de séquençage

Les Actinoptérygiens, ou poissons à nageoires rayonnées, comptent environ 30 000 espèces, soit la moitié de la totalité des vertébrés. La vaste majorité des Actinoptérygiens sont des Téléostéens (96%), un groupe apparu il y a ~ 350 millions d'années et présentant une impressionnante richesse phénotypique et écologique. Les Téléostéens contiennent le poisson-zèbre, le medaka, le platy, le poisson-clown, mais aussi, plus étonnement, d'autres espèces morphologiquement très dérivées comme l'hippocampe. Les Téléostéens ont colonisé des habitats variés et parfois même extrêmes. Par exemple, le molly de l'atlantique peut vivre dans des eaux hautement toxiques, là où la plupart des organismes seraient incapables de survivre plus de quelques minutes (KELLEY et al. 2016). Apprécier la richesse du clade téléostéen passe par l'établissement de leur phylogénie, et la reconstruction de la chronologie de leur diversification. L'augmentation du nombre de ressources génomiques et transcriptomiques a permis d'explorer ces questions, même si le positionnement de certains groupes reste débattu.

La phylogénie des poissons téléostéens

Les premières classifications du vivant se sont basées sur des critères morphologiques. Dès les années 1960, les études morphologiques ont posé les bases de la structure de l'arbre des poissons téléostéens (GOSLINE 1965). Les phylogénies établies à partir de données morphologiques partent de l'identification de structures anatomiques remarquablement conservées entre espèces, prédites homologues. Ces caractères homologues permettent d'établir une matrice d'absence/présence de caractères. Il s'agit ensuite d'identifier l'arbre des espèces le plus parcimonieux, c'est à dire celui impliquant le moins de changements d'états de caractères. Le principe le plus souvent utilisé pour placer les transitions d'états sur un arbre d'espèces candidat est la parcimonie de Dollo (FARRIS 1977), postulant qu'un caractère complexe ne peut être acquis qu'une seule fois. Sous ces critères, différents clades des Téléostes ont pu être résolus (NELSON, GRANDE et WILSON 2016).

Depuis les années 1990, les données moléculaires ont commencé à être utilisées pour reconstruire l'arbre des poissons. Trois études majeures successives constituent aujourd'hui la phylogénie consensus (NEAR et al. 2012; BETANCUR-R et al. 2017; HUGHES et al. 2018, Figure 1.8). Ces études sont en large accord entre elles et intègrent respectivement 230, 2 000 et 300 espèces de poissons. Alors que les deux premières phylogénies sont basées sur un jeu de quelques gènes marqueurs (9 et 21, respectivement), la phylogénie de HUGHES et al. 2018 s'appuie sur un jeu de données de 1 105 gènes orthologues 1-à-1. En partant de l'hypothèse que l'histoire évolutive des gènes suit majoritairement l'arbre des espèces, une phylogénie peut être établie à partir de la séquence des gènes. Deux grandes approches sont possibles : inférer un arbre unique à partir de la concaténation des séquences géniques, ou inférer un arbre pour chaque gène et identifier la topologie majoritaire. HUGHES et al. 2018 ont tiré profit de chacune de ces deux approches. De plus, les auteurs ont effectué des tests de vraisemblance (tests AU, SHIMODAIRA 2002; ARCILA et al. 2017) pour évaluer le support statistique associé aux hypothèses alternatives précédemment reportées par les phylogénies morphologiques (Figure 1.8).

Un point de désaccord entre les études moléculaires et morphologiques concerne la position relative des 3 grands groupes ayant divergé proche de la bases des Téléostéens (Clupeocephala, Elopomorphes et Osteoglossiformes). Les trois phylogénies moléculaires infèrent que les Elopomorphes ont divergé les premiers, suivis des Osteoglossiformes puis des Clupeocephala (Figure 1.8). Inversement, le consensus établi par les études morphologiques place les Elopomorphes en groupe frère des Osteoglossiformes (ARRATIA 1998) Cette topologie alternative a également été supportée par une phylogénie moléculaire (BIAN et al. 2016), même si plusieurs critiques peuvent être adressées à cette étude. En effet, des biais ont pu être induits par des méthodologies hétérogènes de collecte des gènes orthologues entre Clupeocephala et autres groupes, l'échantillon d'espèces est faible et induit de longues branches, et l'inférence est uniquement basée sur la concaténation de gènes. Les

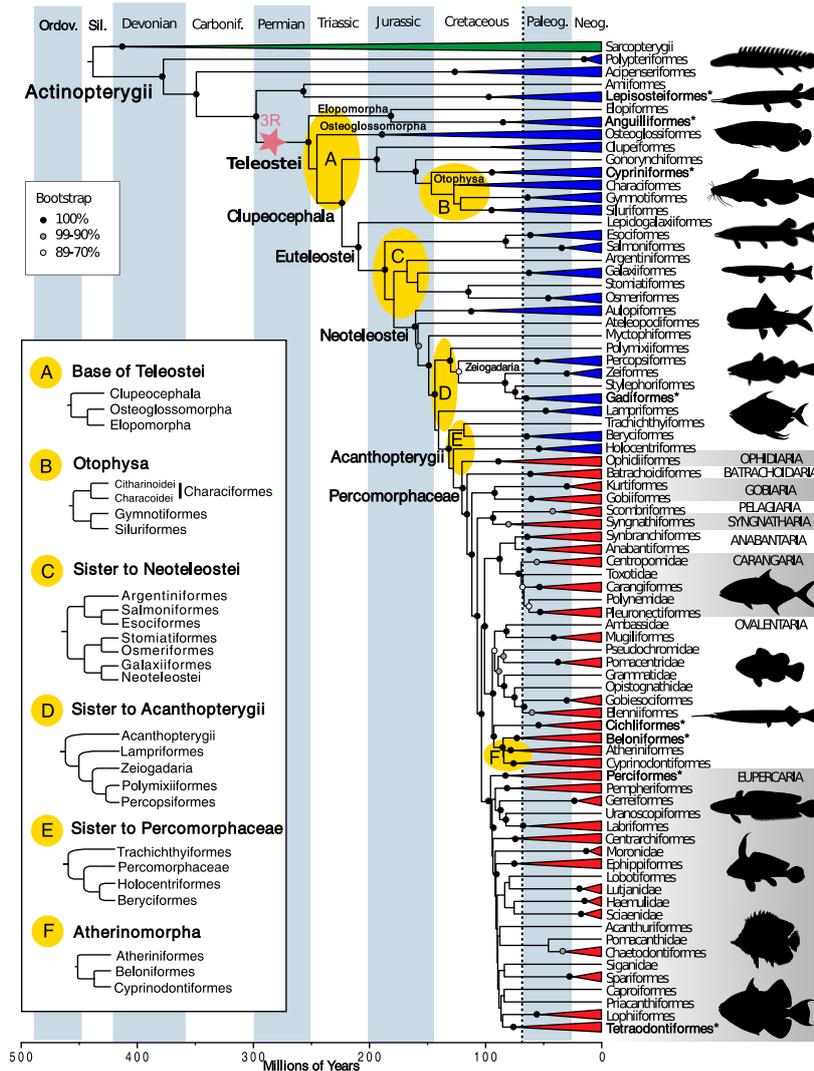


FIGURE 1.8 – Phylogénie des poissons téléostéens, par HUGHES et al. 2018. La phylogénie comprend 3 Sarcoptérygiens (verts, groupes externes) et 300 Actinoptérygiens (bleus et rouges). Les clades de Percomorphes sont indiqués en rouge, pour souligner leur diversification. Les nœuds annotés de A à E correspondent aux hypothèses alternatives présentées dans la littérature et explicitement rejetées par les tests de vraisemblance effectués sur la totalité des arbres de gènes inférés. La position relative de la duplication complète (3R) des poissons téléostéens est indiquée par une étoile rouge (la datation absolue de la duplication n'a pas été évaluée dans cette étude). Figure adaptée de HUGHES et al. 2018.

résultats de Hughes et al. ont montré que la topologie alternative plaçant Elopomorphes et Osteoglossiformes en groupe frère n'est pas statistiquement supportée par les arbres de gènes.

Diversité des poissons et état des lieux des projets de séquençage

L'obtention de l'arbre des poissons permet également de dater la divergence des grands clades, en s'appuyant sur des calibration fossiles. Ces datations établissent la radiation ma-

jeure du groupe des Percomorphes (~ 17 000 espèces) avant la fin du Crétacé : la plupart des espèces de Téléostéens étaient déjà bien établies avant la crise du Crétacé-Paléogène à l'origine de l'extinction des dinosaures. Deux autres épisodes de diversifications majeures sont inférés dans la phylogénie des poissons : un à la base des Téléostes et un second à la base des Otophysii (SANTINI et al. 2009).

Le séquençage et l'assemblage de nouveaux génomes de poissons, en particulier d'Elopomorphes et d'Osteoglossiformes, devraient permettre de préciser l'arbre des espèces, en intégrant à la fois plus de taxons et plus de gènes. Sur les 30 000 espèces de poissons, seules 220 ont été séquencées à ce jour (<http://fish10k.genomics.cn/progress/>). Différents projets de séquençage sont en cours : le Fish 10 000 Genome Project (F10K) vise à séquencer 10 000 génomes de poissons (FAN et al. 2020), avec un objectif à 3 ans de couvrir 3500 espèces représentant les 500 grandes familles. En parallèle, comme évoqué au paragraphe 1.1.1, le projet VGP du G10K Consortium vise, à terme, à séquencer la totalité des espèces de Vertébrés. La génomique comparative chez les poissons va être confrontée, tout comme la génomique des Vertébrés de manière générale, à un déluge imminent de nouveaux génomes. Cependant, de nombreuses difficultés, dues à la complexité des génomes de poissons téléostéens ne sont toujours pas résolues. Les défis relatifs à l'analyse des génomes de poissons sont développés dans la prochaine sous-partie.

1.2.3 Le défi de l'analyse des génomes de poissons

En 1998, avant même la publication du génome du poisson-zèbre, trois études ont indépendamment constaté que les familles de gènes chez le poisson-zèbre étaient plus grandes que chez les mammifères (POSTLETHWAIT et al. 1998 ; PRINCE et al. 1998 ; WITTBRODT, MEYER et SCHARTL 1998). Par la suite, il a été mis en évidence qu'une large fraction des gènes du poisson-zèbre, et des poissons téléostéens en général, existaient en plusieurs copies (MEYER et SCHARTL 1999 ; NARUSE et al. 2000 ; LOH et al. 2004). Depuis, il a été démontré que ces gènes dupliqués proviennent d'un événement de duplication complète de génome, survenu chez l'ancêtre des poissons téléostéens (TAYLOR et al. 2003 ; JAILLON et al. 2004).

Cet événement de duplication complète a laissé une empreinte substantielle sur les génomes de poissons modernes. Les duplications complètes de génomes induisent une copie de tous les chromosomes, et donc de tous les gènes d'une espèce. Il en résulte une complexification des réseaux de gènes, à travers notamment l'expansion de familles de gènes du développement et facteurs de transcriptions. L'ancien événement de duplication complète commun aux poissons téléostéens est suggérée avoir contribué à leur impressionnante diversité morphologique, physiologique et écologique. Du fait de leur contenu élevé en gènes dupliqués (entre 15 et 26%, BRUNET et al. 2006 ; HOWE et al. 2013), comparer les gé-

nomes de poissons représente un défi méthodologique. Cette sous-partie vise à introduire les difficultés que la duplication complète pose pour les études de génomique comparative et fonctionnelle chez les poissons. La mise en évidence et l'impact évolutif des duplications complètes sera discuté plus largement dans la partie 1.3.

Difficultés pour le transfert d'annotations fonctionnelles

Preuve du retard des études de génomique comparative chez les poissons, la résolution de leur phylogénie est récente et contraste avec les connaissances que l'on a quant à la radiation des Tétrapodes (THOMSON et SHAFFER 2010). Les difficultés que posent les événements de duplications dans la reconstruction des arbres de gènes ont déjà été évoquées (voir 1.1.3) : dans le cas d'une duplication complète, le problème est d'autant plus massif. La première difficulté entraînée est que les relations de paralogie et d'orthologie entre espèces sont difficiles à établir avec confiance. En conséquence, la propagation des annotations fonctionnelles, basée sur les relations d'orthologie, est également imprécise.

Un second problème est qu'il est difficile de quantifier l'importance de la divergence indépendante des paralogues (Figure 1.9). Par exemple, il peut arriver qu'une espèce retienne les deux copies par sub-fonctionnalisation tandis qu'une seconde espèce les retient par néofonctionnalisation. Dans ce cas, le transfert des annotations fonctionnelles entre les deux espèces ne sera pas pertinent. Résoudre la question des résolutions indépendantes de paralogues implique d'être capable de résoudre correctement les relations d'orthologie entre espèce. Cela requiert également de caractériser la fonction des gènes dans différentes espèces.

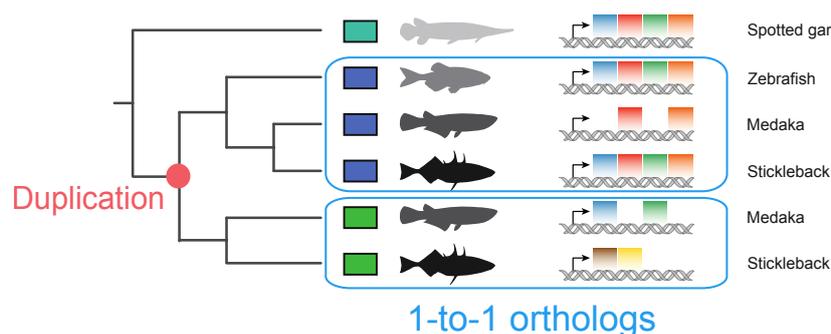


FIGURE 1.9 – Divergence de gènes dupliqués orthologues. Exemple d'une paire de gènes dupliqués au destin différent dans 3 espèces : non-fonctionnalisation chez le poisson-zèbre, sub-fonctionnalisation chez le medaka et néo-fonctionnalisation chez l'épinoche.

Difficultés pour la reconstruction de l'histoire évolutive des chromosomes

Par ailleurs, de part leur origine par duplication complète, les gènes dupliqués forment de larges régions dupliquées dans les génomes de poissons (Figure 1.10). Ces régions cor-

respondent aux chromosomes dupliqués chez l'ancêtre Teleostei, depuis érodés par les processus d'évolution des génomes. Il n'existe pas de cartographie précise des régions dupliquées entre les différents génomes de poissons. Pourtant, confronter les caractéristiques génomiques de ces régions entre espèces aiderait à caractériser la fonction des génomes de poissons.

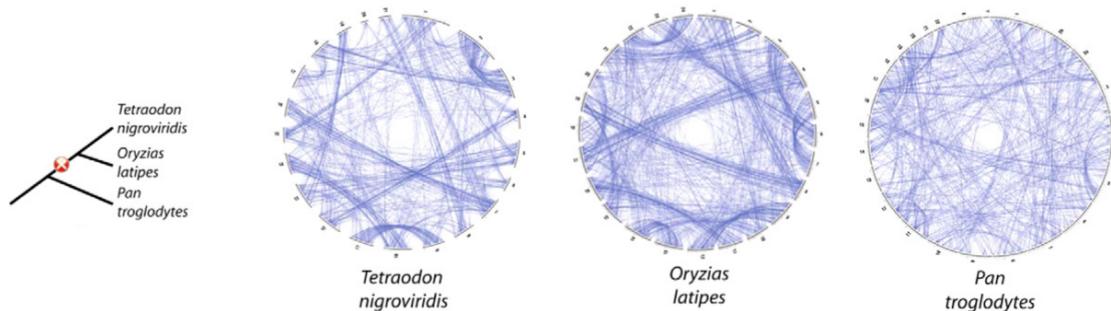


FIGURE 1.10 – Régions dupliquées dans les génomes de poissons téléostéens. Visualisation des gènes dupliqués par la duplication complète chez le tétraodon (*Tetraodon nigroviridis*) et le medaka (*Oryzias latipes*). Les gènes paralogues sont liés par des traits bleus, faisant apparaître un patron complexe de régions dupliquées. Figure tirée de JAILLON, AURY et WINCKER 2009

Identifier précisément les régions dupliquées dans différents génomes de poissons peut également faciliter l'identification des gènes orthologues. Les prédictions d'orthologie basées sur les patrons d'organisation des génomes utilisent le principe de la synténie, ou plus précisément de la synténie conservée. Le terme "synténie" signifie initialement, "sur le même brin" ou "sur le même chromosome". La conservation de la synténie désigne l'existence, dans deux génomes différents, de blocs de gènes conservés dans la même orientation et dans le même ordre. La conservation de la synténie permet de faire l'hypothèse que l'arrangement des gènes était déjà établi dans l'ancêtre commun aux deux espèces comparés et que les gènes des blocs conservés sont orthologues (Figure 1.11). Le principe de la conservation de synténie est communément utilisé pour confirmer l'orthologie de gènes dupliqués entre génomes de poissons. Néanmoins, l'identification de blocs de synténie conservée passe par un squelette de gènes orthologues "sûrs" entre les deux génomes. En l'absence de prédiction d'orthologie de qualité, ce squelette initial peut être difficile à établir. Pour cette raison, les méthodes classiques de reconstruction ancestrales de génomes fonctionnent mal en présence de duplication complète de génome (NAKATANI et MCLYSAGHT 2017).

Absence de ressource dédiée à la génomique comparative des poissons

Comme introduit plus haut (1.2.1), ZFIN est la base de données répertoriant les annotations fonctionnelles du poisson-zèbre. C'est également la base de données de référence pour les annotations de gènes de poissons en général. Du fait des erreurs dans les relations

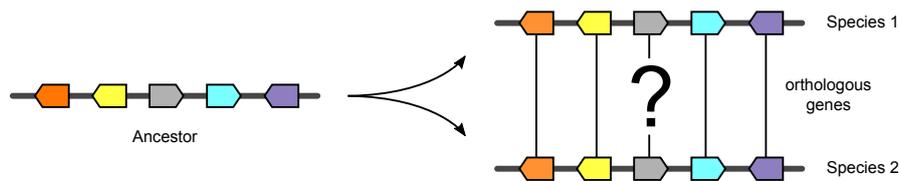


FIGURE 1.11 – Synténie conservée et orthologie. Bloc de gènes retrouvée dans le même ordre et la même orientation dans deux génomes différents. Les gènes colorés ont été prédits orthologues. L'hypothèse de conservation de synténie depuis l'ancêtre commun des deux espèce permet de prédire que les gènes du centre (gris) sont également orthologues.

d'orthologie prédites, la nomenclature des gènes des poissons non-modèles est actuellement instable. La base de données d'Ensembl utilise les noms de gènes définis par ZFIN pour le poisson-zèbre et transfère ces noms par orthologie aux autres espèces. Cependant, à cause des erreurs d'orthologie, les noms de gènes de poissons peuvent être instables d'une version de la base de données à l'autre (c'est le cas notamment des gènes *Sox9a/b* ou encore *cxcl12a/b*). Comme déjà suggéré par Spaink et ses collègues en 2014, le succès des études comparatives futures repose sur l'établissement de ressources communes à toutes les espèces de poissons séquencées, suivant le modèle de ZFIN, et liant les gènes orthologues de façon rigoureuse (SPAINK, JANSEN et DIRKS 2014).

La fraction importante de gènes dupliqués dans les génomes de poissons téléostéens complique significativement leur analyse. Elle rend difficile : l'identification des gènes orthologues, la reconstruction de l'histoire évolutive des chromosomes et le transfert d'annotations fonctionnelles. Le problème à la base de ces difficultés est la complexité d'inférer des arbres de gènes en présence d'événements de duplication. Pourtant, les poissons téléostéens présentent une diversité impressionnante : à la fois au niveau morphologique, écologique, mais aussi génomique. Les génomes de poissons sont très variables, en terme de taille (VENKATESH 2003), de composition en éléments répétés (CHALOPIN et al. 2015 ; SHAO, HAN et PENG 2019), de contenu en GC (SYMONOVÁ et SUH 2019), ou encore de nombre de gènes. Pour les raisons exposées plus haut, cette richesse est actuellement sous-exploitée et nécessite le développement de ressources et de méthodologies spécifiques. L'utilisation de la synténie conservée pour préciser l'histoire évolutive des gènes est un bon point de départ pour répondre à ce problème. Une partie de mon travail de thèse a consisté à intégrer l'information de synténie conservée aux méthodes de reconstruction d'arbre, afin d'améliorer les prédictions d'orthologie.

1.3 Les duplications complètes et leur impact sur les génomes de poissons

1.3.1 Événements de duplications complètes chez les Téléostéens

Duplications complètes chez les Eucaryotes

Les duplications complètes de génomes sont des événements mutationnels dramatiques, générant une copie de la totalité du génome d'une espèce. Produisant de ce fait une copie pour tous les gènes, elles sont une source d'innovations évolutives et influencent la robustesse des génomes. Elles sont récurrentes dans l'histoire évolutive des plantes : la majorité des espèces d'Angiospermes (plantes à fleur) ont connu entre 2 et 3 duplications complètes (VAN DE PEER, MIZRACHI et MARCHAL 2017). En conséquence, la plupart des connaissances proviennent d'observations faites sur les génomes de plantes. L'événement de duplication complète survenu dans l'ancêtre commun des levures du genre *Saccharomyces* a également été intensivement étudié (datée à 80 millions d'années, KELLIS, BIRREN et LANDER 2004). De plus, les paramécies représentent un second modèle unicellulaire d'intérêt, avec plusieurs duplications complètes mises en évidence (MCGRATH et al. 2014).

Les duplications complètes sont étroitement liées à l'évolution des Vertébrés. Deux épisodes successifs de duplications (1R-2R) se sont produits dans l'ancêtre commun à toutes les espèces de vertébrés actuelles (il y a environ 500 millions d'années, DEHAL et BOORE 2005). Des duplications subséquentes ont également eu lieu : plusieurs chez les Xénopes, dont une bien caractérisée chez l'espèce modèle *Xenopus laevis* (datée à 17-18 millions d'années, SESSION et al. 2016), plusieurs chez les poissons non-téléostéens (esturgeons, HAVELKA et al. 2013) et téléostéens (JAILLON et al. 2004 ; BRAASCH et POSTLETHWAIT 2012, Figure 1.12). Les duplications complètes chez les Vertébrés ne sont pas encore bien caractérisées : il n'est pas clair si les observations faites chez les plantes et autres eucaryotes unicellulaires sont généralisables. Cette sous-partie vise à introduire les différents événements de duplications complètes chez les poissons téléostéens et leur mise en évidence. Les sous-parties suivantes développeront leurs implications génomiques (1.3.2) et évolutives (1.3.3).

Mise en évidence de duplications complètes chez les poissons

Les premières observations d'un nombre de copies de gènes anormalement élevé chez le poisson-zèbre (POSTLETHWAIT et al. 1998 ; WITTBRODT, MEYER et SCHARTL 1998 ; POSTLETHWAIT et al. 2000) n'ont initialement pas permis de conclure quant au mécanisme pouvant en être à l'origine (ROBINSON-RECHAVI et al. 2001). Pourtant, des traces d'une possible duplication à grande échelle commençaient à apparaître. Les gènes *hox*, gènes majeurs du développement et responsable de l'établissement du plan d'organisation, sont parmi les gènes les plus étudiés en biologie. Les gènes *hox* fonctionnent en cluster : un cluster unique existe

chez les invertébrés, tandis que chez les Vertébrés, suite aux 1R-2R, on retrouve 4 clusters, chacun sur des chromosomes différents. Sept clusters *hox* ont été mis en évidence chez le poisson-zèbre, suggérant leur possible re-duplication, avec la perte d'une copie (AMORES et al. 1998).

La preuve définitive d'une duplication complète de génome dans l'ancêtre des poissons téléostéens est arrivée avec l'analyse du génome du tétraodon (JAILLON et al. 2004). Les auteurs ont montré que les gènes dupliqués chez le tétraodon étaient significativement associés à des paires de chromosomes : les chromosomes ancestralement dupliqués par un événement de duplication complète daté à 320 millions d'années (JAILLON et al. 2004 ; VANDEPOELE et al. 2004 ; GLASAUER et NEUHAUSS 2014).

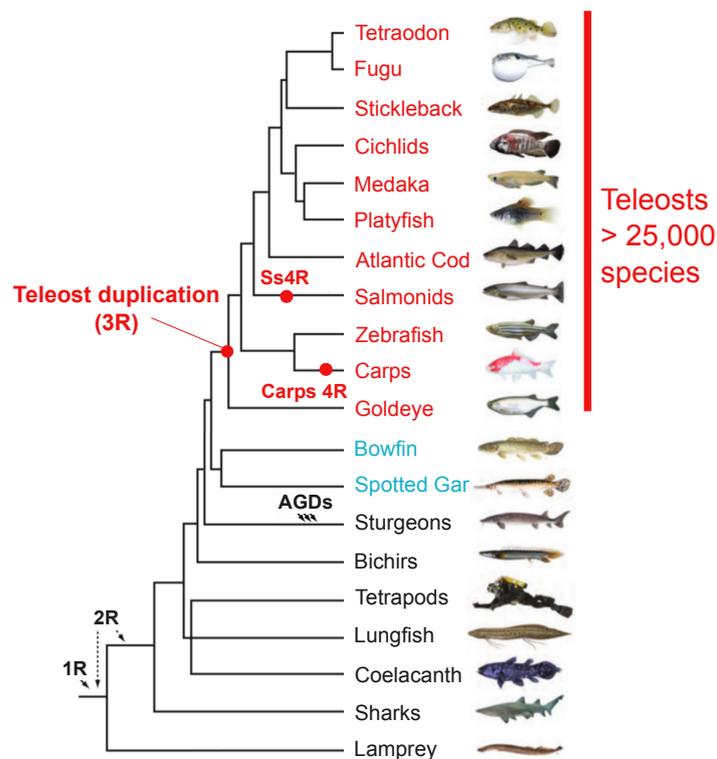


FIGURE 1.12 – Évènements de duplications complètes chez les poissons téléostéens. Phylogénie simplifiée des Vertébrés, montrant les téléostéens en rouge et les Holostéens (plus proche groupe externe non dupliqué) en bleu. Les 3 événements de duplications majeurs chez les poissons téléostéens sont mis en évidence : la 3R des Téléostéens, la Ss4R des Salmonidés et la 4R des Carpes. Les deux duplications à la base des vertébrés sont également représentées (1R, 2R), en plus des duplications des esturgeons (AGDs). Figure adaptée de BRAASCH et POSTLETHWAIT 2012, avec autorisation.

D'autres duplications complètes subséquentes ont été mis en évidence chez les poissons téléostéens : la duplication des Salmonidés (Ss4R estimée à 50-80 millions d'années, LIEN et al. 2016) et la duplication de l'ancêtre commun de la carpe et du poisson rouge (Carpes 4R, estimée à 12.4 millions d'années, XU et al. 2019). En plus de la duplication des carpes, d'autres duplications complètes se sont produites dans d'autres familles de Cypriniformes :

chez les poissons à ventouses (Catostomidae, UYENO et SMITH 1972) et les loches (Cobitidae, SAITOH, CHEN et MAYDEN 2010).

Patron biaisé des duplications complètes dans l'arbre de la vie

Pourquoi les duplications complètes sont-elles si fréquentes chez les plantes et rares chez les animaux ? Posée pour la première fois par Muller en 1925, cette question n'a toujours pas de réponse définitive (MULLER 1925). Les connaissances accumulées depuis suggèrent que la solution passe par la reformulation de la question. Pourquoi les duplications complètes sont-elles prévalentes dans certains groupes d'animaux comme les Amphibiens et les poissons et complètement absentes chez les Mammifères et les Oiseaux ? Le Rat-viscache roux d'Argentine *Tympanoctomys barrerae*, qu'on a longtemps pensé être la seule espèce de mammifère ayant subi un événement de duplication complète, devait permettre de mieux comprendre les mécanismes de tolérance aux duplications de génome (GALLARDO et al. 2004). En réalité, la grande taille du génome du Rat-viscache semble plutôt due à une expansion d'éléments transposables qu'à une duplication complète (EVANS et al. 2017) : il n'y a toujours pas d'exemple de polyploïdie chez les Mammifères. Une hypothèse mise en avant considère l'impact des duplications complètes sur les réseaux de régulation de gènes, perturbés par ce « choc génomique » (détaillé en 1.3.3). En particulier, la perturbation des réseaux ne serait pas tolérée par les contraintes strictes du développement embryonnaire des mammifères (voir WERTHEIM, BEUKEBOOM et ZANDE 2013 pour une revue complète).

Répondre à ces questions nécessite de mieux comprendre les duplications chez les Vertébrés : leurs conséquences génomiques et évolutives à court et long-terme. Mon travail vise à mieux caractériser la duplication commune aux espèces de poissons téléostéens (3R). S'agissant de la duplication la plus ancienne chez les téléostéens, la 3R est encore mal caractérisée parce que difficile à étudier, comparée à la Ss4R des Salmonidés et la 4R des carpes, où les traces des duplications sont encore bien identifiables. Je m'inspire des connaissances grandissantes sur ces deux événements de duplications « récents » pour étudier la 3R.

1.3.2 Conséquences sur l'organisation des gènes et des génomes

Cette sous-partie introduit les conséquences génomiques des duplications complètes. De manière générale, les génomes polyploïdes ne sont pas stables sur le long-terme. Le plus souvent, les polyploïdes reviennent à un état diploïde, en terme de comportement méiotique et, globalement, en terme de nombre de copies de gènes. La question que j'examine ici est la suivante : comment les anciens événements de duplications complètes ont-ils façonné les génomes modernes qui en descendent ?

Évolution des gènes après duplication complète

J'ai déjà introduit les grands modèles qui décrivent l'évolution de gènes après duplication au paragraphe 1.1.3. Brièvement, le destin le plus courant d'un gène dupliqué est d'être non-fonctionnalisé et perdu. Les duplications complètes de génome sont suivies d'une période de perte massive de gènes dupliqués INOUE et al. 2015. En conséquence, dans les génomes des téléostéens actuels, seulement une fraction des gènes sont maintenus en deux copies (26% chez le poisson-zèbre après la 3R, HOWE et al. 2013). Les gènes dupliqués, ou ohnologues, peuvent échapper à la redondance fonctionnelle et perdurer en 2 copies, par sub- ou néofonctionnalisation.

De plus, dans le cas des duplications complètes, un autre modèle est impliqué dans la rétention des paralogues : l'équilibre de dosage. L'hypothèse de l'équilibre de dosage prédit que les gènes codant les sous-unités d'un complexe multiprotéique doivent être maintenues en deux copies, afin d'assurer la bonne stoechiométrie entre les différents partenaires (MAKINO et MCLYSAGHT 2010). Cependant, ce modèle semble plutôt expliquer la préservation initiale des gènes ohnologues, puisque l'équilibre de dosage peut être restauré sur le long terme par l'ajustement du niveau d'expression des protéines en interaction (VAN DE PEER, MIZRACHI et MARCHAL 2017). De fait, le DBH permettra le maintien initial des ohnologues, laissant le temps à l'accumulation de mutations pouvant mener à leur sub- ou néo-fonctionnalisation et leur rétention. Enfin, un dernier modèle propose que certains gènes, dits "dangereux", sont maintenus dupliqués en raison du coût pour la fitness qu'induisent les mutations nécessaires à les emmener sur la voie de la pseudogénéisation (GIBSON et SPRING 1998 ; SINGH et al. 2012 ; MALAGUTI, SINGH et ISAMBERT 2014 ; ROUX, LIU et ROBINSON-RECHAVI 2017). Ces gènes dangereux sont définis par leur susceptibilité aux mutations délétères dominantes, par gain de fonction : leur sur-activation est responsable de désordres cellulaires importants. En particulier, de nombreuses familles d'oncogènes ont été étendues dans le génome humain à la suite des duplications complètes des Vertébrés (1R-2R), comme, par exemple, la famille RAS impliquée dans environ 25% des cancers humains (SINGH et al. 2012).

On ne connaît pas réellement la contribution exacte de chacun des modèles cités précédemment dans la rétention des gènes dupliqués. Cependant, il est clair qu'ils génèrent des biais dans le répertoire de gènes des espèces ancestralement polyploïdes. Notamment, certaines catégories fonctionnelles sont typiquement enrichies dans les paires de paralogues retenues, comme les facteurs de transcription, les récepteurs des voies de signalisation ou encore les gènes impliqués dans le développement (INOUE et al. 2015 ; SINGH, ARORA et ISAMBERT 2015 ; LI et al. 2016).

Conséquences sur l'organisation des génomes

Le retour de la plupart des gènes à un état simple copie et des génomes à un comportement méiotique disomique est nommé la rediploïdisation. Cette rediploïdisation peut s'accompagner d'un taux élevé de réarrangements génomiques (WANG et al. 2005 ; SÉMON et WOLFE 2007a ; DU et al. 2020) et de pertes massives de gènes (INOUE et al. 2015 ; SESSION et al. 2016 ; CHEN et al. 2019). De fait, les chromosomes ancestralement dupliqués, ou homéologues, ne sont plus directement détectables dans les génomes modernes. Cependant, une signature caractéristique est retrouvée dans les génomes d'espèces anciennement tétraploïdes : la synténie doublement conservée (« Double-Conserved Synteny », DCS). Lorsqu'un génome non-dupliqué est comparé à une espèce dupliquée, il apparaît que les blocs de synténie sont conservés en « 1-à-2 ». Autrement dit, un bloc de gènes de l'espèce non-dupliquée correspond à deux blocs dans l'espèce dupliquée. Ce patron de synténie doublement conservée est marqué par des motifs d'entrelacement des orthologues : les orthologues, dans l'espèce dupliquée, se trouvent tantôt sur le premier des deux blocs dupliqués, tantôt sur le second et tantôt sur les deux si les deux paralogues ont été retenus (Figure 1.13). Ces patrons de synténie doublement conservée permettent de confirmer l'origine polyploïde d'une espèce et facilitent également l'identification des gènes dupliqués issus de la duplication complète.

Les mécanismes de rediploïdisation et des retentions de gènes sont encore mal compris. Cette sous-partie visait à introduire les conséquences à long-terme des duplications complètes sur les génomes. Les particularités liées aux différents mécanismes d'origine de duplications complètes de génomes sont détaillées dans la prochaine sous-partie.

1.3.3 Origine et succès évolutifs des polyploïdes

Les duplications complètes de génome sont des événements dramatiques, à l'origine d'une grande instabilité génomique. Longtemps considérées comme une impasse évolutive de part les défis génomiques et évolutifs qu'elles représentent, elles ont pourtant largement impacté l'évolution des vertébrés. Comment les polyploïdes se forment-ils ? Comment parviennent-ils à survivre, puis à établir et maintenir une population ? Dans cette sous-partie, j'introduis les mécanismes à l'origine de la formation de polyploïdes et les hypothèses expliquant comment les polyploïdes peuvent parvenir à surmonter les conflits génomiques auxquels ils sont confrontés, pour finalement s'établir sur le long-terme.

Formation et survie des polyploïdes

Les duplications complètes de génome surviennent majoritairement suite à la fusion de gamètes non-réduits (RAMSEY et SCHEMSKE 1998), transformant effectivement, dans le cas d'une espèce diploïde, le nombre de chromosomes de $2n$ à $4n$. On distingue deux grands

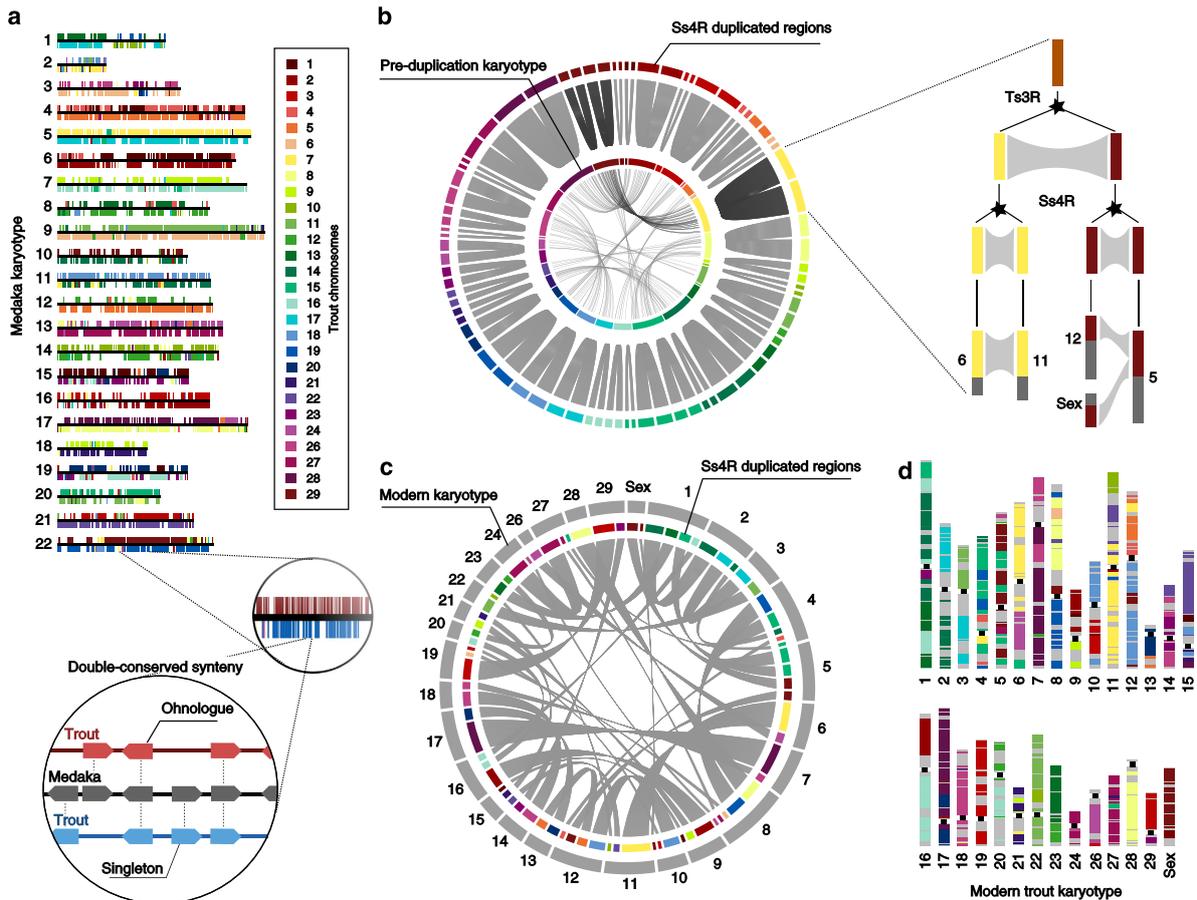


FIGURE 1.13 – Organisation des génomes d'espèces dupliquées : l'exemple du génome de la truite arc-en-ciel. a. Patron de syntenie doublement conservée (DCS) entre le medaka et la truite arc-en-ciel. La truite arc-en-ciel a subi un événement de duplication (la Ss4R) en plus de la duplication 3R qu'elle partage avec le medaka. En conséquence, un chromosome du medaka correspond globalement à deux chromosomes de la truite-arc-en-ciel. b. Paires de larges régions dupliquées dans le génome de la truite arc-en-ciel (cercle extérieur), elles-mêmes groupées en paires de régions précédemment dupliquées par la 3R (cercle interne). c. Représentation circulaire du génome de la truite arc-en-ciel, montrant la position des régions dupliquées présentées en (b). d. Représentation alternative de (c) sous forme de caryotype. Figure tirée de BERTHELOT et al. 2014.

types de duplications complètes : l'allopolyploïdie, où les deux génomes parentaux proviennent d'espèces différentes, et l'autopolyploïdie où les deux génomes parentaux appartiennent à la même espèce (Figure 1.14, STEBBINS 1947). Dans le cas d'une allopolyploïdie, la duplication se produit généralement après l'étape d'hybridation. En effet, les méioses des hybrides sont plus instables et ont une plus grande probabilité de former des gamètes non-réduits. La formation d'hybrides polyploïdes peut également se produire par doublement somatique, durant les premières étapes du développement. Dans la pratique, l'origine des polyploïdes est rarement connue. Les définitions théoriques de l'autopolyploïdie et l'allopolyploïdie représentent plutôt les deux extrêmes d'un spectre continu (MASON et WENDEL 2020). Dans les autotétraploïdes, les 2 jeux de chromosomes identiques forment initialement des tétravalents en méiose. Dans les allotétraploïdes, il est attendu que les génomes

soient suffisamment différents pour former des bivalents. Pourtant, le terme d'allopolyploïde segmentaire est couramment employé dans la littérature pour désigner un allopolyploïde dont les génomes parentaux sont suffisamment proches pour que tous ou partie des chromosomes des parents s'apparient entre eux en méiose (GAUT et DOEBLEY 1997; MARTIN et HOLLAND 2014).

Les néopolyploïdes, ou "jeunes polyploïdes", sont confrontés à de nombreux défis, le premier d'entre eux étant de survivre avec un génome instable. Suite à l'événement de duplication complète, les processus cellulaires sont significativement perturbés : le contenu génétique a doublé de taille, forçant l'augmentation du volume des cellules et des noyaux. Cela a un impact significatif sur l'organisation des chromosomes. En effet, il a été montré, chez les plantes, que les interactions inter- et intra-chromosomes, ou TADs (Topologically Associated Domain), sont désorganisées après duplication complète (WANG et al. 2018; ZHANG et al. 2019). Ce désordre cellulaire dérègle les mécanismes de régulation et d'expression des gènes. On ne sait pas précisément comment les réseaux de régulation sont affectés, ni quelles sont les propriétés qui les rendent robustes aux duplications complètes. Il semblerait que la résolution des conflits d'expression passe par un mécanisme de compensation de dosage, réduisant, en moyenne, l'expression des gènes à un niveau comparable à l'expression diploïde (PALA et al. 2010; SONG et al. 2020)

A partir de ces accidents méiotiques rares et après avoir surmonté le défi de survie initial après la duplication, les polyploïdes doivent réussir à se reproduire et établir une population. La duplication complète érige une barrière reproductive entre les populations diploïde et tétraploïde : les croisements entre eux produisent des individus triploïdes, généralement stériles. Pourtant, ces triploïdes peuvent permettre à la population tétraploïde de s'établir via le principe du « pont triploïde » (RAMSEY et SCHEMSKE 1998; HUSBAND 2004). En effet, les triploïdes produisent une fraction de gamètes diploïdes ou triploïdes viables, qui par fusion avec soit un gamète $2n$ d'un triploïde ou tétraploïde, soit un gamète $1n$ d'un diploïde, donne naissance à un nouvel individu tétraploïde. Ainsi, le pont triploïde offre un second mécanisme permettant la formation d'individus tétraploïdes qui contribuent à l'établissement d'une population tétraploïde.

Enfin, les autopolyploïdes et les allopolyploïdes sont également confrontés à des difficultés spécifiques. Pour les autopolyploïdes, le conflit majeur à résoudre est le retour à une méiose diploïde, puisque la formation de tétravalents rendent les méioses instables et induisent une baisse de fertilité. Les mécanismes de rediploïdisation des méioses sont mal caractérisés. Ils semblent s'opérer par l'accumulation de réarrangements génomiques, voire des pertes de segments de chromosomes, créant ainsi des paires de chromosomes plus similaires entre eux parmi le pool des 4 chromosomes homologues (WANG et al. 2005; DU et al. 2020). Les allopolyploïdes sont en moyenne plus affectés par les perturbations cellulaires et

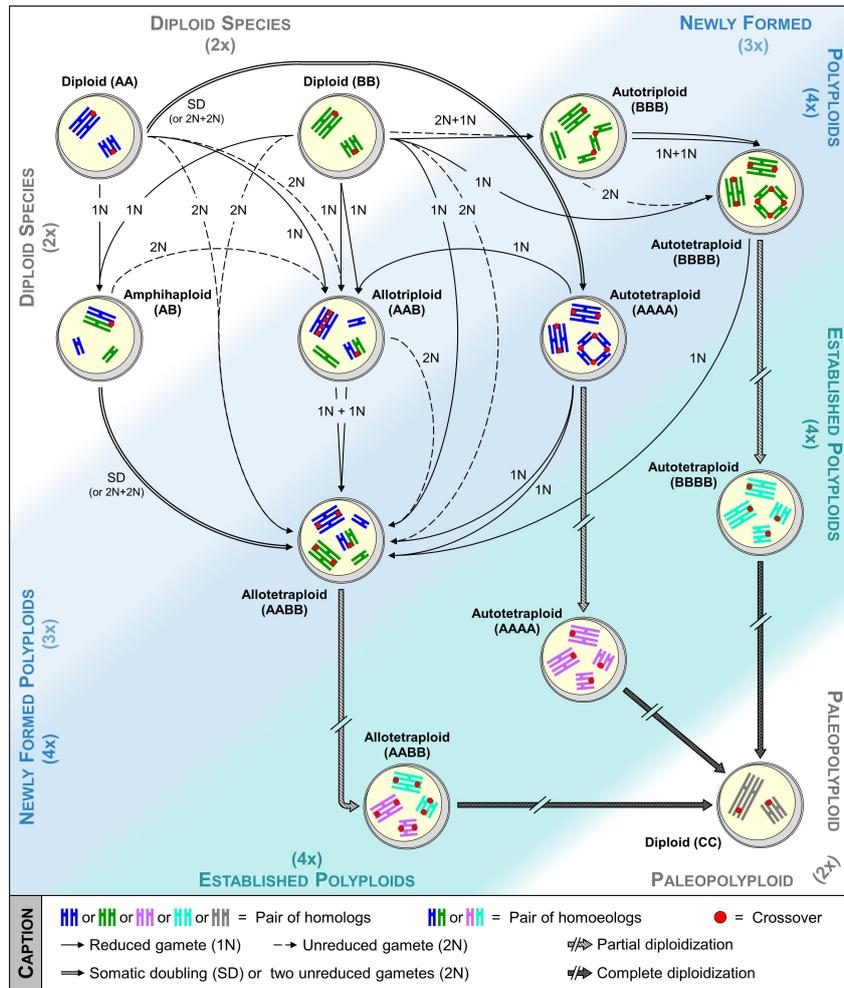


FIGURE 1.14 – Formation, établissement et rediploïdisation des polyplloïdes. Illustration des différentes routes menant à l’allo- et l’autotétraploïdisation, incluant le « pont triploïde ». Les chromosomes d’une même espèce sont représentés de la même couleur. Après des millions d’années d’évolution, les tétraploïdes retournent à un état diploïde. Figure tirée de PELÉ, ROUSSEAU-GUEUTIN et CHÈVRE 2018.

génomiques précédemment citées, en lien avec la cohabitation de différents sous-génomes parentaux. Par exemple, la combinaison de deux populations d’éléments transposables différentes peut altérer leur activité et modifications épigénétiques, impactant la régulation des gènes voisins. Il a été observé à plusieurs reprises que la résolution des conflits entre sous-génomes parentaux passe par la dominance d’un sous-génome sur l’autre (hypothèse de dominance de sous-génome), d’abord à travers l’extinction de l’expression du sous-génome dominé, puis à travers la pseudogénéisation de ses gènes. Chez les plantes, il est suggéré que la composition en éléments transposables (TE) des sous-génomes pourrait moduler cette dominance d’expression, avec le sous-génome le plus « TE-épars » le plus exprimé (FREELING et al. 2012; EDGER et al. 2017). Cependant, il existe des exemples d’allopolyploïdes sans biais d’expression ni de rétention entre sous-génomes (SUN et al. 2017).

Duplications complètes et diversification

Nées de l'observation que beaucoup de clades marqués par une duplication complète ont connu un succès évolutif majeur (Vertébrés, Téléostéens, Angiospermes), même si ce n'est pas systématiquement le cas (KENNY et al. 2016), différentes hypothèses tentent de lier duplication complète et diversification des espèces. L'importance des duplications de gènes dans la génération de nouveautés évolutives est bien établie. En dupliquant les gènes en masse, le potentiel pour générer des nouveautés morphologiques et coloniser de nouveaux milieux semble grand. Cependant, chez les téléostéens, la duplication complète ne peut pas complètement expliquer leur radiation. En effet, le plus important épisode de diversification se situe à la base des Percomorphes (datée à 110 millions d'années, soit environ 200 millions d'années après la duplication). L'augmentation du taux de diversification des Téléostéens après la duplication est également débattu (SANTINI et al. 2009 ; LAURENT, SALAMIN et ROBINSON-RECHAVI 2017). La duplication complète semble cependant précéder une période de forte extinction chez les poissons non-téléostéens, lié à la crise du Permien-Trias (LAURENT, SALAMIN et ROBINSON-RECHAVI 2017 ; VAN DE PEER, MIZRACHI et MARCHAL 2017).

Il semble que les duplications n'induisent pas une radiation évolutive soudaine. Elles pourraient, en revanche, permettre aux espèces de se constituer une « réserve adaptative » (NIETO FELINER, CASACUBERTA et WENDEL 2020). Cette réserve pourrait s'exprimer selon les conditions environnementales, expliquant la survie à long-terme des anciens polyploïdes. Le découplage duplication/diversification peut être concilié avec le modèle majoritaire de diversification après duplication, le "radiation time-lag model". Ce modèle dicte qu'un temps de latence est nécessaire entre la duplication et la radiation, pouvant durer jusqu'à des dizaines voire des centaines de millions d'années. Van de Peer et al. remarquent qu'un grand nombre d'anciens événements de duplications complètes corrèlent avec des crises environnementales majeures (VAN DE PEER, MIZRACHI et MARCHAL 2017). Cette observation rejoint l'hypothèse adaptative selon laquelle le plus grand nombre de copies de gènes pourraient permettre aux polyploïdes de survivre dans des conditions difficiles, là où les diploïdes s'éteindraient. Pour tester cette hypothèse, YAO, CARRETERO-PAULET et PEER 2019 ont réalisé des expériences d'évolution *in silico*. Les auteurs ont montré que les organismes numériques étaient majoritairement diploïdes dans des environnements stables, tandis que le nombre de polyploïdes augmentait drastiquement en cas de changements environnementaux majeurs (YAO, CARRETERO-PAULET et PEER 2019).

Distinguer autotétraploïdie et allotétraploïdie

Chez les poissons téléostéens, la duplication des carpes est une allotétraploïdisation, tandis que la duplication des Salmonidés est une autotétraploïdisation. Pour les carpes, la preuve provient de l'identification d'un des génomes parentaux, rendu possible par la date

relativement récente de la duplication (12.4 millions d'années, XU et al. 2019). L'extinction de l'expression d'un sous-génome parental a également été observée, en accord avec l'hypothèse de dominance de sous-génome (XU et al. 2019 ; LUO et al. 2020). Chez les saumons, une partie du génome continue à être héritée de manière tétrasomique, une indication que la diploïdisation de la méiose de l'ancêtre autotétraploïde n'est pas complètement résolue (LEE et WRIGHT 1981 ; LIEN et al. 2016). De plus, aucun biais de rétention d'ohnologues ne semble présent (BERTHELOT et al. 2014).

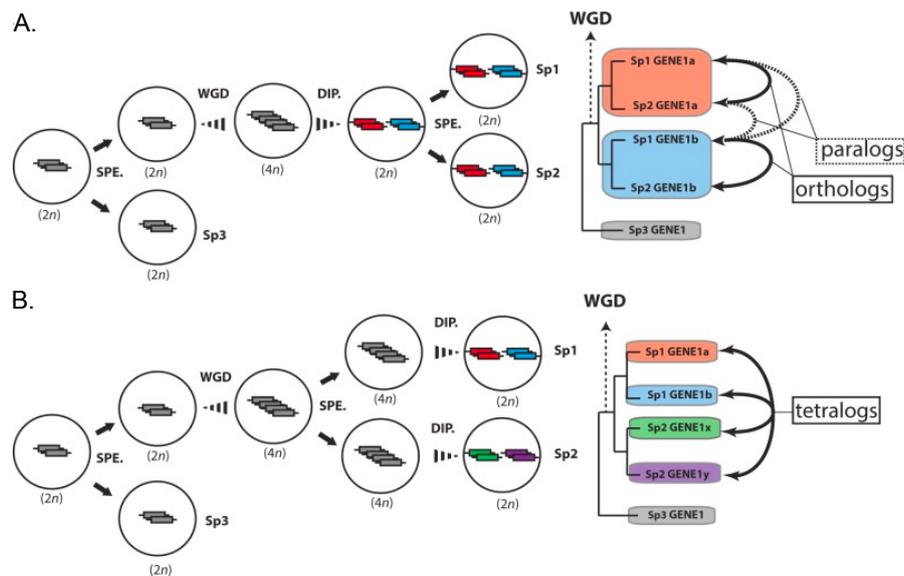


FIGURE 1.15 – Gènes dupliqués hérités de manière tétrasomique. A. Modèle classique d'évolution après duplication complète : la duplication est suivie de la diploïdisation puis de la spéciation. Les gènes descendant de chacune des copies dupliqués sont groupés ensemble, comme attendu, dans l'arbre de gène. B. Modèle d'évolution de gènes dupliqués dans le cas d'héritage tétrasomique : les homéologues continuent de s'apparier en méiose après la spéciation des espèces. Les gènes sont groupés par espèce dans l'arbre de gènes. Figure adaptée de MARTIN et HOLLAND 2014.

Le comportement tétrasomique prolongé des génomes de Salmonidés a notamment été caractérisés à travers l'étude des topologies des arbres des gènes retrouvés dans ces régions (Figure 1.15, ROBERTSON et al. 2017). Dans le cas classique d'un comportement disomique, on s'attend à retrouver dans les arbres de gènes deux grands groupes après l'événement de duplication : un groupe avec la copie 'a' du gène dans toutes les espèces et un groupe avec la seconde copie, 'b'. Dans les régions au comportement tétrasomique, il apparaît que les 2 copies dupliquées sont groupées par espèces dans les arbres, mettant en avant leur plus grande similarité de séquence au sein d'une espèce due à leur recombinaison méiotique qui existait encore au moment de la spéciation.

L'origine de la duplication complète des téléostéens n'est pas connue, les 3 mécanismes possibles (auto-, allo- et allopolyploïde segmentaire) ont été proposés dans la littérature (MARTIN et HOLLAND 2014 ; CONANT 2020). Dans mon travail de thèse, je propose d'analy-

ser l'histoire évolutive des chromosomes de poissons téléostéens, en incluant des d'espèces ayant divergé tôt après la 3R (Osteoglossiformes) pour mieux comprendre les mécanismes de la rediploïdisation chez les poissons téléostéens.

1.4 Problématiques

J'ai mis en avant, dans cette introduction, l'importance de l'inférence de l'histoire évolutive des gènes comme point de départ pour répondre à de nombreuses questions de la génomique comparative. Les arbres de gènes forment une base pour l'annotation fonctionnelle des génomes et l'étude de l'évolution de la structure des chromosomes. Ils permettent également d'examiner des questions plus spécifiques, comme celles liées à l'évolution après duplication complète. Cependant, les méthodologies actuelles de reconstruction d'arbres de gènes sont limitées en présence de nombreux événements de duplications de gènes (TRACHANA et al. 2011).

Dans mon travail de thèse, j'ai développé une nouvelle méthodologie spécifique à la reconstruction d'arbres de gènes dans le contexte de duplications complètes de génome, nommée SCORPiOs (Synteny-guided CORrection of Paralogies and Orthologies). L'idée derrière SCORPiOs est de compléter les méthodes de reconstruction d'arbres basées sur l'évolution de la séquence en utilisant les patrons de conservation de la synténie (DCS) spécifiques aux duplications complètes de génomes. L'application de cette méthodologie à différents jeux de données d'espèces de poissons, m'a permis d'explorer les points suivants :

- (i) préciser l'impact fonctionnel des duplications de gènes,
- (ii) retracer l'histoire évolutive des régions dupliquées chez les poissons,
- (iii) caractériser l'origine de la duplication complète à la base des poissons téléostéens.

La structure de ma thèse s'articule autour de ces trois grands points. La première partie présente la mise au point et la validation de SCORPiOs. Elle montre que la méthode est applicable à la duplication complète des poissons téléostéens et que l'amélioration des arbres permet de lier rétention de paralogues et innovations évolutives des téléostéens. La seconde partie concerne l'application de SCORPiOs à un grand jeu de génomes de poissons. Cette application a pour objectif de faciliter les études de génomiques comparatives au sein des poissons via la résolution de l'histoire évolutive des gènes dupliqués et l'établissement de la carte des régions dupliquées. Enfin, la dernière partie s'intéresse à mieux comprendre les événements de duplications complètes de génome, de manière générale, car ils sont indissociables de la compréhension des génomes de poissons et de vertébrés. En particulier, je m'intéresse à caractériser les patrons de rediploïdisation dans les premiers stades de la divergence des téléostéens.

Chapitre 2

La résolution des arbres de gènes à travers la synténie conservée clarifie l'impact fonctionnel des duplications complètes

2.1 Introduction

Je présente dans cette partie un article que j'ai co-écrit en tant que première auteure, intitulé « Synteny-guided resolution of gene trees clarifies the functional impact of whole-genome duplications » et publié dans *Molecular Biology and Evolution*. Ce travail répond au problème que j'ai introduit dans le chapitre précédent : l'inférence de l'histoire évolutive de gènes, en présence de duplication(s) complète(s). L'article est constitué de deux grandes composantes. La première a trait au développement méthodologique qui a donné lieu à la production d'un nouvel outil de correction d'arbres, nommé SCORPiOs. Ici, je commencerai par introduire, de manière plus détaillée que dans le manuscrit publié, le paysage méthodologique dans lequel SCORPiOs s'insère. Je justifie également les choix que j'ai effectués pour sa conception. La seconde composante correspond à l'application de SCORPiOs à la duplication complète des poissons téléostéens, sur un jeu d'arbres de la base de données d'Ensembl Compara. J'introduis brièvement les résultats de cette application dans la dernière partie de cette introduction, en les mettant en relation avec les connaissances qui nous viennent d'autres événements de duplication complète.

2.1.1 Contexte méthodologique

L'objectif de SCORPiOs est de permettre une identification correcte des gènes orthologues et paralogues, dans le contexte de duplications complètes de génome. Les orthologues sont des gènes qui ont divergé après séparation des espèces (spéciation), tandis que

les paralogues sont séparés par un événement de duplication. Ces relations d'orthologie et de paralogie peuvent être directement extraites des arbres de gènes. Cependant, comme évoqué dans le chapitre précédent (voir le paragraphe 1.1.3), les duplications sont des sources d'erreurs dans les arbres de gènes et il n'existe pas de méthodologie qui prennent en compte l'occurrence de duplication complète de génome. Dans cette sous-partie, je présente les principaux outils généralistes pour l'inférence d'orthologie et motive le développement d'une méthode de correction d'arbres.

Méthodes de prédiction d'orthologie

Comme présenté dans le chapitre d'introduction, l'identification des gènes orthologues est une pierre angulaire de la génomique comparative. De fait, un grand nombre de méthodes ont été développées pour répondre à ce problème. Les principales stratégies employées par les grandes bases de données pour l'inférence d'orthologues sont présentées dans le Tableau 2.1. Elles sont classiquement séparées en deux grandes familles : les méthodes de graphes et les méthodes d'arbres. Dans les deux cas, la conservation de séquence entre gènes de différentes espèces sert à prédire leur origine évolutive commune. La synténie conservée, ou conservation de blocs de gènes ordonnés entre espèces (voir paragraphes 1.2.3 et 2.1.3), constitue un second indicateur d'orthologie rarement directement intégré à ces méthodes, mais permettant leur validation *a posteriori*.

La différence majeure entre les graphes et les arbres est qu'ils ne répondent pas exactement au même problème. Les arbres retracent l'histoire évolutive complète des gènes, c'est à dire la chronologie de tous les événements de duplication et de spéciation. Les graphes répondent en réalité à un sous-problème : ils dérivent des groupes de gènes orthologues à différents niveaux phylogénétiques (« Orthogroupes hiérarchiques », Figure 2.1).

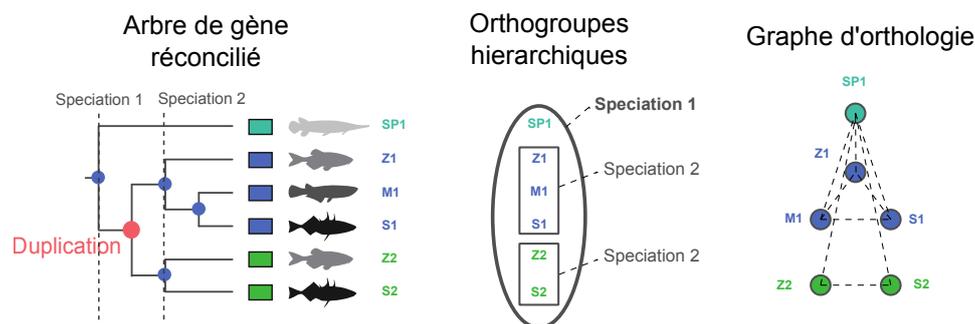


FIGURE 2.1 – Illustration des méthodes d'arbres et de graphes. L'arbre décrit l'histoire évolutive complète d'un gène en identifiant les événements de spéciation (bleu) et de duplication (rouge). Les méthodes de graphes renseignent des groupes hiérarchiques à différentes spéciations, ou une collection de paires d'orthologies (graphe d'orthologie). Ici, les spéciations 1 et 2 forment un total de 3 orthogroupes. A noter que tous les gènes d'un orthogroupe hiérarchique ne sont pas nécessairement orthologues entre eux : ici, à la spéciation 1, Z1 et S2 sont paralogues.

Introduction aux méthodes de graphes

Dans les graphes, les gènes de plusieurs espèces sont représentés par des nœuds, reliés entre eux par des arêtes qui représentent les relations d'orthologie (Figure 2.1). Les méthodes de graphes sont sub-divisées en deux sous-classes : celles qui infèrent des orthologies entre une paire d'espèces et les méthodes multi-espèces. La méthode par paire la plus simple est basée sur le principe du meilleur « hit » bidirectionnel (BBH). Ainsi, deux gènes réciproquement identifiés comme les plus similaires entre deux génomes sont prédits orthologues.

Les méthodes multi-espèces généralisent l'identification d'orthologues en considérant tous les gènes de plusieurs espèces dans le même graphe. En plus d'être plus générales, elles corrigent les limites des méthodes par paire qui peuvent être confondues en cas de comparaison d'espèces distantes ou de pertes différentielles. Différentes approches de clustering de graphe permettent ensuite d'extraire des groupes de gènes orthologues à un niveau phylogénétique donné. L'avantage principal des méthodes de graphes par rapport aux méthodes d'arbres est leur faible complexité algorithmique, permettant d'intégrer plusieurs centaines d'espèces (par exemple OMA intègre plus de 2 000 génomes, ALTENHOFF et al. 2018). Plus de détails concernant l'implémentation des principales méthodes de graphes sont reportés dans le Tableau 2.1.

Introduction aux méthodes d'arbres

La reconstruction d'un arbre de gènes correspond à un problème plus général et plus complexe que la recherche d'orthologues, ce qui se traduit également en terme de complexité computationnelle. En premier lieu, la reconstruction d'arbres nécessite de définir tous les jeux de gènes homologues, c'est à dire toutes les familles de gènes descendant d'une même séquence ancestrale, que ce soit par duplication ou par spéciation. Cette étape est en réalité une méthode de graphe multi-espèces : le clustering est moins conservateur que dans les méthodes introduites plus haut, de manière à assembler un jeu complet de gènes homologues. Ensuite, il s'agit d'inférer un alignement multiple, puis le modèle d'arbre et enfin sa réconciliation à l'arbre des espèces afin d'identifier les événements de duplications.

L'inférence de l'arbre de gènes se fait le plus souvent par maximum de vraisemblance : il s'agit de chercher le modèle le plus probable d'avoir généré les séquences de gènes observées. Trouver l'arbre de maximum de vraisemblance est un problème np-complet, nécessitant l'utilisation de méthodes heuristiques pour parcourir l'espace des topologies d'arbres, paramètres du modèle d'évolution et longueurs de branches. Ce processus, brièvement introduit au paragraphe 1.1.3, est détaillé au paragraphe 2.1.2.

Méthode	Classe	Recherche d'homologie	Clustering	Orthogroupes hiérarchiques	Arbre	Réconciliation	Référence
InParanoid	graphe, paire	BLAST	-	Non	Non	-	SONNHAMMER et ÖSTLUND 2015
RoundUp	graphe, paire	Distance évolutive (ML)	-	Non	Non	-	DELUCA et al. 2012
OMA	graphe, multi-espèces	Distance évolutive (ML)	Clique maximum	Oui	Non	-	ALTENHOFF et al. 2018
OrthoMCL	graphe, multi-espèces	BLAST	Markov	Non	Non	-	CHEN et al. 2006
OrthoDB	graphe, multi-espèces	PARALIGN	Triangles	Oui	Non	-	KRIVENTSEVA et al. 2019
eggNOG	graphe, multi-espèces	BLAST	Triangles	Oui	Non	-	MULLER et al. 2010
HieranoiDB	graphe, multi-espèces	BLAST	-	Oui	Oui*	-	KADUK et al. 2017
PhylomeDB	arbre	BLAST	Seuil sur la e-value	Oui	ML, NJ	species overlap (parcimonie)	HUERTA-CEPAS et al. 2014
Ensembl Compara	arbre	BLAST+, HMM	Hcluster_sg (clustering hiérarchique)	Oui	ML+, NJ	sdi (parcimonie)	VILELLA et al. 2009
OrthoFinder	arbre	DIAMOND	Markov	Oui	ML	species overlap, DLC	EMMS et KELLY 2019

TABLE 2.1 – Principales méthodes de prédiction d'orthologie utilisées à grande échelle. DLC : Duplication-Loss-Coalescent (réconciliation considérant les pertes, duplications et le tri de lignées incomplet). ML : maximum de vraisemblance. ML+ : la vraisemblance contient un terme qui correspond à la réconciliation à l'arbre des espèces. NJ : Neighbor-Joining. * Hieranoid guide les comparaisons par paire de InParanoid en utilisant la structure de l'arbre des espèces et représente les orthologues et in-paralogues obtenus sous forme d'arbre. Adapté de TRACHANA et al. 2011 et ALTENHOFF et DESSIMOZ 2012

Motivation pour la conception d'une méthode de correction d'arbres

Dans la pratique, les méthodes de graphes et méthodes d'arbres sont loin d'être fondamentalement opposées. L'exemple d'OrthoFinder montre bien que la barrière est mince : initialement publiée en tant que méthode de graphes, l'addition d'une étape d'inférence d'arbres dans sa deuxième version a reclassé OrthoFinder dans la famille des méthodes d'arbres (EMMS et KELLY 2015, 2019). De plus, il ne ressort pas de tendance générale de performance entre les deux grandes classes de méthodes, les outils (graphes et arbres confondus) se situent tous à des points différents du compromis précision/rappel (ALTENHOFF et al. 2016). D'une part, les graphes n'informent pas directement quant à la position des duplications de gènes. D'autre part, les topologies d'arbres sont souvent incertaines en présence de duplications. En effet, la présence de nombreuses copies de gènes rend l'espace des solutions d'arbres plus vaste et donc plus difficile à explorer efficacement. De plus, en absence de signal phylogénétique suffisant dans les séquences, la séparation correcte des gènes dupliqués en orthologues et paralogues est parfois mal supportée. Enfin, les duplications complètes sont suivies d'une accélération du taux de perte de gènes, souvent hétérogènes entre espèces, et non pris en compte lors de la reconstruction des arbres. En conclusion, aucune méthode introduite précédemment ne considère spécifiquement le problème de l'identification de gènes orthologues et paralogues dans le cadre d'anciennes duplications complètes. Les seuls développements dans ce sens se limitent à la caractérisation des sous-génomés de néopolyploïdes, chez les plantes, (intégrée, par exemple, dans OMA et Ensembl Compara).

A partir de ce paysage méthodologique, plusieurs possibilités s'offrent pour le développement d'un outil spécifique au problème des duplications complètes de génome. Il serait possible d'intervenir à l'étape des graphes. A partir des familles complètes des gènes homologues, il s'agirait d'identifier, parmi les espèces dupliquées, les paralogues issus de la duplication complète. Cependant, en présence d'autres événements de duplication (familles multigéniques), il n'est alors pas trivial de traduire ces relations d'orthologie et de paralogie au niveau de la duplication complète en terme de topologie pour l'arbre complet (LAFOND et EL-MABROUK 2014). Pourtant, la reconstruction d'un arbre est désirable, car l'arbre de gènes forme un modèle phylogénétique rigoureux permettant de tester un panel d'hypothèses évolutives. Dans cette optique, il est envisageable de partir directement d'arbres de gènes pré-calculés pour identifier et corriger les nœuds peu soutenus. L'avantage de cette méthode est qu'elle permet de traiter des grandes familles de gènes. Un second avantage est que les arbres sont plus informatifs que les graphes : les relations d'orthologie et de paralogie sont déjà prédites et il ne s'agit alors, en théorie, que d'identifier et corriger celles qui sont erronées. De plus, cette méthode se généralise facilement à la correction de plusieurs événements de duplication complète dans un même modèle. Nous avons donc choisi de développer SCORPiOs comme une méthode de correction d'arbres phylogénétiques.

2.1.2 Reconstruction d'arbres de gènes

Inférence d'arbre de gènes par maximum de vraisemblance

L'arbre de gènes décrit un modèle probabiliste permettant d'évaluer la probabilité d'observer les données dans ce modèle, ou vraisemblance du modèle. Les données, ici, sont un alignement de séquence : les colonnes de l'alignement représentent des caractères (acides aminés ou nucléotides) supposés homologues. Le processus de substitutions au cours du temps est formalisé par un modèle d'évolution, décrivant le processus biologique de manière simplifiée. Notamment, l'hypothèse principale des modèles d'évolution de séquence est que les colonnes de l'alignement sont indépendantes, c'est à dire que la vraisemblance est assimilable au produit de la vraisemblance calculée pour chaque site. La vraisemblance d'un site est donnée par l'intégration de tous les scénarios de substitutions pouvant mener à l'observation de la colonne d'alignement. En plus de décrire les taux de substitutions, les modèles d'évolution permettent généralement à différents sites de l'alignement d'évoluer à des vitesses différentes, pour tenir compte de régions évolutivement plus contraintes. En revanche, ils modélisent plus rarement l'hétérogénéité des taux de substitutions entre lignées.

L'heuristique utilisée pour explorer l'espace des solutions est une stratégie de hill-climbing, qui consiste à perturber un arbre initial et, itérativement pour chaque paramètre, accepter les changements qui améliorent la vraisemblance. Les topologies d'arbres sont explorées par des réarrangements de sa structure, par deux stratégies principales : le NNI ("Nearest Neighbor Interchange") qui consiste à échanger la position des 4 sous-arbres connectés à une branche ou le SPR ("Subtree Prune and Regraft") qui détache un sous-arbre et le recolle à une autre position dans l'arbre. Il a été montré que le SPR est moins susceptible d'être piégé par un optimum local mais il est computationnellement plus coûteux. Parmi les méthodes principales d'inférences d'arbres par maximum de vraisemblance, on peut citer RAxML, PhyML et IQ-TREE (GUINDON et al. 2010; STAMATAKIS 2014; MINH et al. 2020).

Réconciliation à l'arbre des espèces

La réconciliation d'un arbre de gènes avec la phylogénie des espèces consiste à expliquer les désaccords observés entre les deux topologies par un jeu d'événements biologiques définis, comprenant classiquement : duplications, pertes, transferts horizontaux et/ou tri de lignées incomplet (ILS, incomplete lineage sorting). Le tri de lignées incomplet peut se produire lorsque plusieurs spéciations se produisent successivement et rapidement dans le temps. Si les polymorphismes présents dans les populations ancestrales se fixent indépendamment (et différenciellement) dans les espèces descendantes, le tri de lignées incomplet peut engendrer des discordances entre l'histoire des gènes et des espèces. Peu de méthodes de réconciliation modélisent l'ILS et il est suggéré, chez les mammifères, que l'ILS ne soit qu'une cause mineure des incongruences entre arbres de gènes et arbres d'espèces

(SCORNAVACCA et GALTIER 2017). Les méthodes de réconciliation sont classées en deux types : les méthodes de parcimonie qui cherchent la réconciliation impliquant le moins d'événements (ou la somme la plus faible d'événements pondérés) et les méthodes probabilistes qui modélisent le processus de génération de l'arbre de gènes sur l'arbre des espèces. Typiquement, les méthodes de parcimonie Duplication-Perte (méthode « sdi » avec un arbre d'espèces ou sans « species overlap ») minimisent le nombre de duplications et pertes, en plaçant les duplications aux nœuds séparant deux gènes de la même espèce (species overlap) ou le plus bas possible sur l'arbre des espèces (sdi). Les méthodes probabilistes sont plus coûteuses car, comme pour l'inférence de l'arbre, elles cherchent le modèle de réconciliation et ses paramètres les plus vraisemblables (taux des événements de perte, duplication et transfert ou autres), parmi l'espace des réconciliations possibles.

Méthodes STA (Species-Tree Aware)

Le terme de réconciliation est parfois utilisé pour décrire à la fois la simple annotation des nœuds internes de l'arbre de gènes ou le réarrangement de sa topologie afin de rendre cette annotation plus parcimonieuse. En effet, les dernières années ont vu fleurir une quantité de nouvelles méthodes proposant de prendre en compte la proximité à l'arbre des espèces pour améliorer la topologie de l'arbre de gènes (pour une revue détaillée voir SZÖLLŐSI et al. 2015). Lorsque l'arbre inféré par maximum de vraisemblance est directement réconcilié à l'arbre des espèces, les incertitudes liées à son inférence ne sont pas considérées. De fait, les méthodes « STA » permettent généralement d'inférer des arbres plus corrects, mais les plus sophistiquées sont trop coûteuses pour être appliquées à des jeux de données de plusieurs dizaines d'espèces. Une première catégorie de méthodes est fondée sur le principe de l'amalgamation : à partir d'une distribution *a posteriori* des topologies d'arbres, l'espace des arbres réconciliés est exploré en combinant les clades observés dans l'échantillon. L'arbre maximisant un score joint séquence-réconciliation est sélectionné. Ces stratégies sont employées par exemple par ALE et (ecce)TERA, avec la différence notable que ALE utilise un modèle probabiliste de réconciliation et ecceTERA une approche par parcimonie (SZÖLLŐSI et al. 2013 ; SCORNAVACCA, JACOX et SZÖLLŐSI 2015).

Une seconde famille de méthodes, moins intensives car elles ne nécessitent pas l'obtention de distribution *a posteriori* d'arbres, parcourt l'espace des arbres réconciliés à partir d'un arbre de départ, en inférant itérativement les paramètres de maximum de vraisemblance du modèle de réconciliation et d'évolution des séquences (Phyldog et Generax, BOUSSAU et al. 2013 ; MOREL et al. 2020). Enfin, avec TreeBeST (VILELLA et al. 2009), la version très récente d'OrthoFinder (EMMS et KELLY 2019) représente la seule méthode STA applicable à des jeux de plusieurs dizaines d'espèces. Le modèle de réconciliation d'OrthoFinder est une méthode hybride combinant modèle probabiliste et parcimonie, « DLCpar » (duplications, pertes, tri de lignées incomplet, WU et al. 2014). Dans une première passe, les nœuds de duplication sont annotés par parcimonie puis localement réarrangés par DLCpar. Le cas

particulier de la méthode TreeBeST est détaillé dans le prochain paragraphe.

Reconstruction d'arbres par la base de données d'Ensembl : TreeBeST

TreeBeST (VILELLA et al. 2009) emploie une stratégie plus simpliste pour répondre au problème de l'inférence d'arbres de gènes réconciliés. La topologie finale de l'arbre inféré par TreeBeST est un arbre consensus, établi à partir de la combinaison de 5 arbres prédits par différentes méthodes (3 arbres « Neighbor-Joining » basés sur une matrice de distance, et 2 arbres inférés par maximum de vraisemblance). Les arbres de maximum de vraisemblance de TreeBeST sont estimés en utilisant l'algorithme de PhyML (avec une exploration de l'espace des topologies par « NNI »), soit une méthode de maximum de vraisemblance « classique » (GUINDON et GASCUEL 2003). Cependant, le calcul de la vraisemblance finale de TreeBeST inclut un second terme, qui correspond à la vraisemblance de la réconciliation la plus parcimonieuse. Les taux de duplications et pertes sont fixés et ne sont pas des paramètres du modèle de réconciliation. En résumé, TreeBeST n'explore que superficiellement l'espace des arbres réconciliés possibles, mais le critère de réconciliation est tout de même intégré dans le choix de la topologie finale. Les arbres sont généralement plus corrects que les arbres purement basés sur l'évolution des séquences, pour un temps de calcul réduit par rapport aux autres méthodes « STA ». De part la faisabilité de TreeBeST pour reconstruire des arbres pour des jeux contenant plus d'une centaine d'espèces, nous avons choisi de développer SCORPiOs comme une méthode de correction d'arbres initiaux générés par TreeBeST.

2.1.3 Indicateurs d'incertitude dans les arbres de gènes

Je présente ici différents indicateurs permettant d'identifier les incertitudes ou erreurs dans les arbres de gènes. Ces indicateurs peuvent ensuite permettre de guider la correction des arbres.

Support de bootstrap

Le support que les données de séquence apportent à un nœud particulier de l'arbre peut être quantifié par une procédure de bootstrap (ou par des probabilités postérieures dans le cadre bayésien). Le principe du bootstrap a été introduit par Efron en 1979 et repris dans le contexte phylogénétique par Felsenstein en 1985 (EFRON 1979 ; FELSENSTEIN 1985). Il s'agit d'une mesure de répétabilité, qui indique à quel point les inférences sont robustes à un ré-échantillonnage des données. La procédure de bootstrap consiste à ré-échantillonner les colonnes de l'alignement de séquences et ré-inférer un arbre de gènes pour chaque réplicat. Le score de bootstrap correspond alors à la proportion des arbres recalculés qui contiennent les nœuds de l'arbre original. De part son utilité et sa simplicité conceptuelle, le bootstrap est très vite devenu un indicateur indispensable en phylogénie moléculaire :

l'article de Felsenstein est désormais parmi les 100 articles scientifiques les plus cités de tous les temps. Il n'existe pas de règle absolue quant au score de bootstrap désirable pour supporter une phylogénie, même si l'utilisation d'un seuil à 70% est très répandue. L'absence de seuil fixe est principalement due au fait que le bootstrap varie en fonction de la taille du sous-arbre sous un nœud considéré. Une méthode récemment développée permet de corriger ce biais, et tendra probablement à se démocratiser dans les études à venir (LEMOINE et al. 2018).

Score de confiance des nœuds de duplication

En complément, les nœuds de duplication qui impliquent un grand nombre de pertes traduisent un scénario d'évolution peu parcimonieux et peuvent également servir à l'identification de nœuds mal résolus. Par exemple, la base de données Ensembl annote certains nœuds de duplication comme incertains (ou "dubious") parce qu'ils impliquent un événement de duplication de gène suivi d'une perte dans toutes les espèces descendantes (Figure 2.2). Il y a une forte corrélation entre les scores de confiance des nœuds de duplications (calculé comme la proportion d'espèces ayant retenu le gène après la duplication) et les scores de bootstrap (VILELLA et al. 2009), ce qui tend à valider leur pertinence respective.

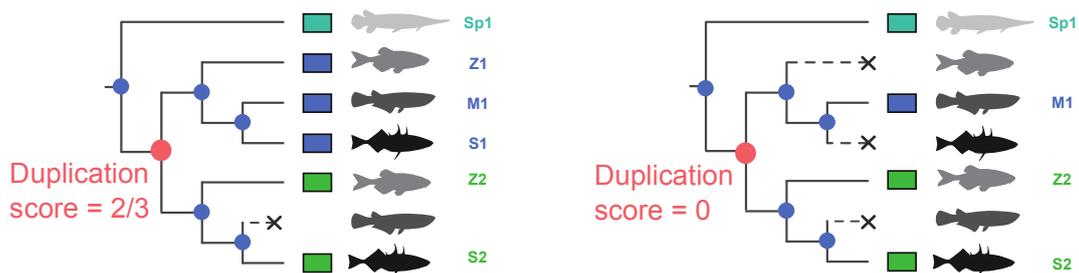


FIGURE 2.2 – Score de confiance des nœuds de duplication. Le score de duplication est donné par le rapport du nombre d'espèces ayant gardé les deux copies du gènes dupliqués et le nombre total d'espèces sous le nœud. Les nœuds induisant une perte dans toutes les espèces impliquent un scénario peu parcimonieux.

Synténie conservée

L'identification de blocs de synténie conservée permet d'identifier des gènes orthologues, sous l'hypothèse que les gènes à l'organisation génomique conservée entre espèces descendent d'un même bloc ancestral. Si ces gènes prédits orthologues par la synténie ne sont pas retrouvés orthologues dans l'arbre de gènes, alors sa topologie est potentiellement incorrecte. Par conséquent, la synténie conservée peut également être utilisée pour identifier des erreurs dans les arbres, même si le lien avec la topologie de l'arbre est moins direct.

Plusieurs outils permettent d'identifier des blocs de synténie conservée entre paires de génomes. Ils se basent sur une étape de BLAST permettant de définir les gènes homologues

(par exemple i-ADHORE ou MCScanX, PROOST et al. 2012; WANG et al. 2012) ou directement sur les familles extraites d'arbres (PhylDiag, LUCAS, MUFFATO et CROLLIUS 2014). MCScanX est extrêmement populaire en génomique des plantes, où les gènes ont des histoires évolutives complexes, avec de nombreuses duplications. MCScanX propose, en plus de l'identification des blocs de synténie, une classification des gènes dupliqués (duplication complète, tandem, segmental) en fonction de leur localisation au sein d'un génome. A l'exception de ParalogyCorrector (LAFOND et al. 2013 voir la prochaine sous-partie), la synténie n'a jamais été utilisée dans le contexte de la correction des arbres de gènes.

2.1.4 Méthodes de correction des arbres de gènes

A partir de l'observation que certains nœuds des arbres de gènes étaient peu soutenus par les indicateurs introduits précédemment, des méthodes de correction ont été développées.

Correction des nœuds aux scores de bootstrap faibles

Une première approche vise à recalculer les sous-arbres les moins bien supportés par l'alignement des séquences. Par exemple, ProfileNJ (NOUTAHI et al. 2016) contracte les nœuds pour lesquels le score de bootstrap est inférieur à un seuil fixé et propose une nouvelle solution. Ce nouveau sous-arbre est inféré de manière à être à la fois parcimonieux en terme de scénario de pertes et de duplications, et supporté par les données de séquences. La résolution des sous-arbres contractés utilise un algorithme polynomial efficace pour inférer l'histoire des duplications, combiné au principe du Neighbor-Joining pour positionner les gènes dupliqués. Dans la pratique, cela permet d'appliquer ProfileNJ à l'échelle de grandes bases de données d'arbres de gènes, comme celle d'Ensembl Compara (NOUTAHI et al. 2016). Plus récemment, Treerecs étend le principe de ProfileNJ en ajoutant un critère de sélection du seuil de bootstrap utilisé pour la contraction des branches, reposant sur la vraisemblance jointe des modèles d'évolution de séquence et de la réconciliation (COMTE et al. 2020).

Correction des arbres au scénario de duplication/perte peu parcimonieux

Une seconde famille de méthodes considère le coût de réconciliation à l'arbre des espèces pour améliorer les arbres. La topologie autour des nœuds de duplication les moins bien supportés sont réarrangés pour obtenir un scénario de duplication plus parcimonieux. C'est le cas notamment de méthodes comme NOTUNG ou Treefix (CHEN, DURAND et FARACH-COLTON 2000; WU et al. 2013). C'est également le cas des arbres de gènes de la base de données Genomicus, où l'édition des nœuds de duplication peu soutenus permet de mieux reconstruire l'ordre des gènes dans les génomes ancestraux (MUFFATO et al. 2010).

Correction basée sur la conservation de la synténie

Enfin, ParalogyCorrector (LAFOND et al. 2013) est la seule méthode qui utilise la synténie conservée pour corriger les arbres de gènes. ParalogyCorrector prend en entrée des relations d'orthologie prédites à travers l'identification de blocs de synténie entre paires de génomes. Lorsque ces relations d'orthologie ne sont pas retrouvées dans les arbres originaux, ParalogyCorrector réarrange les branches de l'arbre de façon à rendre ces gènes orthologues.

Plus précisément, l'algorithme développé permet de trouver l'arbre de gènes le plus proche topologiquement de l'arbre original, au sens de la distance de Robinson-Foulds (RF), tout en respectant les contraintes d'orthologie dérivées de l'analyse de synténie. Cependant, on peut faire un parallèle avec les méthodes de graphes, puisque ParalogyCorrector corrige uniquement les relations d'orthologie entre gènes. Comme plusieurs topologies sont possibles pour un jeu d'orthologies (Figure 2.3), le choix d'un second critère est nécessaire (ici la distance RF). Cela implique que les relations de paralogie ne sont pas nécessairement correctes dans les arbres corrigés. En réalité, nous avons observé, au cours du développement de SCORPiOs, que ParalogyCorrector était biaisé vers le positionnement des nœuds de duplications proche des feuilles (résultat présenté dans l'article en Figure 3D).

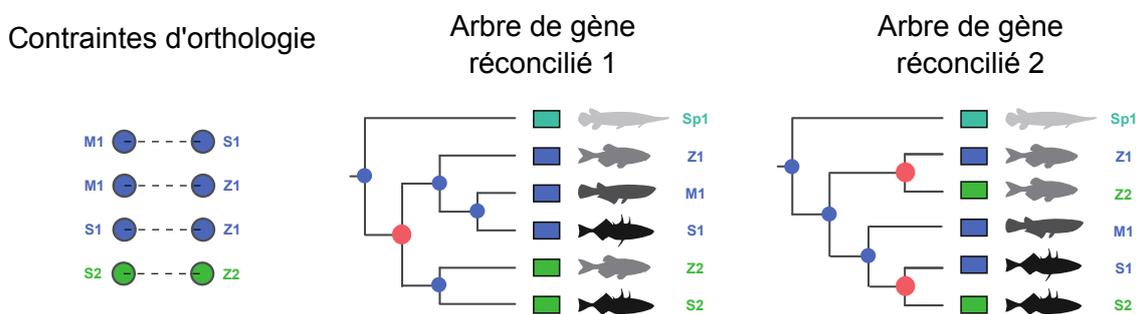


FIGURE 2.3 – Contraintes d'orthologie et topologies d'arbres possibles. Pour un jeu de relations d'orthologies données, plusieurs topologies d'arbres sont possibles, impliquant différentes relations de paralogie.

Le cas des duplications complètes

Dans le cas particulier des duplications complètes, il apparaît évident que considérer les patrons de synténie conservée est un indicateur pertinent, au regard des signatures génomiques spécifiques laissées par la duplication (synténie doublement conservée, voir le paragraphe 1.3.2). Un second avantage du fait de travailler avec les duplications complètes est que la position du nœud de duplication est connu : la duplication est à la base des gènes d'espèces dupliquées. Dans ce contexte, le problème qui consiste à traduire des contraintes d'orthologie et de paralogie en terme de topologie d'arbre devient plus simple : il s'agit de

positionner correctement les gènes de part et d'autre de l'événement de duplication connu.

De plus, si les relations d'orthologie et paralogie dérivées par l'analyse de synténie ne sont pas complètes, il est possible de s'appuyer sur les relations existantes dans l'arbre original. C'est ainsi que fonctionne SCORPiOs : si les patrons de synténie ne permettent pas de prédire l'histoire évolutive d'un gène, il est replacé dans l'arbre dans le groupe de son plus proche homologue dans l'arbre initial. En résumé, nous avons choisi de construire SCORPiOs comme une méthode de correction d'arbre qui s'appuie sur l'analyse des patrons de synténie doublement conservée (DCS).

2.1.5 Principe de SCORPiOs

J'introduis ici brièvement le principe de SCORPiOs sans entrer dans les détails de son implémentation qui se trouvent dans l'article publié. La conception de SCORPiOs a consisté en deux étapes principales : le développement d'une méthode d'analyse des patrons de synténie dans le cadre de duplication complète et son intégration dans un pipeline de reconstruction d'arbres phylogénétiques. Je tire largement profit des logiciels existant pour inférer des arbres de gènes et les tester. Le choix de ces différents outils a été effectué pour garantir la faisabilité de SCORPiOs sur des jeux de données d'une centaine d'espèces, en s'inspirant notamment des méthodes utilisés par Ensembl Compara (TreeBeST, VILELLA et al. 2009). Les outils utilisés ainsi que la validation de SCORPiOs sont présentés en détails dans l'article publié.

Principes de l'analyse des patrons de synténie

SCORPiOs repose sur la connaissance *a priori* de la position d'un événement de duplication dans l'arbre des espèces. Pour l'analyse des patrons de synténie, les deux principes sur lesquels SCORPiOs est fondé sont les suivants : (i) les gènes dupliqués par la duplication complète se trouvent en synténie doublement conservée avec les gènes d'une espèce non-dupliquée, (ii) les blocs de synténie orthologues entre espèces dupliquées peuvent être identifiés par leur patron de perte/rétention et leur évolution moléculaire commune.

L'application de ces deux principes permet d'inférer des relations d'orthologie et de paralogie entre gènes d'espèces dupliquées. Pour une famille de gènes donnée, SCORPiOs confronte l'arbre initialement inféré à partir des séquences aux prédictions d'homologie dérivées de (i) et (ii). L'arbre est ensuite corrigé si les relations prédites n'y sont pas retrouvées.

Intégration de l'information de synténie dans les arbres

Un point essentiel du fonctionnement de SCORPiOs est qu'il ne force pas la topologie inférée par l'analyse des patrons de synténie mais s'en sert pour guider le problème de la

sélection d'arbre (Figure 2.4). Dans la pratique, SCORPiOs compare le support statistique apporté par les séquences à l'arbre corrigé contre celui apporté à l'arbre original.

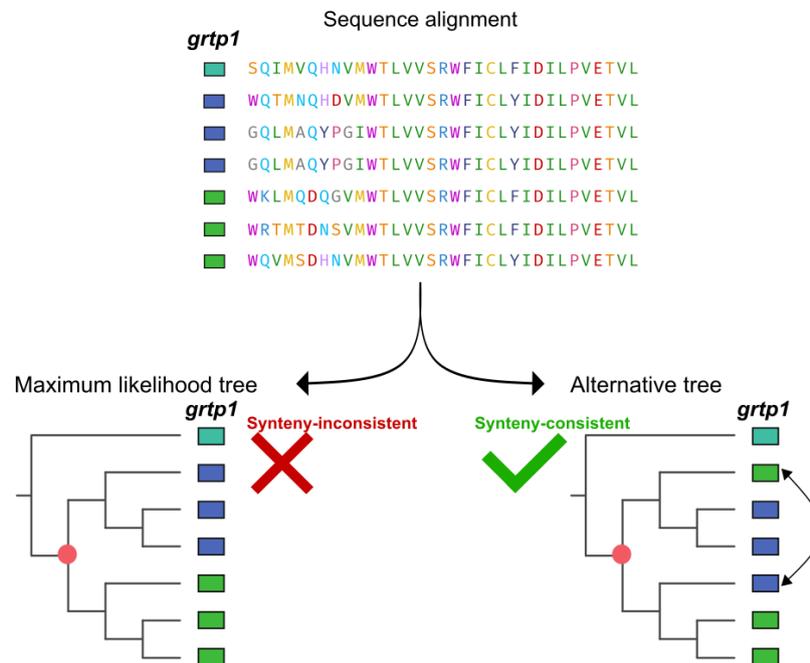


FIGURE 2.4 – Utilisation de la synténie conservée pour guider le problème de sélection d'arbre. A partir de l'alignement de gènes présenté, deux topologies d'arbres sont possibles : l'arbre de maximum de vraisemblance et une topologie alternative statistiquement équivalente. L'analyse des patrons de synténie permet de guider la sélection de la topologie.

Plusieurs tests permettent de comparer la vraisemblance de différents arbres de gènes (pour une revue détaillée, voir PLANET 2006). Le test KH (KISHINO et HASEGAWA 1989), introduit par Kishino et Hasegawa, permet de tester l'hypothèse nulle que deux topologies sont aussi bien supportées par les données : la différence de vraisemblance observée entre les deux arbres est seulement due au biais d'échantillonnage (les séquences sont finies). Une procédure de bootstrap permet d'estimer la distribution de la différence de vraisemblance sous l'hypothèse nulle. Enfin, une p-valeur est dérivée en confrontant la différence de vraisemblance observée à la distribution nulle. Le test KH souffre du problème de « biais de sélection » (GOLDMAN, ANDERSON et RODRIGO 2000) : inclure l'arbre de maximum de vraisemblance parmi les topologies testées invalide la procédure d'estimation de la distribution nulle. Pour corriger ce problème, le test SH a été introduit (Shimodaira et Hasegawa, SHIMODAIRA et HASEGAWA 1999). Cependant, deux limites principales du test SH sont apparues : il est très conservateur et, en théorie, l'espace complet des topologies possibles devrait être considéré pour qu'il soit valide. Enfin, le test AU (« Approximately Unbiased ») est le test de vraisemblance d'arbre le plus populaire, effaçant les limites des deux précédents (SHIMODAIRA 2002). Il permet de calculer, pour un jeu de topologies, la probabilité que chaque arbre soit inclus dans un intervalle de confiance autour du meilleur arbre.

Par conséquent, SCORPiOs utilise le test AU pour comparer les topologies et ne corrige les arbres que si l'arbre dérivé de la synténie est au moins statistiquement équivalent à l'arbre initial.

SCORPiOs en résumé

En résumé, SCORPiOs permet de construire un jeu d'arbres de gènes cohérent avec un ou des événement(s) de duplication complète connu(s), à partir d'un jeu d'arbres de gènes, d'alignements, l'arbre des espèces et les positions des gènes dans les génomes (Figure 2.5). Afin de faciliter la distribution de SCORPiOs, je l'ai implémenté dans un langage de définition de pipeline, Snakemake (KÖSTER et RAHMANN 2012), et intégré avec conda pour la gestion des dépendances. Le code est disponible sur GitHub (<https://github.com/DyogenIBENS/SCORPIOS>) et possède une documentation complète sur ReadTheDocs (<https://scorpios.readthedocs.io/en/stable/>).

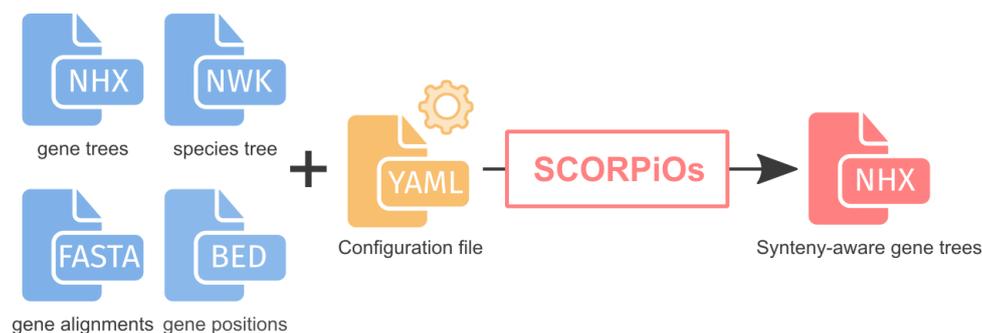


FIGURE 2.5 – Représentation simplifiée de SCORPiOs.

2.1.6 Apport aux connaissances sur l'évolution après duplication complète

La seconde partie de l'article présenté concerne l'application de SCORPiOs aux arbres de gènes disponibles dans la base de données d'Ensembl Compara. Cette application s'est effectuée sur la version 89 de la base de données, qui comprenaient alors 10 espèces de poissons téléostéens pour un total de 70 espèces. La motivation derrière le développement de SCORPiOs était de permettre de mieux comprendre l'impact fonctionnel des duplications complètes de génome. J'ai donc analysé les fonctions des gènes ayant suivi différentes trajectoires évolutives après la duplication. J'ai comparé les signaux fonctionnels dérivés de ces analyses en utilisant les données avant et après correction des arbres, ce qui a permis de confirmer l'utilité de SCORPiOs dans l'étude des duplications complètes de génome. Les résultats détaillés de ces analyses sont présentés dans l'article publié. Ici, je les introduis

brièvement et je les confronte aux connaissances qui nous viennent d'autres duplications complètes.

Patrons d'expression des gènes dupliqués

Comme introduit dans le chapitre précédent, les duplications complètes de génomes introduisent des biais fonctionnels dans le répertoire de gènes des espèces dupliquées. Ces biais s'expriment différemment à court et long-terme et dépendent donc de l'âge des duplications complètes étudiées.

Le patron d'expression d'un gène représente un premier proxy pour caractériser sa fonction. Dans le manuscrit, j'ai comparé trois catégories de gènes : les ohnologues systématiques (gènes restés en deux copies dans toutes les espèces), les ohnologues facultatifs (gènes en deux copies dans au moins une espèce mais pas toutes) et enfin les simples copies systématiques (gènes retrouvés en une seule copie dans toutes les espèces). J'ai ensuite utilisé le patron d'expression des gènes du poisson-zèbre retrouvés dans chacune de ces catégories pour étudier les biais fonctionnels.

La première observation, au regard des différences globales de niveau d'expression, est que les ohnologues sont exprimés à un niveau plus bas et sont plus tissu-spécifiques que les gènes en simple copie. Ces résultats sont accord avec les observations faites chez les plantes (DE SMET et al. 2013) et les mammifères après les 1R-2R vertébrés (GUSCHANSKI, WARNEFORS et KAESSMANN 2017). Dans le manuscrit, nous expliquons ce résultat par l'occurrence de subfonctionnalisation pour au moins une partie des ohnologues. On ne sait pas vraiment si la néofonctionnalisation ou la subfonctionnalisation est plus fréquente après la duplication 3R, même si certains résultats suggèrent que la néofonctionnalisation serait plus fréquente (BRAASCH et al. 2016 ; LIEN et al. 2016 ; SANDVE, ROHLFS et HVIDSTEN 2018). Dans tous les cas, il suffit d'une fraction de gènes subfonctionnalisés pour abaisser le niveau d'expression moyen des ohnologues et retrouver les patrons observés.

En revanche, ces résultats sont en contradiction avec le modèle COSTEX, proposé pour expliquer la rétention des gènes après duplication complète chez les paramécies (GOUT et al. 2010). Dans ce modèle, les gènes à forte expression sont plus souvent maintenus en deux copies, parce qu'ils correspondent à des gènes sous forte sélection négative et à l'expression très contrainte. Leur pseudogénéisation, à travers une baisse du niveau d'expression et/ou une accumulation mutations dans la séquence codante, serait souvent délétère. Ce modèle explique également les biais de rétention de gènes observés chez les levures (SEOIGHE et WOLFE 1999 ; GOUT et al. 2010).

Je propose plusieurs hypothèses pour concilier ces résultats. Premièrement, Gout et al., ont montré un effet de l'âge de la duplication : pour les duplications les plus anciennes,

le biais vers une expression plus forte des gènes dupliqués est masqué, suggérant que les contraintes sur leur dosage peuvent être résolues sur le long terme. Or, les duplications complètes des paramécies et levures sont beaucoup plus récentes que la duplication des poissons téléostéens (100 contre 320 millions d'années, VAN DE PEER, MAERE et MEYER 2009). Gout et al. soulignent également que les contraintes sur le niveau d'expression sont plus fortes chez les micro-organismes comme les paramécies que chez les animaux. De plus, on peut proposer que la subfonctionnalisation jouerait un rôle moins important dans un organisme unicellulaire, puisque l'expression ne peut pas être partitionnée par tissus.

Contribution des gènes dupliqués à l'évolution des tissus

Le second résultat concerne la contribution des différentes catégories de gènes dupliqués à l'évolution des tissus. Le résultat principal est que les ohnologues contribuent davantage à l'expression spécifique aux tissus du cœur et du cerveau. L'expression dans le cerveau a déjà été montrée comme un facteur expliquant la rétention des gènes dupliqués chez les Vertébrés (1R-2R, Figure 2.6, GUSCHANSKI, WARNEFORS et KAESSMANN 2017), Téléostes (ROUX, LIU et ROBINSON-RECHAVI 2017) et Salmonidés (VARADHARAJAN et al. 2018). En effet, Roux et al. ont montré que les gènes exprimés dans le cerveau étaient sous forte sélection purifiante et évoluaient plus lentement que les gènes exprimés dans d'autres tissus (ROUX, LIU et ROBINSON-RECHAVI 2017). Ils proposent que ces gènes seraient également retenus selon le modèle des "gènes dangereux" : leur pseudogénéisation serait ralentie par le coût mutationnel qu'elle implique. De plus, le biais de rétention des ohnologues dans le cœur a également été observé chez l'humain (Figure 2.6).

Enfin, un dernier résultat non-discuté dans l'article est que les gènes exprimés dans les tissus ovariens sont sous-représentés dans chacune des catégories. En effet, ces gènes sont faiblement retrouvés dans les familles de gènes qui descendent de la duplication complètes : ce sont pour la plupart de gènes lignée-spécifique du poisson-zèbre. Ce résultat peut s'expliquer soit parce que la divergence rapide de séquence de ces gènes masque leur origine évolutive soit parce qu'ils sont effectivement apparus récemment dans le génome du poisson-zèbre. La contribution des gènes récemment dupliqués au tissu ovarien a déjà été mis en évidence chez les Vertébrés (Figure 2.6).

Gènes dupliqués et innovations évolutives

Afin de préciser les biais fonctionnels révélés par les patrons d'expression des gènes dupliqués j'ai caractérisé les termes Gene Ontology et voies de signalisations KEGG enrichies dans les différentes catégories de gènes. Cela a permis de mieux caractériser la contribution des gènes dupliqués à l'évolution des poissons téléostéens.

En particulier, deux signaux marquants sont apparus : les ohnologues systématiques et

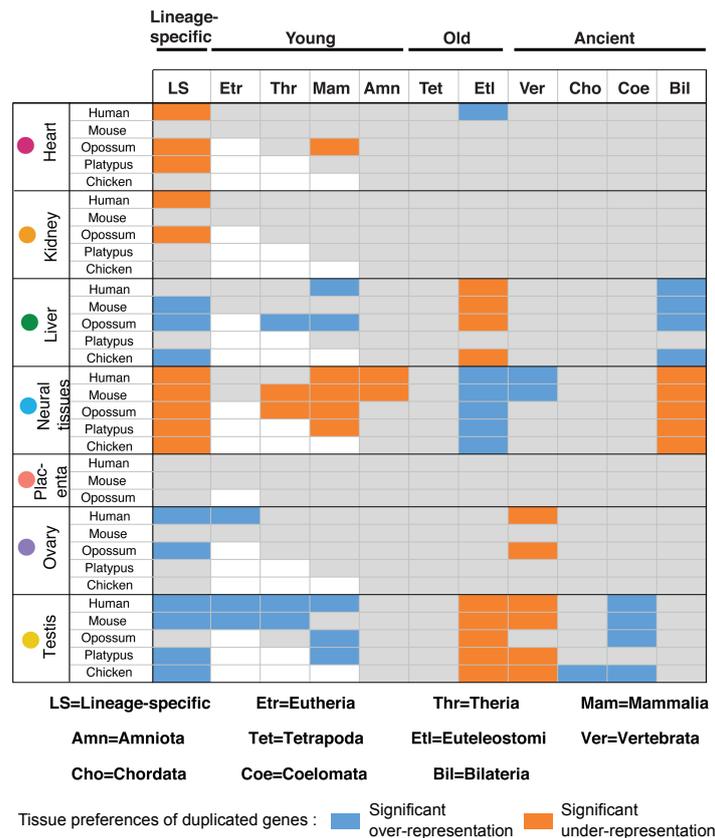


FIGURE 2.6 – Contribution des gènes dupliqués de différents âges à l'évolution des tissus chez les Mammifères. Les paralogues de différents âges présentent des préférences d'expression tissus-spécifiques distinctes. Notamment, les gènes exprimés dans les tissus neuronaux et cardiaques sont sur-représentés parmi les gènes datant de la duplication des Vertébrés (Euteleostomi, 1R-2R). Figure adaptée de GUSCHANSKI, WARNEFORS et KAESSMANN 2017

facultatifs sont enrichis en termes liés à la physiologie de la rétine ainsi qu'au système circulatoire. Or, ces enrichissements peuvent être mis en relation avec deux innovations évolutives documentées chez les téléostéens et coïncidant avec la 3R : une organisation géométrique extrêmement précise des cellules photoréceptrices de la rétine et un nouvel organe, le « bulbus arteriosus » qui sophistique le cœur des poissons téléostéens (Figure 2.7).

Le bulbus arteriosus (BA) est un organe situé en sortie du ventricule cardiaque des poissons téléostéens. Constitué de muscle lisse et non de muscle cardiaque, sa structure élastique agit comme un condensateur qui stabilise la pression artérielle. MORIYAMA et al. 2016 ont caractérisé expérimentalement une paire de gènes dupliqués (*elastin a* et *elastin b*) impliqués dans l'acquisition du bulbus arteriosus. La néofonctionnalisation de *elastin b* à travers l'acquisition de son expression localisée au niveau du bulbus arteriosus est nécessaire à la formation correcte du bulbus arteriosus : elle est impliquée dans la migration des cellules précurseurs et leur détermination en cellules du BA. Nos résultats pourraient suggérer un plus grand nombre de gènes dupliqués potentiellement impliqués dans le fonctionnement de ce nouvel organe.

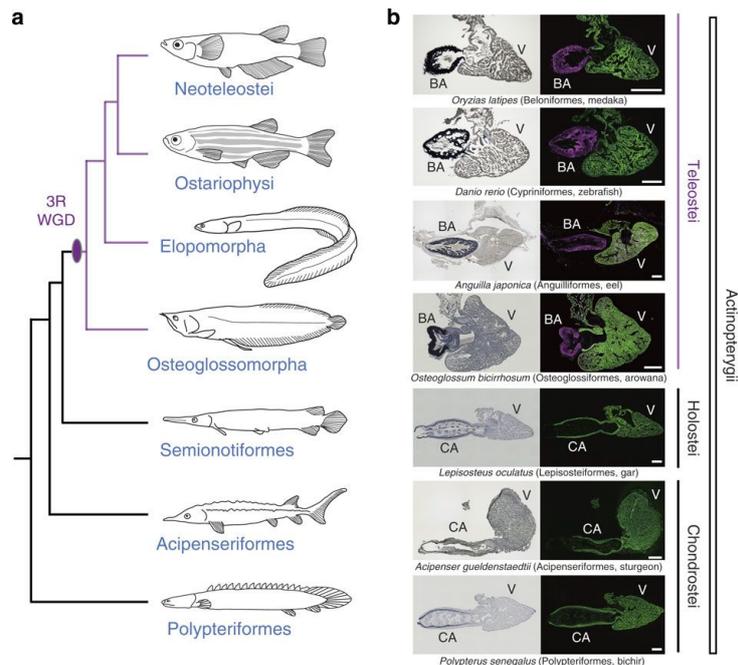


FIGURE 2.7 – Le bulbus arteriosus : une innovation évolutive dans le cœur des poissons téléostéens. a. Position phylogénétique de l’acquisition du bulbus arteriosus sur l’arbre des Actinoptérygiens. b. Anatomie du cœur dans les grands groupes représentés. Le bulbus arteriosus est représenté en violet. Figure tirée de MORIYAMA et al. 2016.

En résumé, la contribution de SCORPiOs à la littérature est double : il complète l’arsenal des méthodes de reconstruction d’arbre de gènes et montre que lier duplications de gènes et innovation évolutive passe par l’amélioration des arbres de gènes. En effet, nous avons montré que plusieurs enrichissements fonctionnels étaient masqués par les erreurs dans les arbres.

2.2 Article

Contribution au travail publié

J’ai entièrement implémenté SCORPiOs et réalisé toutes les analyses présentées dans le papier. Seule la préparation de certaines données a été effectuée par des co-auteurs. J’ai également écrit la première version du manuscrit et contribué à sa remise en forme pour aboutir à la version publiée.

Synteny-Guided Resolution of Gene Trees Clarifies the Functional Impact of Whole-Genome Duplications

Elise Parey ¹, Alexandra Louis,¹ Cédric Cabau,² Yann Guiguen ³, Hugues Roest Crolius ^{*},¹ and Camille Berthelot ^{*},¹

¹Institut de Biologie de l'École Normale Supérieure (IBENS), École Normale Supérieure, CNRS, INSERM, Université PSL, Paris, France

²SIGENAE, GenPhySE, Université de Toulouse, INRAE, ENVT, Castanet Tolosan, France

³INRAE, LPGP, Rennes, France

*Corresponding authors: E-mails: camille.berthelot@bio.ens.psl.eu; hrc@bio.ens.psl.eu.

Associate editor: Koichiro Tamura

Abstract

Whole-genome duplications (WGDs) have major impacts on the evolution of species, as they produce new gene copies contributing substantially to adaptation, isolation, phenotypic robustness, and evolvability. They result in large, complex gene families with recurrent gene losses in descendant species that sequence-based phylogenetic methods fail to reconstruct accurately. As a result, orthologs and paralogs are difficult to identify reliably in WGD-descended species, which hinders the exploration of functional consequences of WGDs. Here, we present Synteny-guided CORrection of Paralogs and Orthologies (SCORPiOs), a novel method to reconstruct gene phylogenies in the context of a known WGD event. WGDs generate large duplicated syntenic regions, which SCORPiOs systematically leverages as a complement to sequence evolution to infer the evolutionary history of genes. We applied SCORPiOs to the 320-My-old WGD at the origin of teleost fish. We find that almost one in four teleost gene phylogenies in the Ensembl database (3,394) are inconsistent with their syntenic contexts. For 70% of these gene families (2,387), we were able to propose an improved phylogenetic tree consistent with both the molecular substitution distances and the local syntenic information. We show that these synteny-guided phylogenies are more congruent with the species tree, with sequence evolution and with expected expression conservation patterns than those produced by state-of-the-art methods. Finally, we show that synteny-guided gene trees emphasize contributions of WGD paralogs to evolutionary innovations in the teleost clade.

Key words: whole-genome duplications, synteny, gene phylogeny, molecular evolution.

Introduction

Whole-genome duplications (WGDs) are dramatic evolutionary events that result in the doubling of a species entire genome. Several ancient WGDs have occurred in land plants, fungi, and animals, in which they represent a major source of functional innovation with long-term impact on the evolution of species (Jaillon et al. 2004; Van de Peer et al. 2017). Genome doubling events have also been uncovered in non-model organisms in recent years (Kenny et al. 2016; Sollars et al. 2017), with more still to be discovered. Although many gene duplicates produced by WGDs are eventually lost, some are retained and thought to provide raw material for evolution to repurpose into new functions (Ohno 1970; Lynch and Conery 2000). For example, 35% of human genes still have ancient duplicates from two WGD events ~550 Ma, which diversified essential multigene families including the MHC (immunity) or the four HOX clusters (anteroposterior development) (Sacerdot et al. 2018; Singh and Isambert 2020). Investigating the fates of duplicate genes in descendant species is crucial to understand how WGDs contribute to

phenotypic robustness, adaptation, and evolvability. This however requires the accurate identification of WGD orthologs, that is, genes descended by speciation after the duplication event, from WGD paralogs, that is, genes descended from different duplicates after the duplication event. WGD paralogs, also called ohnologs, must also be distinguished from paralogs that arose by small-scale duplications or retrotransposition (Hahn 2009).

Orthology and paralogy relationships between genes are generally inferred from a phylogenetic tree. This gene tree represents the most likely evolutionary history from a common ancestral gene based on existing gene sequences. Reconciliation with the species phylogeny then labels gene duplication and speciation events. These phylogeny-reconciled gene trees allow rigorous, model-based tests to examine the evolution of gene sequences and functions. However, gene trees are known to contain errors related to methodological, technical, and biological factors (Som 2015). A prominent source of errors is low phylogenetic signal in sequences, that is, insufficient numbers of substitutions to confidently support one gene tree topology over others

(Rasmussen and Kellis 2007). To address this issue, species-tree-aware methods use proximity to the structure of the species tree, known as the reconciliation cost, to select among statistically equivalent gene tree topologies (Durand et al. 2006; Vilella et al. 2009; Rasmussen and Kellis 2011; Szöllősi et al. 2013; Wu et al. 2013; Scornavacca et al. 2015). Because of computational trade-offs, these methods either rely on heuristic exploration of the gene-tree solution space and result in suboptimal trees, or have an intensive computational cost and are not applicable to large data sets. Critically, these limitations are enhanced in the presence of ancient WGDs. Species descended from a WGD frequently have two paralogs or more per gene family, directly increasing the size of the solution space. Further, reconciliation methods do not account for the acceleration of gene loss rates shortly after the WGD, or for rate heterogeneity across species (Zwaenepoel and Van de Peer 2019).

However, we argue that WGDs also possess underexploited characteristics that can be leveraged to improve gene tree modeling. First, the known timing of well-supported WGD(s) can be integrated as prior knowledge to select or constrain gene tree topologies. Second, WGDs result in specific patterns of genome organization, where pairs of sister regions share duplicated genes in conserved order throughout the genome. These double-conserved syntenic (DCS) regions can be readily uncovered by comparison with unduplicated outgroup genomes (Jaillon et al. 2004; Kellis et al. 2004). Because reconciled gene trees are unreliable, DCS patterns are commonly used to identify WGD-duplicated genes with confidence (Byrne and Wolfe 2005; Catchen et al. 2009; Muffato et al. 2010). In multispecies studies, this evidence has typically been used to exclude thousands of gene families where synteny disagrees with the precomputed tree structure (Kassahn et al. 2009; Berthelot et al. 2014; Braasch et al. 2016), an imperfect solution that reduces statistical power and may lead to technical biases. However, synteny has never been used to systematically select among alternative gene tree topologies in the context of a known WGD event. To the best of our knowledge, only two gene tree-building methods are able to leverage evidence from gene organization to correct orthology and paralogy relationships: SYNERGY, a Neighbor-Joining iterative tree-building algorithm that is no longer available (Wapinski et al. 2007), and ParalogyCorrector, which identifies and corrects gene trees inconsistent with synteny information (Lafond et al. 2013). However, ParalogyCorrector is designed to remove unsupported duplication nodes and recover all true orthologs at the expense of paralogs, which makes it unsuited to WGD studies.

Here, we present Synteny-guided CORrection of Paralogies and Orthologies (SCORPiOs), a synteny-guided gene tree-building algorithm for WGD studies. SCORPiOs builds optimized, species-tree-aware gene trees consistent with known WGD events, local synteny context, and gene sequence evolution. We apply SCORPiOs to the teleost-specific genome duplication (TGD), dated 320 Ma (Jaillon et al. 2004) and find that almost one in four WGD-descended teleost gene trees in Ensembl are incorrect. We propose a corrected,

synteny-consistent tree for 70% of these gene families (2,387 of 3,394 synteny-inconsistent trees). Then, we show that the corrected trees emphasize how duplicate gene retention after the WGD has shaped developmental and signaling pathways involved in known phenotypic innovations in the teleost lineage.

Results

Synteny Is Informative to Describe Gene Phylogenetic Relationships after WGD

After a WGD, gene positions along chromosomes display a particular conservation signature (DCS) that can be leveraged to identify and differentiate orthologous- and paralogous-duplicated regions across species (Jaillon et al. 2004; Kellis et al. 2004). This genomic organization reveals gene trees that are inconsistent with the trees of other syntenic genes in their local neighborhood (fig. 1A). Synteny may therefore be useful to systematically correct erroneous trees by sourcing additional evolutionary information from surrounding genes. To confirm that true WGD orthologs and paralogs can be distinguished based on information from their neighbors, we identified 2,394 reliable WGD gene duplicates in 229 unambiguous DCS regions in zebrafish and medaka covering 23% and 26% of each genome, respectively, using gene homologies from Ensembl as in previous studies (Materials and Methods; Kassahn et al. 2009; Braasch et al. 2016; supplementary table S1, Supplementary Material online). We then tested whether local gene neighborhoods differ between orthologous and paralogous WGD gene copies, in the absence of any correction. In theory, paralogous regions diverge immediately after the WGD event, in particular through massive gene losses, and should be more different in terms of gene retention, loss, and gene sequence evolution across species compared with orthologous genomic segments, which diverge at speciation (fig. 1B; Scannell et al. 2006). We found that orthologs indeed share more orthologous syntenic neighbors and have more similar local gene retention and loss patterns than paralogs, as expected (fig. 1C, Wilcoxon–Mann–Whitney tests, $P < 2.2e-16$). Local syntenic context is therefore informative to distinguish orthologs from paralogs, and can potentially be leveraged to correct erroneous gene trees systematically after a WGD event.

SCORPiOs: Synteny-Guided CORrection of Paralogies and Orthologies in Gene Trees

Integrating sequence and synteny information to build gene phylogenies is challenging, and current methods are not designed to deal with WGDs. To address this issue, we have developed SCORPiOs, an algorithm that improves gene trees in the presence of one or several WGD events, using information from synteny conservation patterns. Because it complements gene sequence evolution with syntenic information, SCORPiOs is suitable to study WGD events where synteny with outgroups species remains detectably conserved, such as the different fish WGDs, the baker's yeast WGD or most plant WGDs (Van de Peer et al. 2017). SCORPiOs is not intended for very ancient events such as the 1R or 2R vertebrate WGDs,

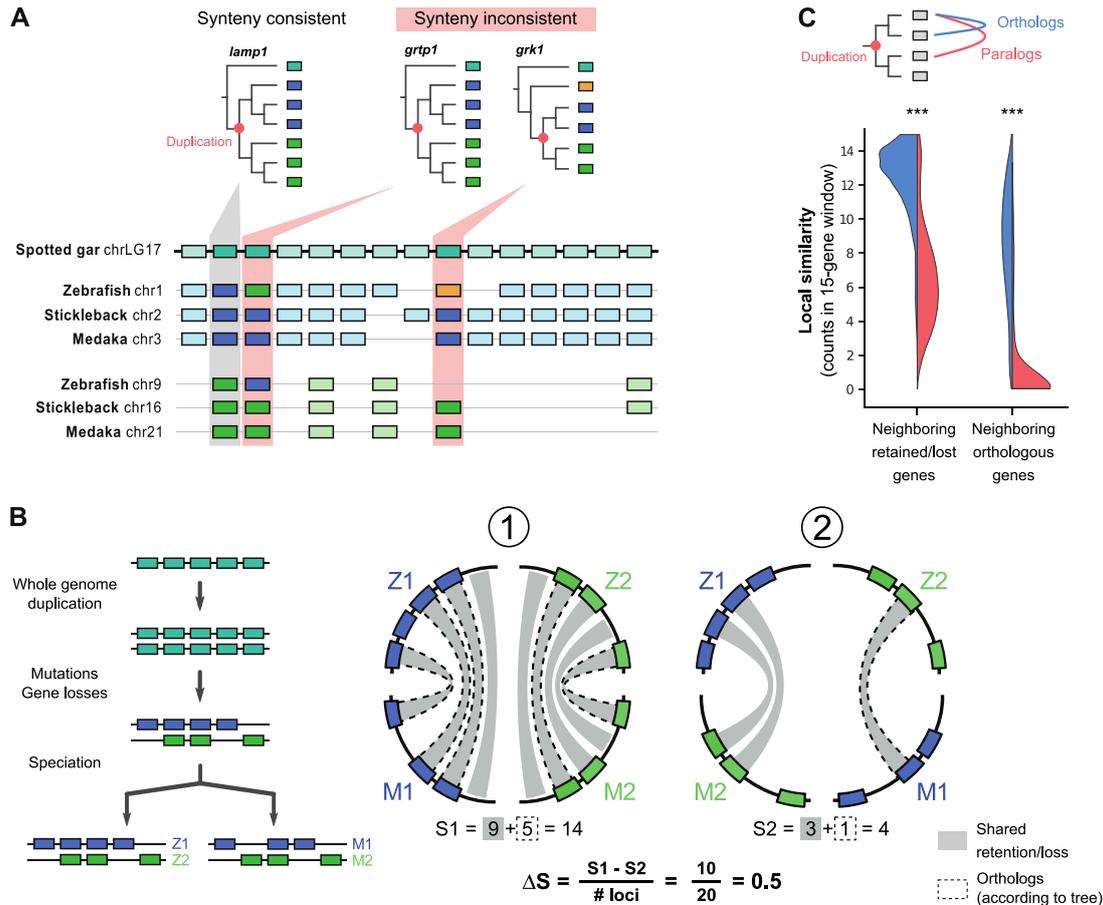


Fig. 1. Local synteny context is informative to sort out gene duplicates after WGD. (A) Gene trees and synteny context of the *lamp1*, *grtp1*, and *grk1* gene families in teleosts and their nonduplicated outgroup (spotted gar). Duplication nodes are labeled in red. Blue and green boxes represent members of the two orthology groups in each original gene tree from Ensembl; yellow denotes genes annotated as nonduplicated. The synteny context identifies gene trees where zebrafish homologs are assigned to the wrong orthology groups (highlighted in red). (B) Schematic evolution of a genomic segment after WGD, with gene colors as in (A). After WGD, the duplicated segments evolve independently and accumulate mutations and gene deletions. Zebrafish and medaka genomic segments are compared under two scenarios: Z1/M1 and Z2/M2 (scenario 1), which corresponds to true orthologs, and Z1/M2 and Z2/M1 (scenario 2), corresponding to paralogs. Under each scenario, genes similarly retained or lost across species are represented by gray links, and genes annotated as orthologs are linked by dotted lines. True orthologous segments (1) share more similar patterns of gene retentions/losses and orthologies than paralogous segments (2), resulting in a high four-way similarity score ΔS . (C) Local syntenic context can differentiate orthologs from paralogs. Distributions of the number of shared gene retentions/losses and annotated orthologs in a 15-gene window around zebrafish/medaka orthologs (in blue) and paralogs (in red), based on the original gene trees from Ensembl. Wilcoxon–Mann–Whitney tests, $n = 2,394$, $***P < 0.001$.

where synteny with outgroups is highly degraded. SCORPiOs is coded in Python 3, implemented as a snakemake workflow (Köster and Rahmann 2012), and takes as inputs: the full set of 1) phylogenetic gene trees and 2) gene positions from extant, WGD-derived genomes along with one or several unduplicated outgroups; 3) the corresponding gene sequence alignments; and 4) the species phylogeny with the putative WGD position(s). For convenience, SCORPiOs also includes an implementation of TreeBeST (Vilella et al. 2009) and can initialize the process from sequence alignments if precomputed gene trees are not available. The major steps of SCORPiOs are outlined in figure 2 (see supplementary fig. S1, Supplementary Material online, for a detailed flowchart):

The implementation details of these different steps in SCORPiOs are developed below.

Comprehensive Identification of Homologous Families after the WGD Event

To identify duplicated regions within a genome, SCORPiOs takes advantage of patterns of DCS with a nonduplicated outgroup. After the WGD, genomic regions of the nonduplicated outgroup give rise to two orthologous genomic segments in each duplicated ingroup species (fig. 1B). However, genomic rearrangements, gene losses or duplications, and errors in the initial gene trees can obscure these 2:1 orthologous relationships in modern genomes. Most WGD studies thus restrain themselves to genomic regions where the WGD event is evident either from the gene trees or from highly conserved DCS patterns (Scannell et al. 2006; Kassahn et al. 2009; Berthelot et al. 2014; Inoue et al. 2015; Braasch et al. 2016). As a consequence, significant subsets of gene families are de facto discarded from analysis (54% not considered in

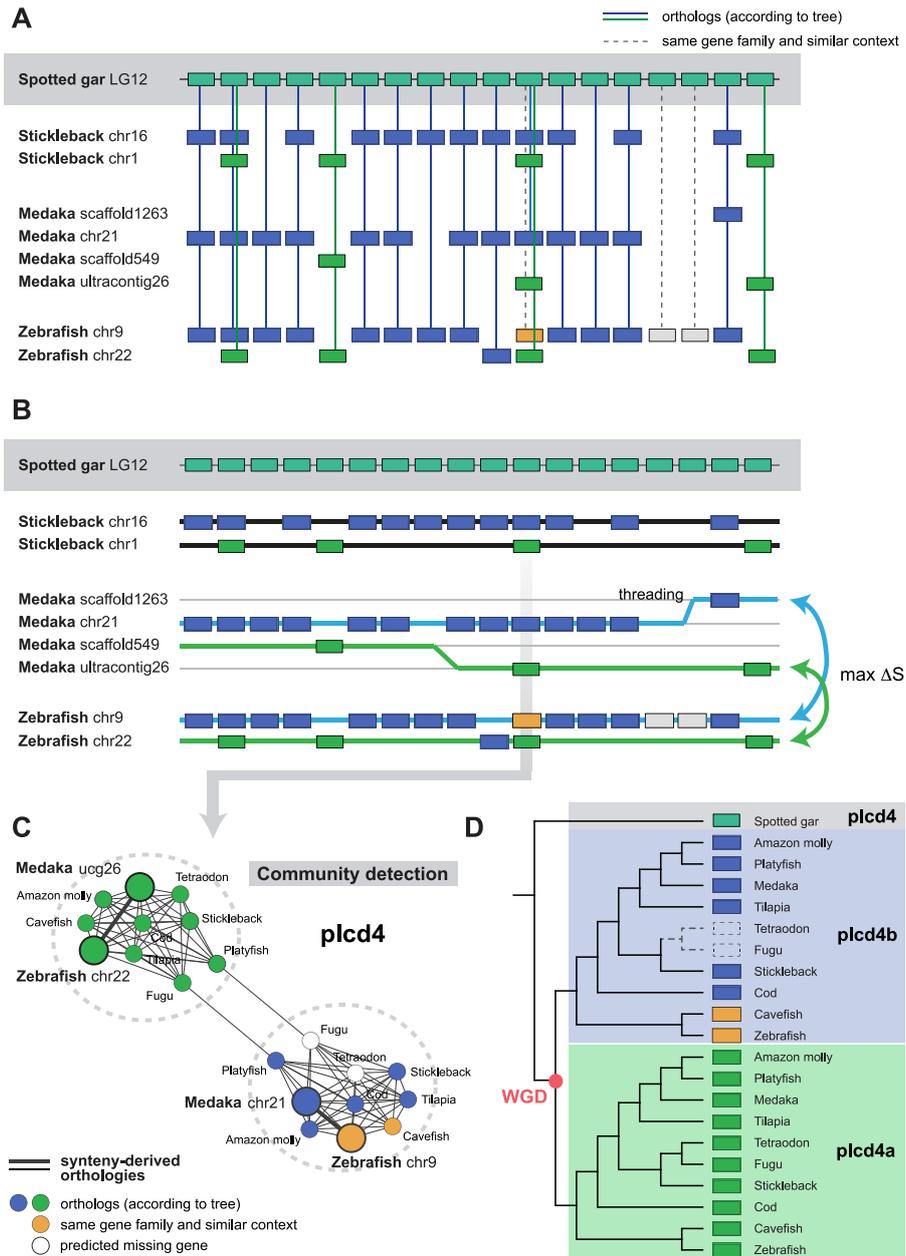


Fig. 2. Overview of the SCORPIOs workflow. (A) For every region in the reference nonduplicated genome, SCORPIOs identifies all potential orthologs in the duplicated genomes (orthologs and syntenic homologs) from the original gene trees. Here, a section of the spotted gar linkage group 12, and genomic locations of its potential orthologous genes in three TGD teleosts (out of ten). For each gene family, color represents genes identified as orthologs in the original gene trees from Ensembl. (B) SCORPIOs maximizes the syntenic context similarity ΔS to thread and pair orthologous genomic segments between duplicated species (see [supplementary note 2](#), [Supplementary Material](#) online). Green and blue threads show the highest scoring configuration for the medaka/zebrafish comparison. (C) For each individual gene, SCORPIOs integrates orthology links derived from the local synteny context across all species into a gene graph, where orthologous communities are identified. Here, the *plcd4* gene, whose genomic position is highlighted in (B). Nodes represent homologs of the *plcd4* gene, colored as in (A) and (B), where the medaka and zebrafish genes are highlighted as large circles. Links represent orthology relationships deduced from the similarity in syntenic contexts, with orthologies between zebrafish and medaka highlighted with a thicker line. SCORPIOs explicitly models nodes corresponding to missing homologs in the gene graphs (in white). Ucg, ultracontig. (D) If the original gene tree is not consistent with the syntenic-derived orthologous gene communities, SCORPIOs proposes a corrected gene tree. Here, the corrected gene tree for the *plcd4* gene family, which includes zebrafish and cavetfish homologs (in yellow) as part of the *plcd4b* subclade based on their genomic location (as seen in B).

Inoue et al.; 43% in [Braasch et al. 2016](#)). Omitting such genes from analysis because of incorrect trees or genomic location is highly problematic, and may result in patterns reflecting technical bias rather than biological insight.

SCORPIOs addresses this problem by first identifying all potentially WGD-descended gene duplicates in the initial gene trees. SCORPIOs uses a loose definition of orthology to build a comprehensive list of potential orthologs between

each gene in the nonduplicated reference genome and all duplicated genomes (fig. 2A and supplementary note 1 and fig. S2, Supplementary Material online). This table retains all gene copies since the ingroup/outgroup speciation node, as well as any other homologs from the same gene tree with a loosely similar syntenic context. Unlike other methods, SCORPiOs therefore first strives to define a comprehensive gene set from each gene family to be searched for potential orthology and paralogy relationships, instead of pruning all genes that do not belong to predefined duplicated segments.

Identification of Orthologous Genomic Segments between Pairs of Species

As a second step, SCORPiOs reconstructs ancestral WGD-duplicated regions, which may have become eroded by genome evolution processes. SCORPiOs uses the nonduplicated outgroup as a proxy for the ancestral gene order. It scans the outgroup genome with a sliding window of user-defined length, and uses the comprehensive list of orthologs to identify and thread the ancestral-duplicated segments in each descendant species (fig. 2B and supplementary note 2, Supplementary Material online). These duplicated genomic segments are then compared between pairs of duplicated species to match orthologous segments. SCORPiOs scores the syntenic similarity between segments using the number of genes annotated as orthologs in the initial gene trees, as well as the pattern of gene retention and losses, as described in figure 1. SCORPiOs is flexible and does not require perfect gene order conservation or contiguity to thread genomic segments, instead attempting to maximize the local similarity between both duplicated species (supplementary note 2, Supplementary Material online). This results in a four-way comparison of threaded, scored genomic segments for every pair of duplicated genomes and every window in the outgroup (fig. 2B).

SCORPiOs interprets the differentiation in four-way similarity scores (ΔS) as a confidence index that two genomic segments can be assigned as orthologs between the two duplicated species. This information is then reflected back to the genes within those segments to identify putative orthologs and paralogs using the window of highest ΔS that they belong to (fig. 1B and supplementary note 2 and fig. S4, Supplementary Material online). Of note, SCORPiOs considers that independent lineage-specific duplications are unlikely to result in duplicated genomic segments across different species that could be confused with the WGD. Paralogs in duplicated genomic regions consistent with the known WGD event are therefore considered as WGD paralogs, to the exclusion of other types of paralogs.

Orthology and Paralogy Constraints across All Species

As the previous step assigns orthologs and paralogs between pairs of species, SCORPiOs then integrates those results to define groups of orthologous genes across all species under study. SCORPiOs exhaustively performs all pairwise comparisons between duplicated genomes as described above, and constructs an orthology graph for every gene family (fig. 2C). SCORPiOs orthology graphs are unweighted graphs where

two nodes (genes) are linked by an edge when they are inferred as orthologs based on their local synteny similarity. The graphs also include “placeholder” nodes, which correspond to gene copies that have been lost since the WGD but whose existence can be inferred from the synteny pattern (supplementary fig. S5, Supplementary Material online). In each graph, we then expect to find two similar-sized orthologous gene communities that are derived from the WGD. If synteny was perfectly informative and the process was error-free, we would expect these communities to be two independent, fully connected cliques. In practice, we observe that some graphs do not result in two isolated communities due to inconsistencies in the orthology assignments. SCORPiOs then uses the Girvan–Newman algorithm (GN), a community detection algorithm that iteratively removes the most central edges of each graph to separate nodes in two communities (Girvan and Newman 2002). In cases where GN is unable to separate genes of a duplicated species in two communities, we apply the Kernighan–Lin algorithm (KL) (Kernighan and Lin 1970). KL is a heuristic aimed at finding two communities of similar sizes, that require cutting the fewest edges. If the KL solution separates the duplicated genes from the same species more frequently, it is preferred over the GN solution. At the end of this step, SCORPiOs has identified the two groups of orthologous genes across extant species that each descend from a single ancestral gene in the common WGD ancestor, derived from local synteny information. Of note, one of these orthology groups can be entirely made of placeholders, corresponding to the loss of one duplicate gene in all species (typically an early loss before the first speciation, supplementary fig. S5, Supplementary Material online).

Gene Tree Correction and Test

Next, SCORPiOs identifies and attempts to correct gene trees that do not fulfill the orthologies and paralogies from their synteny-derived orthology graph. First, SCORPiOs checks whether each gene tree contains a node from which the unduplicated outgroup gene diverges, followed by either 1) a node corresponding to the WGD, under which each subtree encompasses one of the two orthology communities from the orthology graph, or 2) a single subtree, in cases where the same paralog is missing in all species (i.e., placeholder nodes subgraph; supplementary fig. S5, Supplementary Material online). When this topological constraint is not verified in the original gene tree (supplementary fig. S6, Supplementary Material online), SCORPiOs proposes a corrected, fully resolved gene tree that is both species tree and synteny-consistent. We explore the constrained topologies solution space using the fast distance-based approach implemented by the program ProfileNJ (Noutahi et al. 2016). Briefly, ProfileNJ independently resolves each multifurcated orthology group, so as to minimize the reconciliation cost with the species tree. SCORPiOs then compares the likelihoods of the original and corrected tree and accepts the correction if both trees are statistically equivalent (approximately unbiased test, Shimodaira and Hasegawa 2001; branch lengths fitted with PhyML, Guindon et al. 2010; Materials and Methods). If

ProfileNJ fails to find an adequate solution, SCORPiOs uses the TreeBeST-modified version of PhyML, which fits the gene sequences in each post-WGD subtree using maximum likelihood (ML) optimization while accounting for the species phylogeny topology (Vilella et al. 2009), and compares the corrected tree to the original one as above. We correct gene tree topologies only when the sequence similarity data are explained equally well (or better) by the synteny-aware tree: our correction approach is conservative, in the sense that it gives precedence to the likelihood of the molecular evolution model. Note also that SCORPiOs does not attempt to build a corrected tree for very large multigenic families (>1.5 genes per species in one or both post-WGD communities), as these are often overaggregated families in input trees that SCORPiOs cannot solve. As a final step of the pipeline, SCORPiOs can reinsert the corrected subtrees into the original gene tree depicting the evolution of the whole-gene family, and recompute branch lengths (supplementary note 3 and figs. S7–S9, Supplementary Material online). Additional available options when executing SCORPiOs are described in supplementary note 4, Supplementary Material online.

SCORPiOs Corrects a Significant Fraction of Neopterygii Subtrees in Ensembl

We applied SCORPiOs to gene trees from the Ensembl Compara database, version 89 (Vilella et al. 2009), which includes 70 vertebrate genomes, of which eleven are Neopterygii: ten duplicated teleost genomes, and the spotted gar as an outgroup to the TGD (supplementary fig. S10, Supplementary Material online). This set represents a total of 21,431 gene trees reconciled with the vertebrate species tree. About 50% of genes are syntenic between spotted gar and teleost genomes, which makes it a suitable outgroup and proxy for the ancestral gene order (Materials and Methods; Braasch et al. 2016).

We ran SCORPiOs in iterative mode, which corrects the gene trees the first time, thus improving the quality of the synteny information, and then again until convergence (window size: 15 genes, with a step of 1 gene). For this data set, two iterations were sufficient to reach convergence (see supplementary tables S2–S4, Supplementary Material online, for detailed results by iteration). In brief, in iteration 1, we identified 15,476 Neopterygii loose orthology groups within the 21,431 gene trees. SCORPiOs produced 14,576 orthology graphs, of which 10,172 (70%) were already subdivided into two isolated orthology cliques. The vast majority of families not producing graphs (86%) are small families (<3 genes), for which we do not build graphs, as they would result in the same tree topology (supplementary tables S2–S3, Supplementary Material online). For 3,394 gene trees (22%), the orthologies and paralogies relationships were in disagreement with the synteny graph. In total, at the end of the two iterations, we were able to find a synteny-aware and statistically equivalent gene tree for 2,387 of these genes families, thus correcting 70% of all synteny-inconsistent gene trees (15% of Ensembl Neopterygii subtrees). Interestingly, for 672 of the gene trees that we correct (28%), the new tree is better supported by the sequences and results in a

significantly higher likelihood (AU test, $\alpha = 0.05$). This may reflect difficulty of the Ensembl pipeline to explore the tree topology space due to the high number of species in the database, as previously described (Wu et al. 2013). Furthermore, we also applied SCORPiOs to version 94 of the Ensembl Compara database, containing 47 teleost genomes and a total of 43,491 gene trees (supplementary fig. S11, Supplementary Material online). With this increase in phylogenetic coverage, a higher proportion of gene trees was inconsistent with synteny (7,168 synteny-inconsistent gene trees for 15,760 gene families, 45%), and could be improved by SCORPiOs (3,681 corrected gene trees, 23%). Again, this result is in line with the idea that exploration of the topology space becomes limited for larger trees, inducing more errors. Altogether, we demonstrate that insight from synteny can improve a significant fraction of gene trees in the presence of many gene duplicates after a WGD event.

Validation of SCORPiOs Trees and Comparison to Existing Methodologies

SCORPiOs corrects gene trees by providing synteny-derived orthology and paralogy constraints to existing tree-building programs, ProfileNJ and TreeBeST PhyML. To evaluate how SCORPiOs improves over the state of the art, we compared the 2,387 corrected teleost gene trees to alternatives topologies obtained with other methodologies. First, we used RAXML (Stamatakis 2014) to build a pure ML tree for each gene family, as a standard to compare against the likelihood of other trees. Second, we included the original, uncorrected trees from Ensembl Compara (Vilella et al. 2009), which are phylogeny-reconciled consensus trees build with the TreeBeST pipeline. Last, we tested ParalogyCorrector, the only other existing methodology that corrects gene trees using synteny information (Lafond et al. 2013). To compute ParalogyCorrector trees, we provided our synteny-derived orthologies constraints along with the original Ensembl subtree. ParalogyCorrector then finds a new gene tree minimizing the differences to the original tree while satisfying the orthology constraints (but not necessarily paralogy), whereas SCORPiOs fully recomputes each paralog subtree under the corrected duplication node.

We first evaluated whether each tree is well supported by the sequence alignment (AU test, $\alpha = 0.05$; fig. 3A). As expected, RAXML solutions are systematically the best fit to the sequence data. However, SCORPiOs is able to find a solution as good as the pure ML fit for 61% of gene trees, which is significantly better than the original Ensembl trees (47%; proportion test, $P < 2.2e-16$) or ParalogyCorrector (51%; proportion test, $P < 2.2e-16$). Additionally, the SCORPiOs trees are a better fit to the gene sequence information than either the Ensembl or the ParalogyCorrector trees for 9.5% of the 2,387 test gene families.

We then assessed whether the gene trees are concordant with the known species phylogeny. For each software, we identified the species-tree inconsistent nodes, previously referred to as “dubious nodes” or “nonapparent duplication nodes” (Vilella et al. 2009; Lafond et al. 2014; Materials and Methods). We find that SCORPiOs outperforms both RAXML

and the original Ensembl trees in terms of species-tree congruence (Wilcoxon signed rank test, P values $< 2.2e-16$, fig. 3B). Both SCORPiOs and ParalogyCorrector produce trees that are highly congruent to the species phylogeny, with a slight advantage to ParalogyCorrector (Wilcoxon signed rank test, $P < 2.2e-16$, 88.6% and 99.7% of fully congruent trees, respectively; proportion test $P < 2.2e-16$). However, the noncongruent solution from SCORPiOs was better supported by sequence information for 146 of 269 gene families for which ParalogyCorrector reported a fully congruent solution, suggesting that the additional duplication nodes inferred by SCORPiOs are largely correct.

Lastly, we assessed whether gene trees inferred by all four methods are parsimonious by calculating their total number of duplication nodes. We found that SCORPiOs infers the

lowest number of duplications in gene trees, compared with all other methods (Wilcoxon signed rank test, P values $< 2.2e-16$, fig. 3C). By design, SCORPiOs explicitly replaces the WGD node at its known location in the tree when both gene copies survive in modern genomes (fig. 3D), resulting in fewer inferred duplications downstream of the WGD node. In contrast, ParalogyCorrector places duplications closer to the leaves and wrongly identifies most WGD duplicates as species-specific duplications. For 38% of Ensembl gene trees, the duplication is placed at the Neopterygii node and incorrectly encompasses the spotted gar gene (nonduplicated outgroup). In conclusion, SCORPiOs significantly improves gene trees in an evolutionary context where other methods struggle, due to the high number of gene copies and nonparsimonious tree structures created by WGD events.

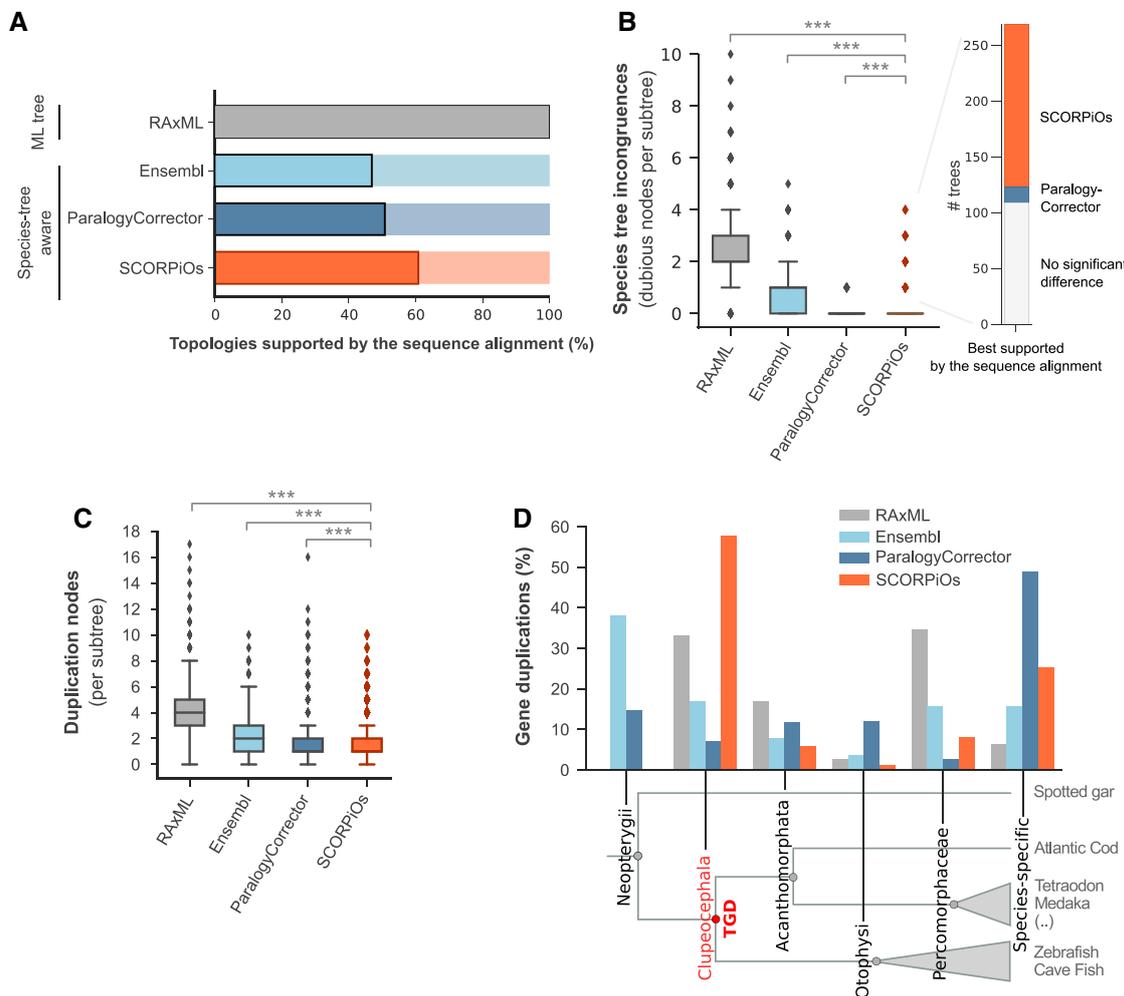


FIG. 3. SCORPiOs performs better than state-of-the-art methods on WGD gene trees. (A) SCORPiOs produces gene trees with higher phylogenetic support than either ParalogyCorrector or the Ensembl Compara pipeline. Bars represent the fraction of trees where the topology proposed by either of the three species-tree aware methods is not significantly less likely than the maximum-likelihood, nonreconciled solution from RAxML (AU tests, $P \geq 0.05$). (B) SCORPiOs and ParalogyCorrector both produce fewer species tree inconsistencies than RAxML or Ensembl Compara in teleost gene subtrees. Boxplots represent the distributions of dubious, low-support duplication nodes per tree; dots represent the top 5% outlier trees. Right inset, stacked barplot: when SCORPiOs introduces more dubious duplication nodes than ParalogyCorrector, we compared the likelihood of both solutions according to sequence evolution and found that the solution from SCORPiOs is equally or more likely for 255 out of 269 gene families. (C) Number of duplication nodes for each method. (D) Position of duplication nodes in the species tree for each method. The Clupeocephala ancestor corresponding to the expected TGD node is highlighted in red. Intermediate ancestors with fewer than 5% of duplications for all methods are not displayed.

SCORPiOs Corrections Improve Correlation of Orthologous Gene Expression

In addition to their consistency with sequence evolution and phylogeny, we evaluated whether orthologs corrected by SCORPiOs are functionally more similar than paralogs, as previously reported (Koonin 2005; Altenhoff et al. 2012; Chen and Zhang 2012). We used gene expression data from 11 tissues in zebrafish and medaka (Pasquier et al. 2016) to investigate whether SCORPiOs-corrected zebrafish/medaka orthologs display higher similarity in expression patterns (Materials and Methods). For instance, SCORPiOs modified the orthology relationships in the *cxcl12* gene family for zebrafish and medaka (fig. 4A), grouping together medaka *cxcl12a* and zebrafish *cxcl12b* in one orthology group and medaka *cxcl12b* and zebrafish *cxcl12a* in the other based on their local syntenic surroundings. Indeed, gene expression patterns support the orthology reassignment, with one gene copy expressed in bones, brain, embryo, gills, and liver, and the other expressed at high level, predominantly in kidney, in both species.

Overall, SCORPiOs modified the orthology relationships between zebrafish and medaka for 761 gene families. Of these, 210 correspond to orthology/paralogy reassignments (as in fig. 4A), whereas the remaining 551 correspond to removal of errors or addition of new homology relationships. The correction increased the number of 1-to-1 orthologs between zebrafish and medaka (13,463 vs. 14,008) as well as the total number of genes with an ortholog in the other species (16,150 zebrafish and 15,543 medaka genes with an ortholog before correction, vs. 16,316 and 15,748 after). For the 210 gene families where medaka and zebrafish orthologies were reassigned, we find that orthologs are expressed at closer average levels after correction (Wilcoxon signed rank test, P value = 0.0171, fig. 4B) and also significantly more correlated across tissues (Wilcoxon signed rank test, P value = 0.0050, fig. 4C). These results support that SCORPiOs measurably improves orthology and paralogy relationships. Additionally, they suggest that erroneous orthology relationships may obfuscate functional investigation of gene evolution after genome duplication, especially when their effects get compounded over dozens of species and thousands of gene families, as reported above for teleosts.

SCORPiOs Correction Emphasizes WGD Contributions to Evolutionary Innovations in Teleost Fish

Numerous studies have suggested a link between the function of a gene and retention or loss after a WGD. Yet, in the absence of systematic gene tree correction methods, the fate of TGD duplicates has only been investigated in a restricted set of ~6,000 high-confidence teleost gene families (Kassahn et al. 2009; Inoue et al. 2015; Braasch et al. 2016), potentially introducing biases in subsequent conclusions. Here, we used the full set of 21,431 gene trees from Ensembl corrected by SCORPiOs to investigate gene retention across ten teleost species (zebrafish, cavefish, tetraodon, fugu, stickleback, medaka, tilapia, platyfish, amazon molly, and cod). We classified genes into three categories with respect to their fate after the TGD (Materials and Methods). Briefly, we grouped genes

retained in two copies across all ten teleost species (“systematic ohnologs,” $n = 1,828$), genes found in single copy in all species (“singletons,” $n = 13,895$), and genes retained in two copies in at least one teleost species but not in all (“facultative ohnologs,” $n = 7,265$) (supplementary fig. S12, Supplementary Material online). We then used expression levels and functional annotations in zebrafish to explore how gene function relates to evolutionary trajectory after the TGD (Materials and Methods).

Overall, we find that singletons have slightly higher average expression levels and broader expression patterns than both systematic and facultative ohnologs (Wilcoxon–Mann–Whitney test, $P < 0.001$, fig. 5A, supplementary fig. S13, Supplementary Material online, Materials and Methods). This is in line with previous observations on paralogs and may reflect cases of subfunctionalization between the two groups of ohnologs, for which duplicated genes have partitioned the ancestral function, becoming expressed in fewer tissues and/or at lower level (Huminiecki and Wolfe 2004; De Smet et al. 2013; Guschanski et al. 2017). We next investigated whether tissue-specific singletons and ohnologs display preferential expression in different tissues, reflecting different contributions to teleost evolution. We find that tissue-specific systematic ohnologs are overrepresented in brain, heart, and muscle, and depleted in liver, intestine, ovary, and testis, compared with all zebrafish tissue-specific genes ($\tau > 0.9$; hypergeometric tests, corrected $P < 0.05$, fig. 5B). In contrast, tissue-specific singleton genes are overrepresented in liver, kidney, intestine, and testis (fig. 5B). For facultative ohnologs, we observe an enrichment in brain-specific genes, but also in muscle-specific expression (fig. 5B). Interestingly, enrichment in brain- and heart-specific genes, as well as depletion in liver- and testis-specific genes, have been already observed in human ohnologs retained after the 1R and 2R vertebrate WGDs (Guschanski et al. 2017). This result ties in with previous reports that some gene families and functional categories are recurrently amplified by independent WGDs, possibly because they offer adaptive advantages when duplicated en masse (van Hoek and Hogeweg 2009; De Smet and Van de Peer 2012).

Additionally, we investigated whether TGD ohnologs and singletons belong to different biological pathways using Gene Ontology Biological Processes (GO BP) and KEGG pathway enrichment analyses (Materials and Methods). Systematic and facultative ohnologs are enriched in general molecular processes previously found in WGD duplicates, linked to transcriptional regulation and metabolic processes, as well as terms related to the nervous system, consistent with the brain-specific expression patterns reported above (fig. 5C and supplementary tables S5–S8, Supplementary Material online) (Blomme et al. 2006; Inoue et al. 2015; Singh et al. 2015; Li et al. 2016; Pasquier et al. 2017; Roux et al. 2017). In contrast, singletons are enriched in housekeeping functions, with some of the most significant GO terms being “cell cycle,” “nucleic acid metabolic process,” and “cellular localization,” along with the KEGG pathways “Ribosome” and “DNA replication” (fig. 5C and supplementary tables S9 and S10, Supplementary Material online) (De Smet et al. 2013; Li

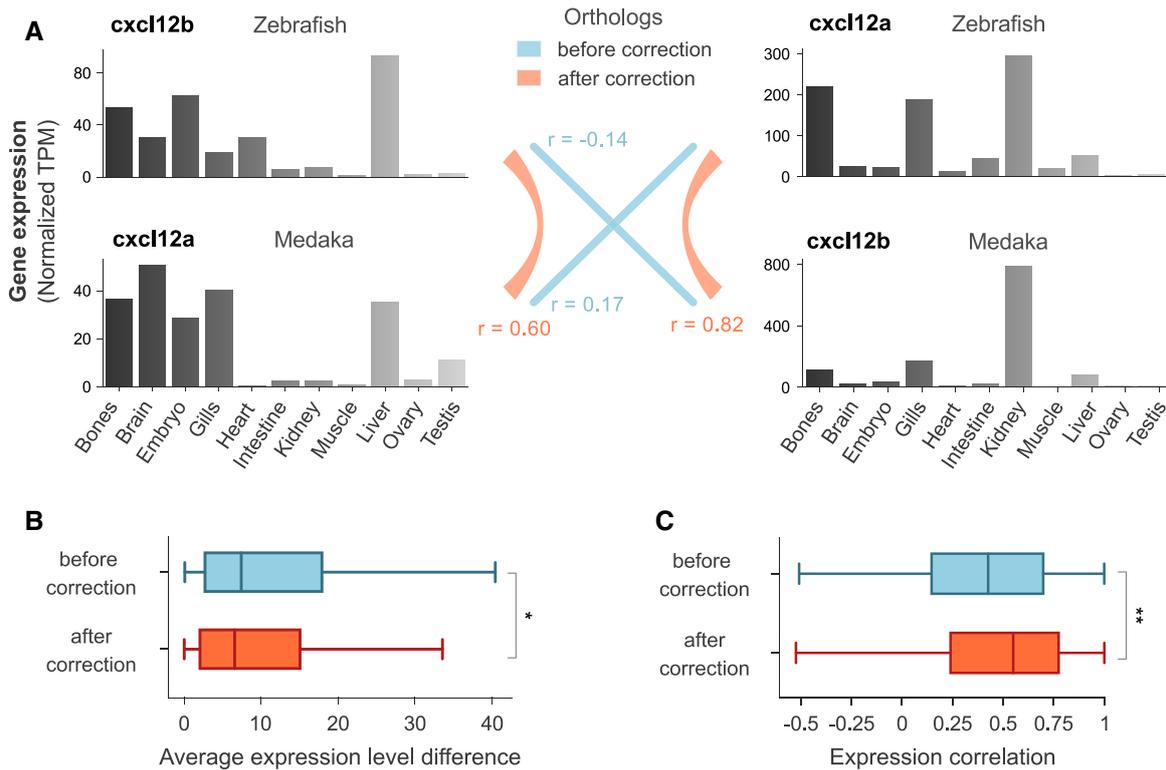


Fig. 4. Functional similarity of orthologous genes after SCORPiOs correction. (A) Expression of zebrafish and medaka *cxcl12* homologs in 11 tissues (quantile-normalized transcripts per million, TPM). Orthology relationships and expression level correlations before and after correction by SCORPiOs are noted in color (r , Pearson correlation coefficient). (B) Difference in average expression levels across 11 tissues between orthologous zebrafish/medaka genes, before and after correction (paired Wilcoxon test, $n = 210$, $*P < 0.05$). (C) Average correlation of expression levels for orthologous zebrafish/medaka genes before and after correction (paired Wilcoxon test, $n = 210$, $**P < 0.01$).

et al. 2016). However, we discover here that systematic duplicates are also enriched in more specific functions, especially related to retina physiology. Interestingly, the teleost WGD coincides with functional innovations in the retina specific to this clade, where photoreceptor cells are organized in a regular pattern described as a “cone mosaic” (Lyll 1957; Engström 1963; Sukeena et al. 2016). These results are mirrored in the KEGG analysis with the overrepresentation of the taurine and hypotaurine metabolism pathway, suggested to have a functional role in the teleost retina (Lima et al. 1998; Omura and Inagaki 2000). Our results therefore support that the amplification of retinal genes during the teleost WGD was important in the acquisition of this evolutionary innovation.

Finally, some functional enrichments become prominent only after gene tree correction with SCORPiOs (in bold on fig. 5). In particular, we observe an enrichment for both systematic and facultative ohnologs toward terms related to the circulatory system (“Adrenergic signaling in cardiomyocytes,” “Vascular smooth muscle contraction,” “VEGF signaling pathway”; fig. 5C, supplementary tables S6 and S8, Supplementary Material online). This extends broader support to a previous report that TGD-derived duplicates, especially those of the *elastin* gene, have led to morphological sophistication of the teleost heart and circulatory system (Moriyama et al. 2016). Lastly, genes in the facultative ohnolog category are enriched in the “Melanogenesis” pathway, also consistent with the expansion of the pigmentation repertoire

after the TGD (Lorin et al. 2018). Taken together, our results suggest a strong contribution of TGD duplicates to functional novelty in this clade, mediated by the fixation of specialized duplicated genes. Importantly, many of these enrichments were fully obscured by errors in the gene evolutionary histories downloaded from as respected a reference database as Ensembl, which is widely sourced for comparative and evolutionary studies (Alföldi and Lindblad-Toh 2013; Herrero et al. 2016). SCORPiOs therefore fulfills its purpose in the arsenal of tree-building tools and has potential to dramatically further investigations into the evolutionary and functional outcomes of WGD events.

Discussion

Gene retention and loss after a WGD is a poorly understood interplay between functional redundancy, increased evolvability, and mutational cost. Several complementary models of gene evolution have been proposed to account for duplicate retention after WGD, including neofunctionalization, where one copy acquires a new function, whereas the other maintains the original one (Ohno 1970; Lynch and Conery 2000); subfunctionalization, where both copies partition the ancestral function between themselves (Force et al. 1999); dosage balance, where subunits of macromolecular complexes are maintained as duplicates to ensure proper stoichiometry between interacting partners (Blomme et al. 2006; Veitia et al. 2008; Makino and McLysaght 2010); and cost of

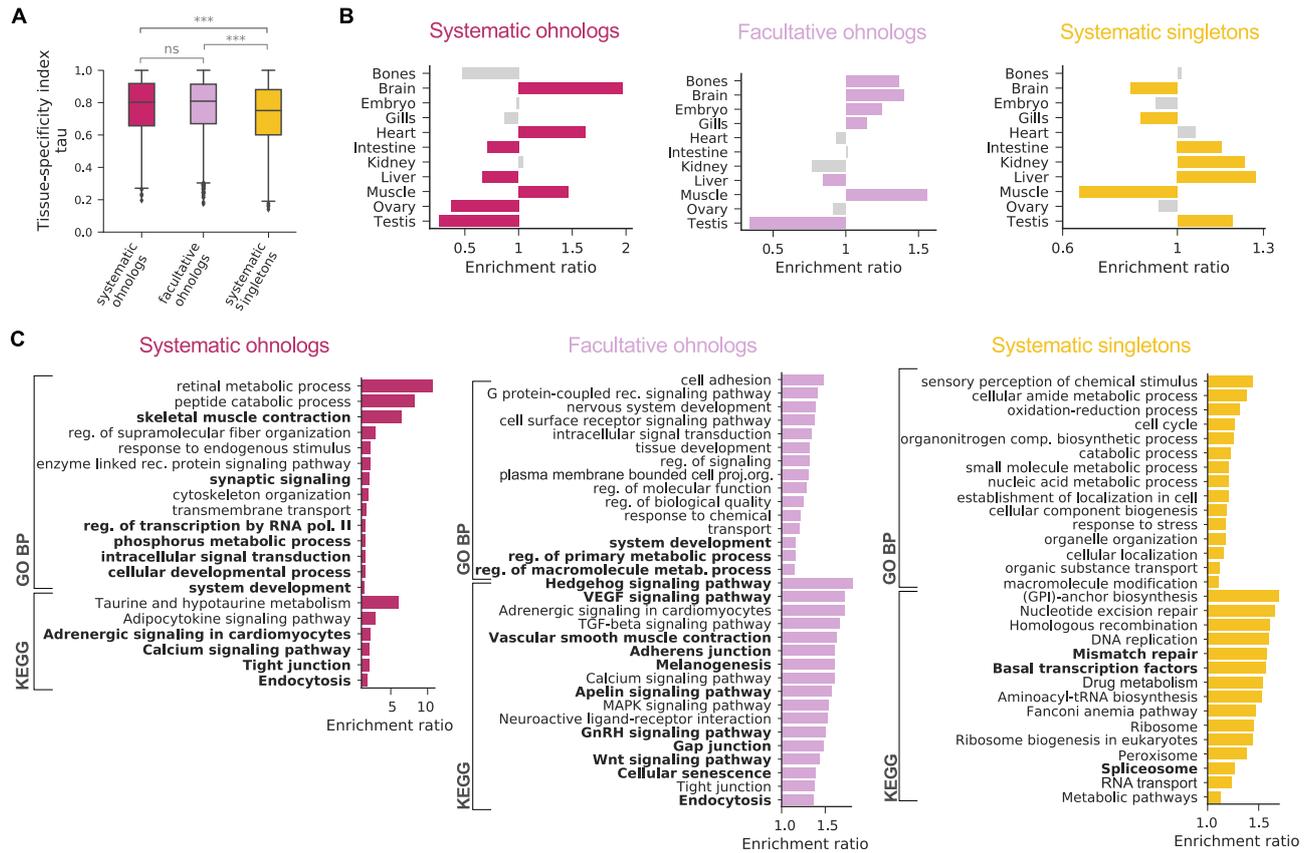


FIG. 5. Functional analysis of genes with different evolutionary trajectories after the TGD. (A) Tissue specificity of zebrafish genes with different evolutionary trajectories. Systematic ohnologs: WGD duplicates retained in two copies in all ten teleost species under study ($n = 1,828$). Facultative ohnologs: WGD duplicates retained in two copies in at least one species ($n = 7,265$). Systematic singletons: WGD duplicates returned to single-copy state in all ten species ($n = 13,895$). (B) Preferential tissue of expression for tissue-specific genes ($\tau > 0.9$) from each evolutionary trajectory. Colors denote statistical significance (hypergeometric test with BH correction, $P < 0.05$). (C) Gene ontology biological process and KEGG pathways enrichments for tissue-specific genes from each evolutionary trajectory (hypergeometric test with BH correction, $P < 0.05$). Bold, only enriched after gene tree correction with SCORPiOs.

deleterious mutations, impeding pseudogenization, and loss (Gout et al. 2010; Singh et al. 2012). Overall, the relative contributions of these processes to short- and long-term gene evolution remain unclear, although it is generally agreed that the evolutionary fate of genes following WGDs is tightly intertwined with their ancestral functions.

Polyploidizations are widespread through eukaryotic evolution, representing many independent opportunities to characterize gene evolution after WGD. Yet, WGDs represent a serious challenge to current gene tree reconstruction methods due to the high volume of duplicates and gene losses that they produce. As a result, uncertainties in gene phylogenies have been a limiting factor to all WGDs studies, whether they investigate the incidence and timing of ancient WGDs (Van de Peer et al. 2010; Ruprecht et al. 2017; Zwaenepoel and Van de Peer 2019), biases in duplicate gene retention (Scannell et al. 2006; Kassahn et al. 2009; Inoue et al. 2015), or genome organization evolution (Varadharajan et al. 2018). Studying WGD occurrences and consequences calls for the development of specific methodologies to characterize the complex histories of WGD genes. Illustrating on-going efforts, the recently published WHALE approach accounts for uncertainty in gene tree reconciliations when identifying plausible WGDs

in a species tree (Zwaenepoel and Van de Peer 2019). SCORPiOs fills another methodological gap by integrating insights from genome evolution to improve reconciled gene trees.

Genome evolution operates through three major mechanisms: nucleotide substitutions, gene duplications and losses, and genomic rearrangements. To date, integrating these different evolutionary events into a unified framework remains an open challenge (Chauve et al. 2013). The inference of reconciled gene trees, which jointly depict the history of substitutions and gene gains and losses, has been growingly addressed in recent years (Szöllösi et al. 2015). Synteny conservation has the potential to neatly complement sequence similarity in gene evolution studies, because gene order evolves via independent mechanisms, and is more resilient to saturation at deep evolutionary times (Rokas and Holland 2000). Genome organization information still remains difficult to incorporate in gene phylogenies, largely due to the lack of well-supported evolutionary models (Chauve et al. 2013) and the need for contiguous genome assemblies. The most notable effort to use extant synteny to correct gene trees in a general context showed mixed results (ParalogyCorrector within the RefineTree framework; Lafond et al. 2013;

Noutahi et al. 2016). However, we show here that in contexts where additional priors on genome organization can be leveraged, synteny patterns can be highly informative and effectively improve reconciled gene trees. As high-quality reference genome assemblies are becoming affordable and straightforward, we expect that synteny will become increasingly useful to gene history resolution in an ever more complex comparative genomics landscape. For instance, synteny-aware methods will allow the investigation of other significant biological events, such as gene conversions, which introduce discordances in the history of a gene sequence and the history of its locus.

Lastly, assessing the quality of gene trees remains a challenging task, simply because the true evolutionary history of a gene is unknown. Although statistical likelihood is widely used as a goodness-of-fit criteria to evaluate gene trees, the tree of ML according to sequence evolution is frequently incorrect (Shimodaira 2002; Szöllösi et al. 2015). It is generally assumed that the correct tree falls within an interval of equally supported trees, but numerous factors can invalidate this hypothesis, ranging from errors in sequences or their alignment to unrealistic assumptions of evolutionary models. Consequently, other goodness-of-fit metrics have been introduced, including measures of species-gene tree discordance, parsimony of the duplication and loss scenario, distances to gold standard or simulated trees, functional similarity of orthologs, and power to reconstruct ancestral genomes (Altenhoff et al. 2016; Noutahi et al. 2016). Their use has been heterogeneous across studies, guided by specific aims, relevance, and feasibility. Here, we validate SCORPIOs on real data, taking full advantage of computable metrics to demonstrate the improved quality of SCORPIOs corrected trees. In the future, efforts toward standardized benchmarking, as led by the Quest for Orthologs community, will be instrumental in producing ever more accurate gene phylogenetic trees.

Materials and Methods

Synteny Similarity of WGD Orthologs and Paralogs in the Absence of Tree Correction

Gene trees constructed with TreeBeST were downloaded from Ensembl v.89 (Vilella et al. 2009) and used to define initial orthology and paralogy relationships before correction. We extracted a set of 2,394 high-confidence, WGD-descended homologous gene pairs in zebrafish and medaka using synteny criteria similar to approaches in previous studies (Kassahn et al. 2009; Braasch et al. 2016). Well-defined WGD duplicated regions were identified in the medaka and zebrafish genomes, where one contiguous window of 15 genes or more in the spotted gar genome has homologs on exactly two different chromosomes in both medaka and zebrafish. In total, these 229 regions include 5959 genes in zebrafish (23%) and 5193 in medaka (26%). Medaka and zebrafish gene pairs were considered high-confidence WGD duplicates when they are located at the midpoint of one such 15-gene window. The orthology and paralogy relationships between those gene copies were extracted from the original gene trees (1696 orthologous and 698 paralogous gene pairs). For each medaka-zebrafish gene pair (orthologs or paralogs),

we counted across the 15-gene window: 1) the number of orthologs genes between medaka and zebrafish, according to the original trees, and 2) the number of homologs to spotted gar genes similarly retained or lost in both species.

Genome-wide synteny conservation was calculated between spotted gar (used as the outgroup) and other teleost genomes in the absence of tree correction using PhylDiag with default parameters (Lucas et al. 2014).

Gene Tree Comparisons

Nucleotide sequence alignments containing teleost fish sequences were downloaded from Ensembl v.89 and pruned of non-Neopterygii sequences. For each tree topology inferred by either TreeBeST, ParalogyCorrector or SCORPIOs, phylogenetic likelihood was computed with PhyML using the HKY85 model (Guindon et al. 2010). Likelihoods for alternative topologies were compared using the AU test implemented in Consel, at $\alpha = 0.05$ (Shimodaira and Hasegawa 2001).

Dubious nodes (or nonapparent duplication nodes) are gene duplications inferred by the reconciliation procedure where no descendant species actually contains two genes copies. They correspond to inconsistencies between the gene and species trees and are likely errors in the gene tree topology. To find dubious nodes, we used treebest sdi to reconcile gene trees with the species tree and identified all duplication nodes with a confidence score of 0 (Vilella et al. 2009).

Gene Expression Analysis

We used RNA-seq data sets from the PhyloFish database (Pasquier et al. 2016), which provides transcriptomes in fish for the following tissues: bones, brain, embryo, gills, heart, intestine, kidney, liver, muscle, ovary, and testis. We used kallisto (Bray et al. 2016) with default parameters to quantify transcript abundances for the full set of Ensembl transcripts in zebrafish and medaka. We summed transcripts per million (TPM) values of alternative transcripts to obtain a quantification of the expression of their corresponding gene. Finally, TPM values were quantile normalized within each species to obtain equivalent distributions of gene expression levels across tissues (Bolstad et al. 2003).

Functional Similarity of Orthologs

We assessed the functional similarity of zebrafish-medaka orthologs before and after gene tree correction. From the 2,387 corrected gene families, we selected the subset of 210 trees where zebrafish and medaka orthologies were reassigned by SCORPIOs. Orthologous gene expression levels were compared before and after correction using Pearson correlation and differences in mean expression across tissues. Average correlation and average expression difference before and after correction were compared using Wilcoxon signed rank tests. All tests are paired to ensure that results are unbiased by heterogeneous evolutionary rates across gene families.

Evolutionary Categories of Genes

We used the corrected set of gene trees to classify zebrafish genes with respect to their evolutionary fate across species after the TGD. We extracted all teleost gene clades from the trees and used the duplication status of the root node to determine the fate of the descending genes across species. If the root node is not a duplication, then all genes returned to a single-copy state after the TGD and we classify descending zebrafish genes as “systematic singletons.” If the root node is a duplication (corresponding to the TGD), and all descending species retained both duplicated copies (duplication confidence score = 1), we defined them as “systematic ohnologs.” Finally, if more than one but not all species retained the two copies (duplication confidence score < 1), we classify genes as “facultative ohnologs.” We excluded from this classification 2,086 zebrafish genes with no other teleost homolog in their respective subtrees.

Expression of Genes with Different Trajectories after the TGD

We used the tau index (Yanai et al. 2005) to assess the degree of tissue specificity of the expression of zebrafish genes. Tau varies between 0 and 1, where 0 means broad expression and 1 specific expression. We defined tissue-specific genes as genes with a tau index >0.9 and tested for enrichment of genes specific to particular tissues using hypergeometric tests with Benjamini and Hochberg correction for multiple testing.

Functional Enrichment of Genes with Different Trajectories after the TGD

We used WebGestalt (Liao et al. 2019) to search for functional enrichment in each evolutionary category of genes, with all zebrafish protein-coding genes as background. For systematic ohnologs, 786 and 496 out of 1,828 genes were mapped to an annotation in GO BP and KEGG pathway, respectively, 5,496 and 3,392 out of 13,895 in systematic singletons, and 2,698 and 1,625 out of 7,265 in facultative ohnologs. WebGestalt uses hypergeometric tests, corrected for multiple testing with the Benjamini and Hochberg procedure, to test for significant enrichment. We report significant enrichments at a threshold of corrected P value <0.01. For visualization purposes (fig. 5C), we reduced redundancy of functional GO terms to a maximum of 15, using the weighted set cover method implemented in Webgestalt. In all cases, the reduced set covers >97% of the total genes in each category. We repeated the analysis starting from the uncorrected Ensembl gene tree set to determine if the enriched annotations differ after applying SCORPiOs.

Availability and Implementation

SCORPiOs is coded in Python 3 and implemented as a snake-make workflow, supported on Linux and macOS. Code is publicly available on Github at <https://github.com/DyogenIBENS/SCORPIOS>. SCORPiOs is distributed under the GNU GPLv3 license. Teleost gene trees, before and after correction, have been deposited to Zenodo (doi:10.5281/zenodo.3727519).

Supplementary Material

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

Acknowledgments

We thank Pierre Vincens for the coordination of computing resources and all members of the GenoFish consortium for fruitful discussions. This work is funded by ANR GenoFish (Grant No. ANR-16-CE12-0035), and was supported by grants from the French Government and implemented by ANR (ANR-10-LABX-54 MEMOLIFE and ANR-10-IDEX-0001-02 PSL* Université Paris).

References

- Alföldi J, Lindblad-Toh K. 2013. Comparative genomics as a tool to understand evolution and disease. *Genome Res.* 23(7):1063–1068.
- Altenhoff AM, Boeckmann B, Capella-Gutierrez S, Dalquen DA, DeLuca T, Forslund K, Huerta-Cepas J, Linard B, Pereira C, Przych LP, Quest for Orthologs Consortium, et al. 2016. Standardized benchmarking in the quest for orthologs. *Nat Methods.* 13(5):425–430.
- Altenhoff AM, Studer RA, Robinson-Rechavi M, Dessimoz C. 2012. Resolving the ortholog conjecture: orthologs tend to be weakly, but significantly, more similar in function than paralogs. *PLoS Comput Biol.* 8(5):e1002514.
- Berthelot C, Brunet F, Chalopin D, Juanchich A, Bernard M, Noël B, Bento P, Silva CD, Labadie K, Alberti A, et al. 2014. The rainbow trout genome provides novel insights into evolution after whole-genome duplication in vertebrates. *Nat Commun.* 5(1):10.
- Blomme T, Vandepoele K, De Bodt S, Simillion C, Maere S, Van de Peer Y. 2006. The gain and loss of genes during 600 million years of vertebrate evolution. *Genome Biol.* 7(5):R43.
- Bolstad BM, Irizarry RA, Åstrand M, Speed TP. 2003. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics* 19(2):185–193.
- Braasch I, Gehrke AR, Smith JJ, Kawasaki K, Manousaki T, Pasquier J, Amores A, Desvignes T, Batzel P, Catchen J, et al. 2016. The spotted gar genome illuminates vertebrate evolution and facilitates human-teleost comparisons. *Nat Genet.* 48(4):427–437.
- Bray NL, Pimentel H, Melsted P, Pachter L. 2016. Near-optimal probabilistic RNA-seq quantification. *Nat Biotechnol.* 34(5):525–527.
- Byrne KP, Wolfe KH. 2005. The Yeast Gene Order Browser: combining curated homology and syntenic context reveals gene fate in polyploid species. *Genome Res.* 15(10):1456–1461.
- Catchen JM, Conery JS, Postlethwait JH. 2009. Automated identification of conserved synteny after whole-genome duplication. *Genome Res.* 19(8):1497–1505.
- Chauve C, El-Mabrouk N, Guéguen L, Semeria M, Tannier E. 2013. Duplication, rearrangement and reconciliation: a follow-up 13 years later. In: Chauve C, El-Mabrouk N, Tannier E, editors. *Models and algorithms for genome evolution*. computational biology. London: Springer London. p. 47–62. Available from: https://doi.org/10.1007/978-1-4471-5298-9_4.
- Chen X, Zhang J. 2012. The ortholog conjecture is untestable by the current gene ontology but is supported by RNA sequencing data. *PLoS Comput Biol.* 8(11):e1002784.
- De Smet R, Adams KL, Vandepoele K, Van Montagu MCE, Maere S, Van de Peer Y. 2013. Convergent gene loss following gene and genome duplications creates single-copy families in flowering plants. *Proc Natl Acad Sci U S A.* 110(8):2898–2903.
- De Smet R, Van de Peer Y. 2012. Redundancy and rewiring of genetic networks following genome-wide duplication events. *Curr Opin Plant Biol.* 15(2):168–176.
- Durand D, Halldórsson BV, Vernot B. 2006. A hybrid micro-macroevolutionary approach to gene tree reconstruction. *J Comput Biol.* 13(2):320–335.

- Engström K. 1963. Cone types and cone arrangements in teleost retinae. *Acta Zool.* 44(1–2):179–243.
- Force A, Lynch M, Pickett FB, Amores A, Yan YL, Postlethwait J. 1999. Preservation of duplicate genes by complementary, degenerative mutations. *Genetics* 151(4):1531–1545.
- Girvan M, Newman M. 2002. Community structure in social and biological networks. *Proc Natl Acad Sci U S A.* 99(12):7821–7826.
- Gout J-F, Kahn D, Duret L, Paramecium Post-Genomics Consortium. 2010. The relationship among gene expression, the evolution of gene dosage, and the rate of protein evolution. *PLoS Genet.* 6(5):e1000944.
- Guindon S, Dufayard J-F, Lefort V, Anisimova M, Hordijk W, Gascuel O. 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst Biol.* 59(3):307–321.
- Guschanski K, Warnefors M, Kaessmann H. 2017. The evolution of duplicate gene expression in mammalian organs. *Genome Res.* 27(9):1461–1474.
- Hahn MW. 2009. Distinguishing among evolutionary models for the maintenance of gene duplicates. *J Hered.* 100(5):605–617.
- Herrero J, Muffato M, Beal K, Fitzgerald S, Gordon L, Pignatelli M, Vilella AJ, Searle SMJ, Amode R, Brent S, et al. 2016. Ensembl comparative genomics resources. *Database* 2016:bav096.
- Huminiecki L, Wolfe KH. 2004. Divergence of spatial gene expression profiles following species-specific gene duplications in human and mouse. *Genome Res.* 14(10a):1870–1879.
- Inoue J, Sato Y, Sinclair R, Tsukamoto K, Nishida M. 2015. Rapid genome reshaping by multiple-gene loss after whole-genome duplication in teleost fish suggested by mathematical modeling. *Proc Natl Acad Sci U S A.* 112(48):14918–14923.
- Jaillon O, Aury J-M, Brunet F, Petit J-L, Stange-Thomann N, Mauceli E, Bouneau L, Fischer C, Ozouf-Costaz C, Bernot A, et al. 2004. Genome duplication in the teleost fish *Tetraodon nigroviridis* reveals the early vertebrate proto-karyotype. *Nature* 431(7011):946–957.
- Kassahn KS, Dang VT, Wilkins SJ, Perkins AC, Ragan MA. 2009. Evolution of gene function and regulatory control after whole-genome duplication: comparative analyses in vertebrates. *Genome Res.* 19(8):1404–1418.
- Kellis M, Birren BW, Lander ES. 2004. Proof and evolutionary analysis of ancient genome duplication in the yeast *Saccharomyces cerevisiae*. *Nature* 428(6983):617–624.
- Kenny NJ, Chan KW, Nong W, Qu Z, Maeso I, Yip HY, Chan TF, Kwan HS, Holland PWH, Chu KH, et al. 2016. Ancestral whole-genome duplication in the marine chelicerate horseshoe crabs. *Heredity* 116(2):190–199.
- Kernighan BW, Lin S. 1970. An efficient heuristic procedure for partitioning graphs. *Bell Syst Tech J.* 49(2):291–307.
- Koonin EV. 2005. Orthologs, paralogs, and evolutionary genomics. *Annu Rev Genet.* 39(1):309–338.
- Köster J, Rahmann S. 2012. Snakemake—a scalable bioinformatics workflow engine. *Bioinformatics* 28(19):2520–2522.
- Lafond M, Chauve C, Dondi R, El-Mabrouk N. 2014. Polytoymy refinement for the correction of dubious duplications in gene trees. *Bioinformatics* 30(17):i519–i526.
- Lafond M, Semeria M, Swenson KM, Tannier E, El-Mabrouk N. 2013. Gene tree correction guided by orthology. *BMC Bioinformatics* 14(S15):S5.
- Li Z, Defoort J, Tasdighian S, Maere S, Van de Peer Y, Smet RD. 2016. Gene duplicability of core genes is highly consistent across all angiosperms. *Plant Cell* 28(2):326–344.
- Liao Y, Wang J, Jaehnic EJ, Shi Z, Zhang B. 2019. WebGestalt 2019: gene set analysis toolkit with revamped UIs and APIs. *Nucleic Acids Res.* 47(W1):W199–W205.
- Lima L, Obregón F, Matus P. 1998. Taurine, glutamate and GABA modulate the outgrowth from goldfish retinal explants and its concentrations are affected by the crush of the optic nerve. *Amino Acids* 15(3):195–209.
- Lorin T, Brunet FG, Laudet V, Volff J-N. 2018. Teleost fish-specific preferential retention of pigmentation gene-containing families after whole genome duplications in vertebrates. *G3 (Bethesda)* 8:1795–1806.
- Lucas JM, Muffato M, Crollius HR. 2014. PhylDiag: identifying complex synteny blocks that include tandem duplications using phylogenetic gene trees. *BMC Bioinformatics* 15(1):268.
- Lyall AH. 1957. Cone arrangements in teleost retinae. *J Cell Sci.* s3–s98:189–201.
- Lynch M, Conery JS. 2000. The evolutionary fate and consequences of duplicate genes. *Science* 290(5494):1151–1155.
- Makino T, McLysaght A. 2010. Ohnologs in the human genome are dosage balanced and frequently associated with disease. *Proc Natl Acad Sci U S A.* 107(20):9270–9274.
- Moriyama Y, Ito F, Takeda H, Yano T, Okabe M, Kuraku S, Keeley FW, Koshiba-Takeuchi K. 2016. Evolution of the fish heart by sub/neofunctionalization of an elastin gene. *Nat Commun.* 7(1):10.
- Muffato M, Louis A, Poisnel C-E, Crollius HR. 2010. Genomicus: a database and a browser to study gene synteny in modern and ancestral genomes. *Bioinformatics* 26(8):1119–1121.
- Noutahi E, Semeria M, Lafond M, Seguin J, Boussau B, Guéguen L, El-Mabrouk N, Tannier E. 2016. Efficient gene tree correction guided by genome evolution. *PLoS One* 11(8):e0159559.
- Ohno S. 1970. Polyploidy: duplication of the entire genome. In: Ohno S, editor. *Evolution by gene duplication*. Heidelberg (Berlin): Springer. p. 98–106. Available from: https://doi.org/10.1007/978-3-642-86659-3_17.
- Omura Y, Inagaki M. 2000. Immunocytochemical localization of taurine in the fish retina under light and dark adaptations. *Amino Acids* 19(3–4):593–604.
- Pasquier J, Braasch I, Batzel P, Cabau C, Montfort J, Nguyen T, Jouanno E, Berthelot C, Klopp C, Journot L, et al. 2017. Evolution of gene expression after whole-genome duplication: new insights from the spotted gar genome. *J Exp Zool Mol Dev Evol.* 328(7):709–721.
- Pasquier J, Cabau C, Nguyen T, Jouanno E, Severac D, Braasch I, Journot L, Pontarotti P, Klopp C, Postlethwait JH, et al. 2016. Gene evolution and gene expression after whole genome duplication in fish: the PhyloFish database. *BMC Genomics* 17(1):368.
- Rasmussen MD, Kellis M. 2007. Accurate gene-tree reconstruction by learning gene- and species-specific substitution rates across multiple complete genomes. *Genome Res.* 17(12):1932–1942.
- Rasmussen MD, Kellis M. 2011. A Bayesian approach for fast and accurate gene tree reconstruction. *Mol Biol Evol.* 28(1):273–290.
- Rokas A, Holland P. 2000. Rare genomic changes as a tool for phylogenetics. *Trends Ecol Evol.* 15(11):454–459.
- Roux J, Liu J, Robinson-Rechavi M. 2017. Selective constraints on coding sequences of nervous system genes are a major determinant of duplicate gene retention in vertebrates. *Mol Biol Evol.* 34(11):2773–2791.
- Ruprecht C, Lohaus R, Vanneste K, Mutwil M, Nikoloski Z, Van de Peer Y, Persson S. 2017. Revisiting ancestral polyploidy in plants. *Sci Adv.* 3(7):e1603195.
- Sacerdot C, Louis A, Bon C, Berthelot C, Roest Crollius H. 2018. Chromosome evolution at the origin of the ancestral vertebrate genome. *Genome Biol.* 19(1):166.
- Scannell DR, Byrne KP, Gordon JL, Wong S, Wolfe KH. 2006. Multiple rounds of speciation associated with reciprocal gene loss in polyploid yeasts. *Nature* 440(7082):341–345.
- Scornavacca C, Jacox E, Szöllösi GJ. 2015. Joint amalgamation of most parsimonious reconciled gene trees. *Bioinformatics* 31(6):841–848.
- Shimodaira H. 2002. An approximately unbiased test of phylogenetic tree selection. *Syst Biol.* 51(3):492–508.
- Shimodaira H, Hasegawa M. 2001. CONSEL: for assessing the confidence of phylogenetic tree selection. *Bioinformatics* 17(12):1246–1247.
- Singh PP, Affeldt S, Cascone I, Selimoglu R, Camonis J, Isambert H. 2012. On the expansion of “dangerous” gene repertoires by whole-genome duplications in early vertebrates. *Cell Rep.* 2(5):1387–1398.
- Singh PP, Arora J, Isambert H. 2015. Identification of ohnolog genes originating from whole genome duplication in early vertebrates, based on synteny comparison across multiple genomes. *PLoS Comput Biol.* 11(7):e1004394.
- Singh PP, Isambert H. 2020. OHNOLOGS v2: a comprehensive resource for the genes retained from whole genome duplication in vertebrates. *Nucleic Acids Res.* 48:D724–D730.

- Sollars ESA, Harper AL, Kelly LJ, Sambles CM, Ramirez-Gonzalez RH, Swarbreck D, Kaithakottil G, Cooper ED, Uauy C, Havlickova L, et al. 2017. Genome sequence and genetic diversity of European ash trees. *Nature* 541(7636):212–216.
- Som A. 2015. Causes, consequences and solutions of phylogenetic incongruence. *Brief Bioinform.* 16(3):536–548.
- Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30(9):1312–1313.
- Sukeena JM, Galicia CA, Wilson JD, McGinn T, Boughman JW, Robison BD, Postlethwait JH, Braasch I, Stenkamp DL, Fuerst PG. 2016. Characterization and evolution of the spotted gar retina. *J Exp Zool Mol Dev Evol.* 326(7):403–421.
- Szöllösi GJ, Rosikiewicz W, Boussau B, Tannier E, Daubin V. 2013. Efficient exploration of the space of reconciled gene trees. *Syst Biol.* 62(6):901–912.
- Szöllösi GJ, Tannier E, Daubin V, Boussau B. 2015. The inference of gene trees with species trees. *Syst Biol.* 64(1):e42–e62.
- Van de Peer Y, Maere S, Meyer A. 2010. 2R or not 2R is not the question anymore. *Nat Rev Genet.* 11(2):166–166.
- Van de Peer Y, Mizrachi E, Marchal K. 2017. The evolutionary significance of polyploidy. *Nat Rev Genet.* 18(7):411–424.
- van Hoek MJA, Hogeweg P. 2009. Metabolic adaptation after whole genome duplication. *Mol Biol Evol.* 26(11):2441–2453.
- Varadharajan S, Sandve SR, Gillard GB, Tørresen OK, Mulugeta TD, Hvidsten TR, Lien S, Asbjørn Vøllestad L, Jentoft S, Nederbragt AJ, et al. 2018. The Grayling genome reveals selection on gene expression regulation after whole-genome duplication. *Genome Biol Evol.* 10(10):2785–2800.
- Veitia RA, Bottani S, Birchler JA. 2008. Cellular reactions to gene dosage imbalance: genomic, transcriptomic and proteomic effects. *Trends Genet.* 24(8):390–397.
- Vilella AJ, Severin J, Ureta-Vidal A, Heng L, Durbin R, Birney E. 2009. EnsemblCompara GeneTrees: complete, duplication-aware phylogenetic trees in vertebrates. *Genome Res.* 19(2):327–335.
- Wapinski I, Pfeffer A, Friedman N, Regev A. 2007. Automatic genome-wide reconstruction of phylogenetic gene trees. *Bioinformatics* 23(13):i549–i558.
- Wu Y-C, Rasmussen MD, Bansal MS, Kellis M. 2013. TreeFix: statistically informed gene tree error correction using species trees. *Syst Biol.* 62(1):110–120.
- Yanai I, Benjamin H, Shmoish M, Chalifa-Caspi V, Shklar M, Ophir R, Bar-Even A, Horn-Saban S, Safran M, Domany E, et al. 2005. Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification. *Bioinformatics* 21(5):650–659.
- Zwaenepoel A, Van de Peer Y. 2019. Inference of ancient whole-genome duplications and the evolution of gene duplication and loss rates. *Mol Biol Evol.* 36(7):1384–1404.

2.3 Résultats complémentaires

Je présente, dans cette partie, quelques résultats complémentaires qui n'ont pas été intégrés au manuscrit publié. Ils représentent des pistes complémentaires également explorées, certaines avant et d'autres après la publication du manuscrit.

2.3.1 Extension du benchmark des arbres corrigés

Effet de l'échantillonnage des espèces

Nous avons montré, dans le manuscrit publié, que 15% des sous-arbres de gènes téléostéens présents dans la base de données d'Ensembl contiennent des erreurs de placement du nœud de duplication complète 3R. Une question qui peut se poser face à ce résultat est de savoir si les erreurs peuvent être en partie liées à l'échantillonnage des espèces. En effet, SCORPiOs recalcule les sous-arbres à corriger, qui ne contiennent que les espèces de poissons, alors que, dans la base de données d'Ensembl, TreeBeST cherche à résoudre l'arbre pour un jeu d'espèces beaucoup plus large. L'inférence réalisée par TreeBeST est donc compliquée par un espace de solutions plus grand. Pour confirmer l'apport de l'analyse de synténie, j'ai également directement recalculé avec TreeBeST la topologie des 2 387 sous-arbres corrigés. Dans les sous-arbres ainsi inférés, la position de la duplication complète n'est pas améliorée, preuve que ces topologies sont effectivement difficiles à résoudre pour TreeBeST en l'absence de l'information de synténie.

Comparaison à Generax

Generax est un outil de reconstruction d'arbres phylogénétiques par maximum de vraisemblance, développé très récemment et publié au même moment que le manuscrit décrivant SCORPiOs (publié ~ 10 jours avant, MOREL et al. 2020). Generax représente la première implémentation efficace d'une heuristique de recherche d'arbre qui maximise la vraisemblance jointe. Le terme de vraisemblance jointe intègre à la fois la vraisemblance « phylogénétique » (probabilité d'observer l'alignement étant donné l'arbre de gènes) et la vraisemblance de réconciliation (probabilité d'observer l'arbre de gènes étant donné l'arbre des espèces). Comparer les topologies d'arbres inférés par SCORPiOs et Generax est intéressant pour évaluer si une méthode de vraisemblance jointe rigoureuse est capable de corriger les erreurs systématiques retrouvées dans les arbres inférés par TreeBeST. Nous avons d'abord tenté d'appliquer Generax pour reconstruire tous les arbres de la base de données d'Ensembl Compara : le temps de calcul n'était pas faisable au regard de nos ressources de calcul. J'ai donc reconstruit avec Generax la topologie des 2 387 sous-arbres corrigés afin de les comparer aux arbres inférés par SCORPiOs, en utilisant les mêmes indicateurs qu'en Figure 3 du manuscrit présenté (support apporté par l'alignement, nombre et position des duplications et nombre de duplications non apparentes). Le résultat principal est

que Generax donne un plus grand poids au modèle d'évolution de séquences qu'au scénario de perte/duplication qui est moins parcimonieux que dans les arbres « SCORPiOs » (Figure 2.8). En revanche, Generax permet bien d'observer un excès de duplications correspondant à la position du nœud 3R (Figure 2.8), ce que les autres méthodes testées dans le manuscrit ne retrouvaient pas. Enfin, un tiers des topologies inférées sont identiques entre SCORPiOs et Generax, ce qui tend à les valider mutuellement sur un jeu d'arbres difficiles à inférer. Les deux tiers restant présentent des niveaux de désaccord variables, où Generax infère généralement des scénarios peu parcimonieux au vue de la connaissance de la duplication 3R. Une difficulté que rencontre Generax (et toutes les méthodes de reconstructions d'arbres phylogénétiques actuelles) est le fait qu'il n'estime qu'un unique taux de duplication - et un de perte - à la fois sur toutes les branches de la phylogénie des espèces et toutes les familles, ce qui ne permet pas de tenir compte d'une accélération des taux liée à l'événement de duplication complète.

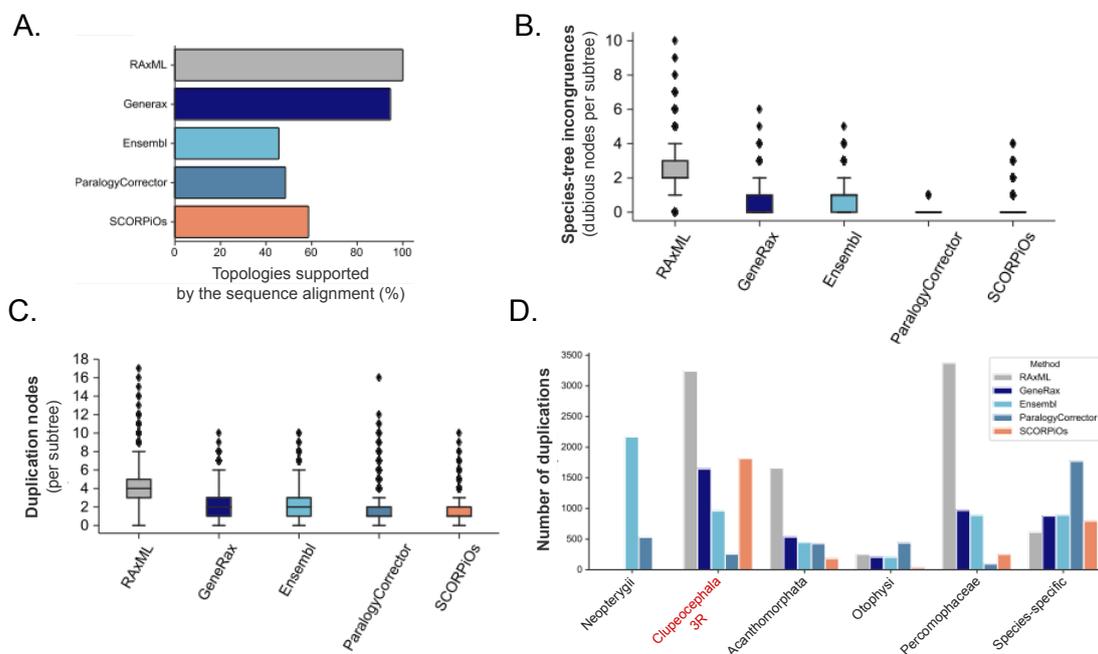


FIGURE 2.8 – Comparaison des arbres de gènes inférés par SCORPiOs et Generax. Extension de la Figure 3 du manuscrit pour inclure Generax, une nouvelle méthode de reconstruction d'arbres publiée au même moment que SCORPiOs. Les arbres de Generax sont mieux supportés par les alignements de séquences mais présente un scénario de duplications et pertes moins parcimonieux. Voir la légende de la Figure 3 du manuscrit pour le détail des panneaux.

2.3.2 Pertes réciproques de gènes

Les pertes réciproques ou pertes différentielles désignent les pertes, dans différents individus, populations ou espèces, de copies distinctes d'une paire de gènes dupliqués. Il en résulte des gènes en copie unique qui ne sont pas orthologues. Il a été proposé que de telles pertes pouvaient être une source d'incompatibilités génétiques et qu'elles pouvaient mener

à l'établissement d'une barrière reproductive au sein d'une population (LYNCH 2000). En effet, si deux sous-populations ont chacune gardé une seule copie de gène, sur des chromosomes différents, les hybrides produits pourraient ne pas recevoir le gène en question (1/4 des hybrides). Dans le cas d'un gène essentiel, l'hybride ne serait alors pas viable. Chez les levures, ce phénomène a été lié à des événements de spéciation après duplication complète (SCANNELL, BUTLER et WOLFE 2007). Chez les poissons téléostéens, l'impact des pertes réciproques est mal caractérisé, avec des études aux résultats variables quant à l'importance de leur occurrence : ~7% et ~1% de pertes réciproques, respectivement dans SÉMON et WOLFE 2007b; KASSAHN et al. 2009.

Ces pertes réciproques sont difficiles à identifier dans les arbres de gènes. Même dans le cas où la topologie de l'arbre est correcte, si toutes les espèces n'ont retenu qu'une seule copie du gène, alors la duplication pourra être complètement masquée (Figure 2.9). Cependant, les graphes d'orthologie de SCORPiOs, basés sur des analyses de synténie, permettent de les détecter : s'il s'agit d'une perte réciproque alors les copies gardées seront dans deux orthogroupes distincts. Sur le jeu de 10 poissons présentés dans l'article publié, nous avons identifié, dans les graphes de SCORPiOs, 37 loci avec des pertes de gènes réciproques et coïncidant avec la spéciation *Clupeocephala* (voir l'exemple en Figure 2.10). Nous avons remarqué que ces 37 loci étaient enrichis en gènes essentiels (test hypergéométrique, ** p-value = 5.73e-07, sur la base de listes de gènes essentiels au développement embryonnaire du poisson-zèbre de la base de données OGEE, CHEN et al. 2017). Ce résultat permet de spéculer quant à une contribution des pertes réciproques à la spéciation *Clupeocephala*.

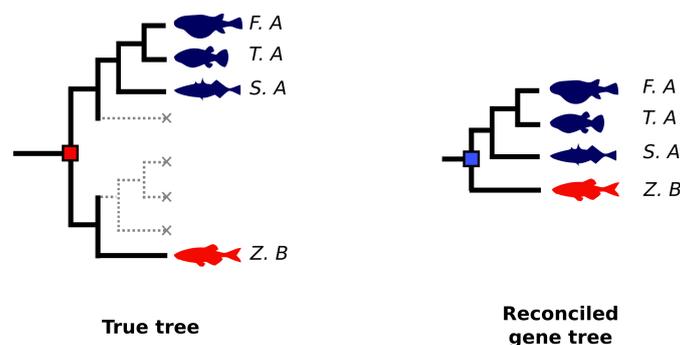


FIGURE 2.9 – Perte réciproque de gène masquée par la topologie de l'arbre des espèces. L'histoire évolutive réelle du gène présenté à gauche implique un scénario peu parcimonieux d'une duplication (carré rouge) suivi d'une perte de gènes dans tous les descendants. En conséquence, la réconciliation par parcimonie à l'arbre des espèces infère un nœud de spéciation à la place du nœud de duplication.

Néanmoins, quand nous avons par la suite commencé à appliquer SCORPiOs à des jeux de données contenant plus d'espèces, il est apparu qu'une partie de ces pertes différentielles ne coïncidaient pas avec la spéciation *Clupeocephala* : l'addition de taxons intermédiaires a précisé la position des pertes de gènes. Nous avons donc décidé de ne pas inclure ce résultat

préliminaire dans le manuscrit publié et de préciser notre analyse sur les jeux de données les plus complets possibles. De fait, la détection des pertes différentielles de gènes à l'aide de SCORPiOs reste une piste que nous souhaitons ré-explorer dans le futur.



FIGURE 2.10 – Exemple d’une perte différentielle de gène coïncidant avec la spéciation Clupeocephala. Deux copies différentes du gène *taf3* (vert clair, au centre), une sous-unité de TFIID, un facteur de transcription nécessaire à l’initiation de la transcription par l’ARN pol 2, ont été retenues chez les Otophysii (poisson-zèbre et tétra mexicain) et les autres Clupeocephala. Les gènes en synténie conservée avec l’espèce de référence, le lépisosté tacheté ("spotted gar", piste du haut) sont colorés : les gènes de même couleur appartiennent à la même famille. La perte différentielle est soutenue par des patrons de synténie différents, suggérant que les gènes appartiennent bien à des segments non-orthologues. A noter que cette perte différentielle de gènes est également datée à la spéciation Clupeocephala dans notre jeu de données de 74 poissons téléostéens (voir prochain chapitre pour des détails sur ce jeu de données).

2.3.3 Accélération de SCORPiOs

La conception de SCORPiOs a relativement bien respecté l’adage le plus populaire de la sphère du développement, attribué à Kent Beck, programmeur Unix : "Make it work, make it pretty, make it fast". Aux premières étapes, l’objectif principal était d’évaluer si intégrer une analyse de synténie dans un pipeline de reconstruction d’arbres pourrait effectivement permettre d’améliorer leur qualité. De ce fait, une priorité initiale a été donnée aux algorithmes et outils bien documentés et facile à prendre en main, tout en restant faisable sur nos jeux de données. La seconde étape a consisté au packaging et à la documentation de SCORPiOs, en vue de sa distribution (voir le paragraphe 2.1.5).

La troisième étape correspond à l’optimisation plus fine des scripts et outils utilisés, afin d’accélérer SCORPiOs. Je développe ici deux modifications que j’ai récemment apportées dans cette optique :

- Le goulot d’étranglement computationnel dans le pipeline de SCORPiOs est l’étape de recalcul des sous-arbres de gènes ou de leur vraisemblance pour tester les topologies prédites par la synténie. Dans ce sens, j’ai remplacé les étapes de calculs de vraisemblance sur topologie fixée, qui utilisaient PhyML, par RAxML. En effet, RAxML est 3 fois

plus rapide sur des alignements CDS (ZHOU et al. 2018). Ce changement n'impacte pas les résultats de SCORPiOs puisque j'utilise des modèles d'évolution comparables : seules les approximations numériques de chacun des outils diffèrent légèrement, ce qui n'impacte pas les résultats des tests de vraisemblance lorsque les vraisemblances comparées sont calculées par le même programme.

- Dans la dernière version de SCORPiOs, une option est disponible afin de séparer les graphes d'orthologie par clustering spectral à la place de l'algorithme de Girvan-Newman. L'algorithme de Girvan-Newman, bien que très intuitif, ne possède pas d'implémentation à la fois bien documentée et efficace. En effet, l'implémentation intégrée à SCORPiOs est celle proposée par la librairie NetworkX, implémentée en pur python et montrée très peu performante. Le clustering spectral peut résoudre le même problème du « minimal cut graph », à savoir séparer un graphe en deux composantes de manière à supprimer le moins d'arêtes. Le clustering spectral repose sur des opérations d'algèbre linéaire, dont la complexité est celle d'une décomposition en éléments propres. Même si la complexité est moins bonne que celle de Girvan-Newman ($\sim O(n^3)$ contre $\sim O(n^2)$ avec n le nombre de nœuds), en pratique, sur nos graphes de seulement une centaine de nœuds, l'efficacité des implémentations existantes ainsi que le coût réduit de l'opération de base rendent le clustering spectral plus rapide. Sur un jeu de données complexe (74 espèces dupliquées), j'ai validé que les découpages de graphes étaient similaires avec les deux approches (95% de solutions identiques), et environ 35 fois plus rapide avec le clustering spectral.

Enfin, un dernier point concerne la gestion des dépendances de SCORPiOs, en vue de faciliter sa distribution. Conda est probablement le système le plus complet pour gérer tout type de dépendances logicielles. De plus, il peut être directement intégré à Snakemake, le langage dans lequel SCORPiOs est implémenté. Récemment, la chaîne bioconda a connu un tel essor que la plupart des logiciels de bioinformatique sont disponibles sur conda. Cependant, conda rencontre durant ces dernières années un problème majeur. Victime de son succès, le nombre de logiciels disponibles est à l'origine de graphes de dépendances complexes que les algorithmes de conda ne résolvent pas efficacement. Cela s'est également ressenti au cours du développement de SCORPiOs : la création de l'environnement virtuel durait une dizaine de minutes au début de sa création et prend maintenant plusieurs heures, dû à la croissance de la base de données. Récemment, nous avons introduit dans les instructions d'installation l'utilisation de mamba (<https://github.com/mamba-org/mamba>) pour installer les dépendances de SCORPiOs. Mamba est un outil de résolution de dépendances développé par l'équipe QuantStack et écrit au-dessus de conda. Mamba permet de résoudre les dépendances d'un environnement conda à l'aide d'algorithmes efficaces, inspirés du système d'exploitation Fedora. De cette façon, la durée de l'installation des dépendances est repassée à moins de 5 minutes.

Chapitre 3

Cartographie à haute résolution des régions anciennement tétraploïdes chez les poissons téléostéens

3.1 Introduction

Je présente, dans ce chapitre, un article en préparation que j'ai co-écrit en tant que première auteure et intitulé « High-resolution map of ancient tetraploidy across 74 teleost fish genomes ». Ce manuscrit décrit l'application de SCORPiOs à un jeu de données de 101 espèces, dont 74 poissons téléostéens. La qualité des arbres de gènes inférés nous a permis d'établir la première carte de paralogie, à grande échelle et à haute résolution, chez les poissons téléostéens. Cette cartographie renseigne sur la localisation des régions de syntenie doublement conservée (DCS), dérivées de chromosomes ancestralement dupliqués. La carte lie la position des régions dupliquées à la fois au sein d'une espèce (régions paralogues) et entre espèces (régions orthologues). Ici, je présente l'état de l'art des stratégies actuellement employées pour caractériser les régions dupliquées chez les téléostéens. Par la suite, j'introduis les méthodes utilisées pour la reconstruction de génomes ancestraux, les difficultés qu'elles rencontrent en présence de duplications complètes et enfin l'information qu'elles fournissent pour l'établissement de la carte de paralogie. Enfin, je présente brièvement la ressource riche et unique que constitue cette cartographie évolutive et les possibilités qu'elle ouvre pour la génomique comparative des poissons.

3.1.1 Érosion des chromosomes dupliqués

Les duplications complètes sont des événements mutationnels dramatiques, générant une copie de l'ensemble des chromosomes d'une espèce. Durant les millions d'années qui suivent la duplication complète, la signature nette de paires de chromosomes dupliqués, ou homéologues, est érodée par les processus d'évolution des génomes. La rediploïdisation, ou retour à l'état diploïde, s'effectue à travers des pertes massives de gènes (INOUE

et al. 2015; SESSION et al. 2016; CHEN et al. 2019), par la perte de segments entiers de chromosomes homéologues (WANG et al. 2005; DU et al. 2020) ainsi que la divergence des séquences d'ADN. De plus, les duplications complètes et l'instabilité génomique qu'elles entraînent sont proposées engendrer un taux élevé de remaniements génomiques (SÉMON et WOLFE 2007a; HUFTON et PANOPOULOU 2009).

Les remaniements chromosomiques, à l'origine du mélange de la position de segments d'ADN le long du génome, sont traditionnellement séparés en deux classes : les réarrangements interchromosomiques (transposition, translocation et fusion) et intrachromosomiques (inversions et fissions). On distingue également les remodelages dits déséquilibrés, qui induisent une modification du contenu d'ADN (délétions et duplications), des remaniements équilibrés qui ne font que déplacer des morceaux d'ADN. L'un des mécanismes moléculaires principaux à l'origine des réarrangements est la réparation erronée des cassures double brin de l'ADN, soit à travers une recombinaison illégitime, soit à travers un défaut de la réparation par jonction d'extrémités non homologues. Lorsqu'ils surviennent dans la lignée germinale, les réarrangements génomiques peuvent être transmis à la descendance et finir par être fixés dans une population, par dérive ou par sélection.

Il est proposé que les duplications complètes seraient à l'origine de taux élevés de réarrangements chromosomiques, à travers un taux élevé soit de leur occurrence, soit de leur fixation, les deux étant non-mutuellement exclusifs. Dans le premier cas, il est proposé qu'ils soient induits par l'augmentation des contacts en méiose du fait de la présence de copies surnuméraires de chromosomes, ainsi qu'à l'activation d'éléments transposables (voir le chapitre 1, paragraphe 1.3.3), qui fournissent un substrat aux recombinaisons illégitimes (FESCHOTTE et PRITHAM 2007). Dans le second cas, les réarrangements permettant de différencier les chromosomes homéologues et promouvant ainsi la formation de bivalents seraient favorisés par la sélection naturelle, car ils stabilisent les méioses (LEVY et FELDMAN 2004; SÉMON et WOLFE 2007a; HUFTON et PANOPOULOU 2009).

Chez les poissons téléostéens, il n'est pas définitivement démontré qu'il y ait eu une réelle augmentation des taux de réarrangements directement corrélée à la 3R (SÉMON et WOLFE 2007a; HUFTON et PANOPOULOU 2009). Par ailleurs, l'analyse du génome du lépisosté tacheté (groupe externe des téléostéens), via sa comparaison au génome du poulet, a permis de mettre en évidence un certain nombre d'événements de fusion de chromosomes dans la lignée téléoste, avant la duplication complète. Les génomes du lépisosté tacheté et du poulet contiennent des micro-chromosomes, hérités de l'ancêtre Vertébré. Les caryotypes des poissons téléostéens n'en contiennent pas : la comparaison au génome du lépisosté a révélé que les micro-chromosomes ancestraux ont été fusionnés avant la duplication complète 3R (Figure 3.1). Additionnés aux réarrangements post-3R, ces événements de fusion brouillent le signal de chromosomes dupliqués lors de la comparaison des téléostéens à des

génomés non-dupliqués (Figure 3.1). En effet, même si, localement, les génomes du lépisosté tacheté et des poissons téléostéens sont en synténie doublement conservée, à l'échelle des chromosomes, la synténie est dégradée et la correspondance 1-à-2 n'est pas directe.

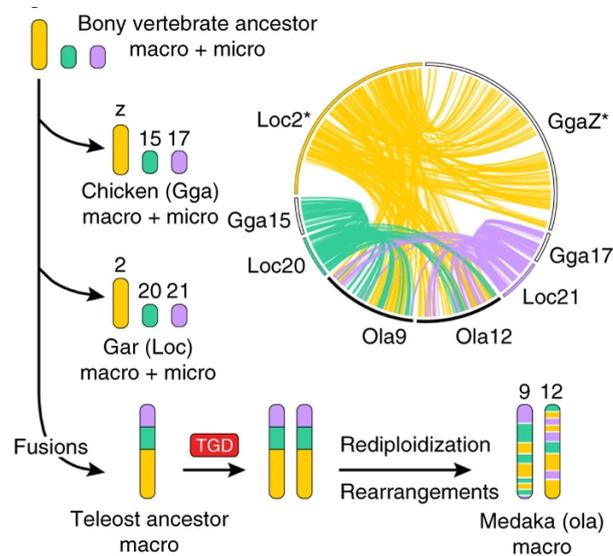


FIGURE 3.1 – Fusions de micro-chromosomes avant la duplication 3R. Exemple d'une fusion de micro-chromosomes (en vert et violet) avec un macro-chromosome (en jaune), dans la lignée des téléostéens. Suite à des réarrangements intra-chromosomiques, la correspondance entre les chromosomes du medaka et du lépisosté n'est plus directement évidente. Figure tirée de BRAASCH et al. 2016.

Du fait de ces remodelages génomiques accumulés dans la lignée des téléostéens, caractériser les régions dupliquées revient à identifier, dans les espèces modernes, les reliques des chromosomes ancestralement dupliqués. Cette tâche est d'autant plus compliquée que la duplication est ancienne. Dans le prochain paragraphe, je décris les approches utilisées pour caractériser les régions de synténie doublement conservée chez les poissons.

3.1.2 Caractérisation des régions de synténie doublement conservée

Caractériser l'organisation des génomes permet de mettre en lumière leur fonction et les processus évolutifs qui les façonnent. Motivés par ces objectifs, différents outils, approches et ressources ont été développés afin d'identifier les régions dupliquées qui parsèment les génomes de poissons et autres espèces anciennement polyploïdes.

Identification des DCS pour valider les prédictions d'orthologie

Les patrons de synténie doublement conservée (DCS) peuvent être directement identifiés à travers la comparaison d'un génome dupliqué contre un génome non-dupliqué : un segment génomique d'une espèce non-dupliquée aura des orthologues sur deux segments distincts d'un génome dupliqué. Ces DCS servent, en premier lieu, à valider les prédictions

d'orthologie et de paralogie entre gènes. Plusieurs bases de données ont vu le jour pour permettre de visualiser et comparer les génomes, comme Genomicus (MUFFATO et al. 2010), YGOB pour les levures (BYRNE et WOLFE 2005) ou, plus spécifiquement pour les génomes de poissons, la Synteny Database (CATCHEN, CONERY et POSTLETHWAIT 2009). Cependant, Synteny Database n'est plus mise à jour et contient uniquement 6 espèces de poissons téléostéens. Par ailleurs, la base fonctionne à travers la comparaison d'une espèce dupliquée contre un groupe externe non-dupliqué et ne permet pas directement de lier les DCS d'espèces dupliquées entre elles.

Identification des DCS pour étudier les mécanismes de rediploïdisation

POInT (« Polyploid Orthology Inference Tool ») est un outil pour étudier l'évolution des génomes après duplication complète (CONANT et WOLFE 2008). Initialement développé pour la duplication complète des levures, POInT a récemment été appliqué à 8 génomes de poissons téléostéens (CONANT 2020). L'identification des DCS par POInT repose sur le placement des gènes des espèces dupliquées dans chacun des deux blocs dupliqués générés par la duplication complète, de façon à maximiser les adjacences de gènes. Par la suite, l'auteur modélise le processus de diploïdisation à travers des modèles probabilistes imbriqués décrivant les pertes de gènes. Il met en évidence un biais significatif de rétention au sein des paires de blocs dupliqués. Ce signal le mène à proposer que la duplication 3R est une allopolyploïdie (voir le chapitre d'introduction 1.3.3).

Néanmoins, peut-être dû à l'état beaucoup plus réarrangé des génomes de poissons que de ceux de levures, POInT ne permet pas d'obtenir une collection exhaustive de régions dupliquées. En effet, moins de 7000 gènes de chaque espèce sont placés dans les blocs dupliqués, et seuls 8 blocs sont d'une taille supérieure à 100 gènes. En conséquence, les blocs de gènes étudiés sont loin de représenter les « sous-génomes » issus de la duplication complète. De fait, le biais de rétention local observé pourrait également s'expliquer par la perte simultanée de plusieurs gènes co-régulés d'un même bloc (pertes de gènes clusterisées, BUGGS et al. 2012; MAKINO et MCLYSAGHT 2012). Cette étude souligne le besoin de mieux caractériser la structure des génomes de poissons avant de pouvoir disséquer les mécanismes de la diploïdisation.

En résumé, les stratégies actuellement utilisées pour caractériser les DCS chez les poissons téléostéens, que ce soit pour valider des relations d'orthologie ou pour étudier les modes d'évolution des génomes après duplication, ne permettent pas de couvrir exhaustivement les génomes de poissons. Dans la prochaine sous-partie, j'introduis l'état de l'art des méthodes de reconstructions ancestrales de génomes, ce qu'elles nous apprennent sur les génomes de poissons, et comment elles peuvent permettre de préciser les patrons de synténie doublement conservée.

3.1.3 Reconstruction de génomes ancestraux en présence de duplication complète

Contexte méthodologique

A partir de l'arrangement de gènes, ou autres marqueurs, dans les espèces modernes, les méthodes de reconstructions ancestrales visent à inférer l'organisation de ces marqueurs dans leur(s) ancêtre(s) commun(s). Basées sur le principe de parcimonie, leur objectif est d'inférer le génome de l'ancêtre impliquant le moins de réarrangements évolutifs pour arriver aux génomes modernes. Le problème de l'identification des régions dupliquées pourrait être résolu à travers la reconstruction du génome ancestral post-duplication. La connaissance de ce génome ancestral permettrait, directement, de dessiner l'histoire évolutive des chromosomes dupliqués jusqu'aux espèces modernes.

On peut citer deux classes principales d'approches pour la reconstruction de génomes ancestraux : les approches basées sur l'analyse des points de cassure et celles fondées sur la conservation de la synténie. Les approches combinatoires, ou d'analyse des points de cassures, reposent sur des heuristiques pour inférer les génomes ancestraux optimaux, ainsi que la suite d'opérations (réarrangements génomiques) à leur appliquer pour obtenir les génomes modernes. Dans les formulations les plus courantes, l'ensemble des réarrangements est modélisé par une unique opération, le "Double-Cut and Join", qui découpe un/des segment(s) génomique(s) à partir de deux points de cassure et le(s) recolle ailleurs (YANCOPOULOS, ATTIE et FRIEDBERG 2005). Malgré la simplification apportée par le modèle DCJ et le développement de nouvelles heuristiques dans ce sens, les approches combinatoires sont encore mal définies en présence de marqueurs dupliqués (EL-MABROUK et SANKOFF 2012 ; FENG, ZHOU et TANG 2017). En conséquence, les méthodes employées pour reconstruire les génomes ancestraux chez les poissons ne modélisent pas explicitement les événements de réarrangements évolutifs et s'appuient sur la conservation de synténie.

Le principe sur lequel repose cette seconde classe de méthodes est le suivant : si des marqueurs sont trouvés dans le même ordre dans deux génomes, alors on fait l'hypothèse que cet ordre était déjà établi dans leur ancêtre commun. Suivant le niveau auquel elles résolvent les génomes ancestraux, elles proposent des reconstructions macrosynténiques (inférence du caryotype ancestral à partir de blocs de synténie ; ANGES, DESCHRAMBLER, JONES et al. 2012a ; KIM et al. 2017) ou microsynténique (inférence de l'ordre ancestral des gènes dans les chromosomes ancestraux ; AGORA, GapAdj, DecoStar, MUFFATO 2010 ; GAGNON, BLANCHETTE et EL-MABROUK 2012 ; DUCHEMIN et al. 2017). Les méthodes microsynténiques s'appuient sur les adjacences de gènes observées dans les génomes modernes. Ces adjacences sont collectées et la reconstruction ancestrale consiste à chaîner les adjacences les mieux soutenues pour reconstruire l'ordre des gènes dans l'ancêtre. Les méthodes macrosynténiques, quant à elles, découpent les génomes modernes en blocs de synténie

conservée et cherchent l'arrangement de ces blocs dans les génomes ancestraux. Des approches micro- et macrosynténiques ont été appliquées au problème des reconstructions ancestrales en présence de duplication complète de génome. J'introduis, dans le prochain paragraphe, les développements particuliers au problème des duplications complètes.

Le cas des duplications complètes

Le cas particulier des duplications complètes a suscité un grand intérêt pour le développement de méthodes plus sophistiquées, capables de tenir compte du doublement des chromosomes entre l'ancêtre pré-duplication et les espèces modernes (MUFFATO et CROLLIUS 2008 ; GAGNON, BLANCHETTE et EL-MABROUK 2012 ; EL-MABROUK et SANKOFF 2012). Les méthodes microsyténiques, basées sur les adjacences de gènes, rencontrent une difficulté supplémentaire dans le contexte de duplication complète. En effet, les pertes de gènes massives qui suivent la duplication cassent les adjacences de gènes (Figure 3.2). Pour pallier ce problème, Gagnon et al. ont développé GapAdj, une méthode qui considère des adjacences « relâchées » : les adjacences sont définies entre gènes situés en dessous d'une certaine distance fixée (GAGNON, BLANCHETTE et EL-MABROUK 2012). Plus la distance autorisée est grande, moins la reconstruction est précise, mais plus elle permet de reconstruire de longs segments ancestraux contigus (CARs). GapAdj a notamment été appliqué avec succès pour reconstruire des génomes ancestraux pré-duplication chez les levures et les plantes (GAGNON, BLANCHETTE et EL-MABROUK 2012).

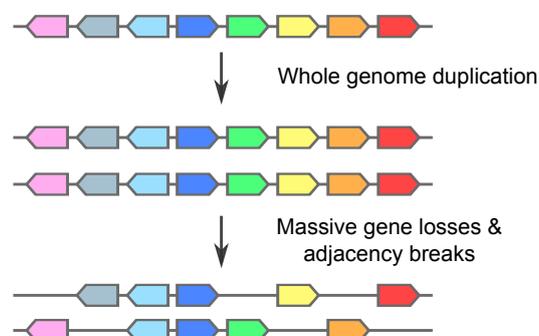


FIGURE 3.2 – Perturbation des adjacences de gènes après duplication complète. Après duplication complète, les pertes massives de gènes perturbent les adjacences de gènes sans qu'il n'y ait de réarrangement génomique.

Néanmoins, la qualité des reconstructions par GapAdj est dégradée en présence de taux de réarrangements et de perte de gènes élevés (GAGNON, BLANCHETTE et EL-MABROUK 2012 ; NAKATANI et MCLYSAGHT 2017). Aucune méthode de reconstruction de l'ordre des gènes n'a pu être appliquée avec succès pour reconstruire le génome de l'ancêtre Teleostei. Cet échec est relié au fort niveau de réarrangements et de pertes de gènes survenus à la suite de la 3R (INOUE et al. 2015). De plus, l'ancienneté de la duplication rend également

difficile l'inférence des arbres de gènes et relations d'orthologie, réduisant le signal exploitable pour inférer les adjacences ancestrales.

En conséquence, les reconstructions ancestrales de l'ancêtre Téléoste sont uniquement des reconstructions macrosynténiques. En effet, les signatures de synténie doublement conservée peuvent être exploitées pour reconstruire les génomes pré-duplication. Dans le paragraphe suivant, j'introduis les reconstructions de l'ancêtre Téléoste.

Reconstructions du génome de l'ancêtre Teleostei

Alors que reconstruire les ancêtres dans le contexte de duplication complète est compliqué par les pertes de gènes, la reconstruction de l'ancêtre pré-duplication est facilitée par l'analyse des blocs de DCS. Les premières reconstructions macrosynténiques de l'ancêtre Teleostei pré-duplication sont venues de la comparaison de différentes combinaisons de génomes dupliqués/non dupliqués : poisson-zèbre, medaka et génome humain (NARUSE et al. 2004) ; tétraodon et génome humain (JAILLON et al. 2004) ; tétraodon, poisson-zèbre et génome humain (WOODS et al. 2005) ; poisson-zèbre, medaka, tétraodon et génome humain (KASAHARA et al. 2007). Ces reconstructions reposent toutes sur le même grand principe de combinaison des DCS. A travers l'identification de blocs de synténie doublement conservée, rassemblés par la suite en groupes de DCS alternant sur les mêmes paires (ou triplets) de chromosomes des espèces dupliquées, elles ont permis d'esquisser entre 11 et 13 chromosomes pré-duplication chez l'ancêtre Teleostei.

La dernière reconstruction en date (NAKATANI et MCLYSAGHT 2017), qui constitue la référence actuelle du génome pré-duplication, infère, comme (KASAHARA et al. 2007), un caryotype à 13 chromosomes (Figure 3.4). Cette reconstruction a plusieurs avantages par rapport aux précédentes. Premièrement, elle tire évidemment profit d'assemblages de génomes plus récents et donc de meilleure qualité, et combine un plus grand nombre de comparaison entre génomes dupliqués et génomes non-dupliqués. La reconstruction intègre également une plus grande portion des génomes modernes : entre 70 et 90% des gènes sont placés dans des blocs descendant des chromosomes ancestraux. De plus, la formulation du problème dans un cadre probabiliste permet d'attribuer des probabilités postérieures aux segments inférés (Figure 3.3).

L'approche développée par NAKATANI et MCLYSAGHT 2017 s'inspire des algorithmes utilisés dans le domaine de l'analyse de texte. L'élégance de la méthode réside dans l'illustration directe qu'elle représente de la métaphore des génomes comme documents historiques de leur évolution (Figure 3.3). Comme les reconstructions précédentes, Nakatani et McLysaght utilisent les signatures laissées par la distribution des gènes orthologues entre un génome dupliqué et un génome non dupliqué (qui dessine donc les DCS). Cette distribution d'orthologues représente le « texte » dans des « documents » que sont les segments de génome.

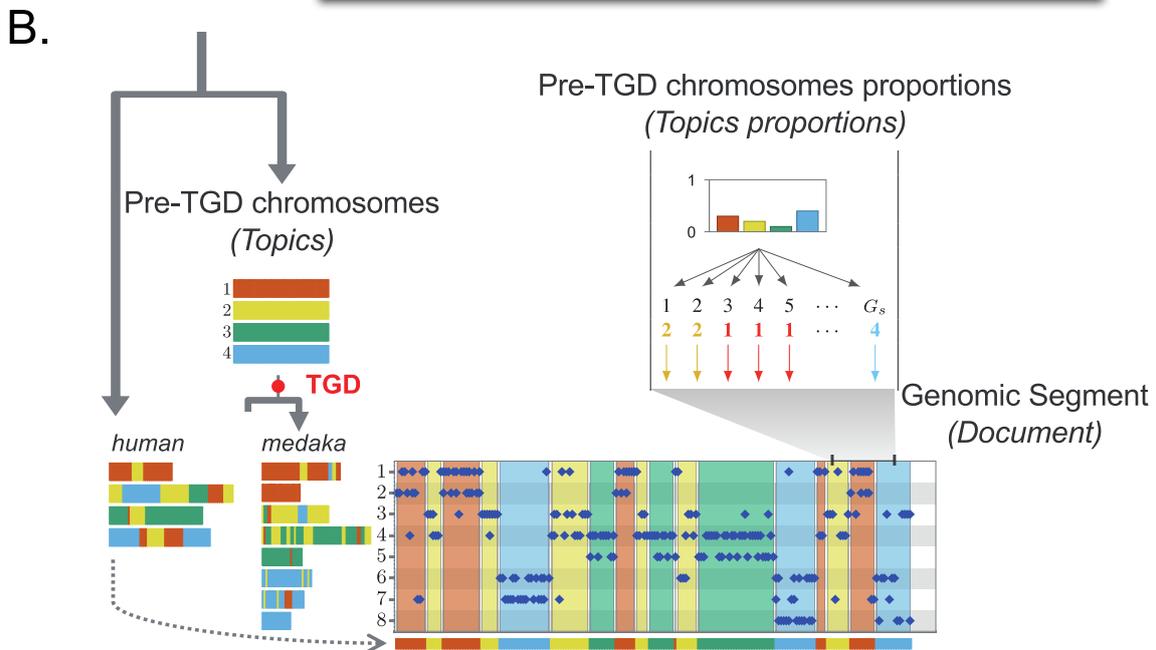
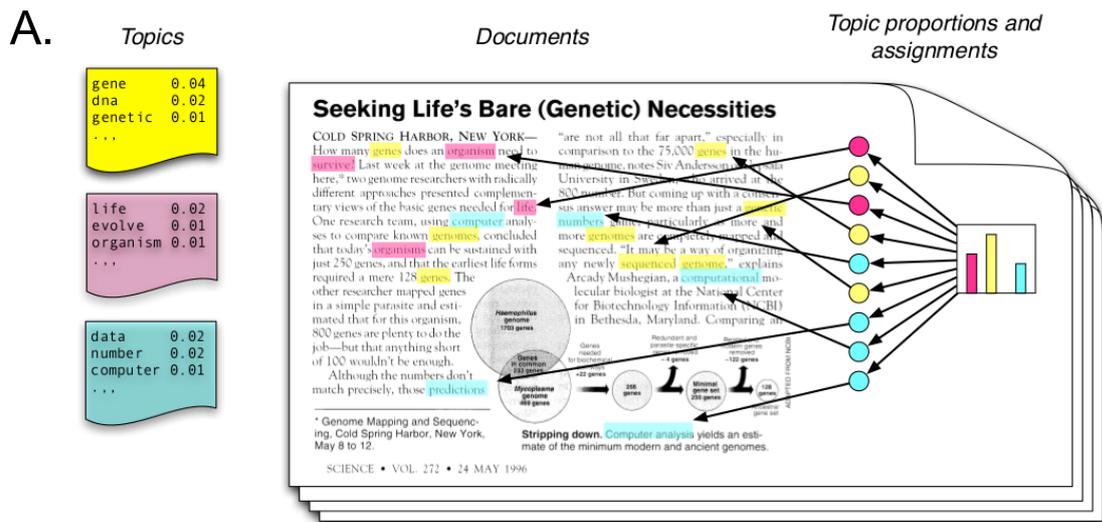


FIGURE 3.3 – Analogie entre les modèles thématiques et le modèle macrosynténique développé par NAKATANI et MCLYSAGHT 2017. A. Représentation schématique d'un modèle thématique décrivant la structure des sujets d'un document (modèle d'allocation de Dirichlet latente). Le document contient plusieurs thèmes (la génétique, le vivant et l'analyse de données), chacun défini par une distribution de probabilité sur des mots. L'application de méthodes bayésiennes variationnelles permet d'estimer les thèmes qui constituent la structure cachée du document. Figure adaptée de BLEI 2012. B. Modèle macrosynténique : les distributions des gènes orthologues entre espèces dupliquées et espèce non-dupliquées (points bleus) sont les « mots » qui constituent les segments génomiques ou « documents ». Les orthologues sont globalement retrouvés sur deux chromosomes suite à la duplication. A partir de la distribution des orthologues, il est possible d'inférer les chromosomes pré-duplication. Figure simplifiée à partir de NAKATANI et MCLYSAGHT 2017.

L'objectif de la reconstruction est de regrouper les blocs génomiques de même distribution d'orthologues en tant que descendants d'un même chromosome ancestral pre-duplication. Pour revenir au problème de l'analyse de texte, cela est équivalent à grouper en grandes thématiques les sujets abordés dans différents documents. Les auteurs ont également tenté de reconstruire l'ordre des gènes dans le génome ancestral, en utilisant une méthode hybride qui combine l'approche macrosynténique à une reconstruction microsyténique obtenue avec l'outil GapAdj (GAGNON, BLANCHETTE et EL-MABROUK 2012). Cependant, à cause du taux élevé de réarrangements, cette approche n'a pas fonctionné : pour obtenir de grands segments ancestraux, la reconstruction demande d'utiliser une distance relâchée qui ne permet pas d'obtenir des CARs cohérents avec le caryotype à 13 chromosomes inféré.

Les reconstructions du génome de l'ancêtre Teleostei ont permis de dériver plusieurs observations concernant l'évolution des génomes de poissons. L'occurrence de réarrangements interchromosomiques ont été relativement rares chez les téléostéens, en comparaison aux génomes de mammifères (KASAHARA et al. 2007 ; NAKATANI et MCLYSAGHT 2017). Additionnellement, le génome du poisson-zèbre a accumulé un plus grand nombre de remaniements interchromosomiques que le medaka, le tétraodon et l'épinoche (KASAHARA et al. 2007 ; NAKATANI et MCLYSAGHT 2017). Cependant, jusque-là, les reconstructions ancestrales ne s'appuient que sur un petit jeu d'espèces. De plus, la combinaison des blocs de DCS ne permet que la reconstruction de l'ancêtre pré-duplication. Aucune ressource ne permet d'identifier les relations entre chromosomes dupliqués de différentes espèces. J'introduis, dans la prochaine sous-partie, comment ces reconstructions ancestrales peuvent être exploitées pour étendre nos connaissances dans ce sens.

3.1.4 Établissement et impact de la carte des régions dupliquées

A première vue, il pourrait sembler que l'information d'orthologie et de paralogie entre segments de génomes dupliqués est calculée par SCORPiOs. Cependant, SCORPiOs n'est pas une méthode de reconstruction ancestrale. SCORPiOs utilise l'information de synténie conservée la plus locale possible pour prédire avec précision l'histoire évolutive d'un gène en particulier. L'étape dite de « threading » permet de définir des segments orthologues uniquement pour une paire de génomes et sur une fenêtre très petite (15 gènes). Les conflits de « threading » entre les différentes paires d'espèces ne sont résolus qu'à l'échelle d'une famille de gènes, à travers la découpe des graphes d'orthologie. Aucune reconstruction multi-espèces n'est effectuée. De plus, SCORPiOs ignore la présence possible de points de cassure dans la fenêtre considérée : son seul objectif est de placer les gènes orthologues dans des blocs orthologues (réciproquement pour les paralogues). En revanche, l'amélioration des relations d'orthologie et de paralogie des arbres représente un appui puissant qui permet de faire passer les informations d'orthologie de l'échelle de la famille de gènes à l'échelle du bloc de gènes.

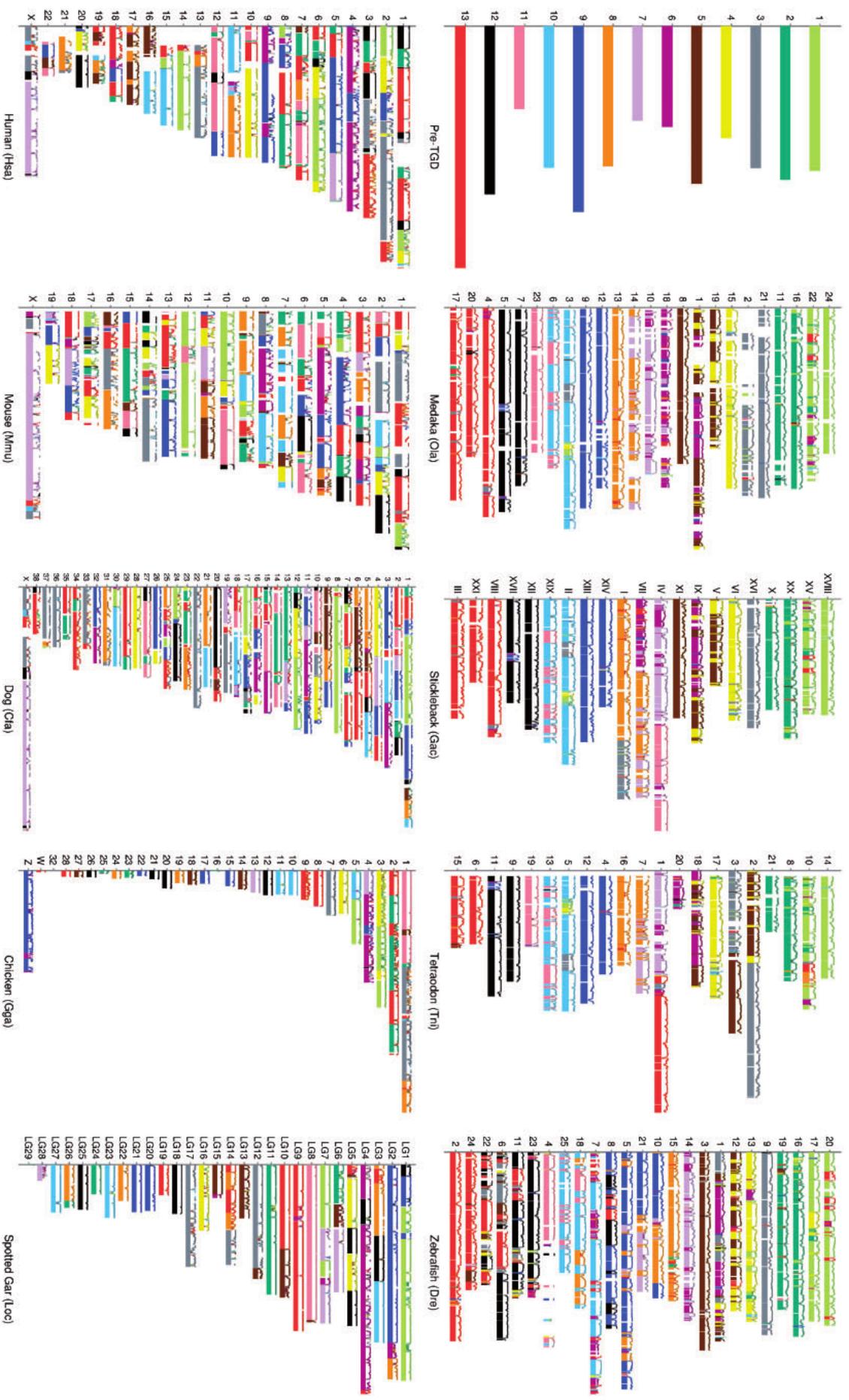


FIGURE 3.4 – Reconstruction du caryotype de l'ancêtre Teleostei pré-duplication. Dans les espèces téléostes modernes (medaka, épinouche, tetraodon et poisson-zèbre), les blocs qui descendent des 13 chromosomes ancestraux sont représentés avec les mêmes couleurs. Les segments orthologues dans les génomes de 5 autres Vertébrés sont également représentés. Figure tirée de NAKATANI et MCLYSAGHT 2017.

Les reconstructions ancestrales introduites précédemment fournissent un cadre idéal pour l'identification des blocs dupliqués descendant des anciens homéologues. Pour simplifier, l'établissement de la carte de paralogie peut se résumer à un passage de la reconstruction pré-duplication à un niveau post-duplication, en séparant les blocs en deux ensembles 'a' et 'b'.

Principe

J'introduis ici les principes sur lesquels repose l'établissement de la carte de paralogie, en justifiant les choix effectués, sans rentrer dans le détail de l'implémentation des différentes étapes qui sont présentées dans le manuscrit.

Au vu des difficultés posées par l'événement de duplication complète, nous n'avons pas reconstruit de caryotype post-duplication directement, de la manière bottom-up classique. Pour simplifier le problème, nous avons choisi de nous appuyer sur la reconstruction du caryotype pré-duplication de référence NAKATANI et MCLYSAGHT 2017. La stratégie mise en place est conceptuellement assez simple et repose sur (i) la connaissance du caryotype pré-duplication et (ii) les arbres de gènes des espèces considérées. La reconstruction ancestrale de l'ancêtre pré-duplication renseigne sur les blocs d'espèces modernes descendant de chromosomes homéologues, entre lesquels il faut établir les relations d'orthologie et de paralogie. Ces relations sont directement extraites des arbres de gènes, dont la qualité est primordiale pour établir ces relations entre blocs. Nous avons montré, dans le manuscrit, que la correction des arbres par SCORPiOs augmente drastiquement le nombre de gènes pouvant être annotés dans la carte de paralogie (il passe de 61% de gènes annotés en moyenne sur toutes les espèces à 83%).

La carte de paralogie est établie en deux étapes. En premier lieu, sur les 4 espèces de références intégrées dans NAKATANI et MCLYSAGHT 2017 et pour chaque chromosome pré-duplication, les blocs 'a' et 'b' qui en descendent sont inférés à partir des relations d'orthologie et de paralogie des arbres de gènes. Dans un second temps, ces annotations sont transférées à toutes les autres espèces présentes dans les arbres de gènes à travers l'annotation des nœuds ancestraux.

Contribution à l'état-de-l'art

La carte de paralogie représente une ressource unique pour la génomique comparative des génomes de poissons : il n'existe pas de référence répertoriant de manière précise les relations d'orthologie et de paralogie entre gènes et segments génomiques de poissons téléostéens. Le jeu d'espèces intégrées aux arbres de gènes a été spécifiquement sélectionné pour donner une représentation à tous les clades majeurs de poissons. Ce jeu est constitué

d'un mélange de génomes disponibles publiquement, ainsi que des génomes séquencés *de novo* dans le cadre du consortium « Genofish » dans lequel mon projet de thèse s'inscrit. La nouveauté par rapport aux reconstructions précédentes est également l'intégration d'un génome à l'assemblage chromosomique d'une espèce des Osteoglossiformes, un groupe qui a divergé rapidement des autres poissons téléostéens après la duplication 3R.

Les génomes séquencés dans le cadre projet « Genofish » ne sont pas encore tous disponibles. Pour cette raison, l'approche computationnellement simple à l'origine de la carte de paralogie permettra sa mise à jour au fil de l'avancée du projet. Nous avons également montré que la carte de paralogie était robuste aux potentielles imprécisions des limites de blocs pré-duplication. L'implémentation permet également de facilement mettre à jour la reconstruction référence, à chaque progression de l'état de l'art.

3.2 Article

Contribution au travail présenté

J'ai conçu et implémenté la méthodologie qui a permis d'établir la carte de paralogie et réalisé les analyses présentées dans le manuscrit. J'ai participé au processus de reconstruction des arbres de gènes initiaux que j'ai par la suite améliorés à travers l'application de SCORPiOs. J'ai également écrit la première version du manuscrit et contribué à sa remise en forme pour aboutir à la version présentée.

High-resolution map of ancient tetraploidy across 74 teleost fish genomes

Elise Parey¹, Alexandra Louis¹, Jérôme Monfort², Yann Guiguen², Hugues Roest Crollius^{1*},
Camille Berthelot^{1*}

¹ Institut de Biologie de l'Ecole normale supérieure (IBENS), Département de Biologie, Ecole normale supérieure, CNRS, INSERM, Université PSL, 75005 Paris, France.

² INRAE, LPGP, 35000, Rennes, France.

* Corresponding authors

Abstract

The impressive species richness and diversity of the teleost clade makes it an outstanding group to study processes of evolutionary, ecological and functional genomics. Yet, despite a growing number of well-established model species, the full potential of the teleost clade is still far from being realized. As a legacy of the teleost whole genome duplication dated 320 million years ago, teleosts are characterized by an elevated level of genomic redundancy. Consequently, the precise identification of orthologous genes and genomic regions across teleost species is particularly challenging, which in turn has hindered large-scale comparative studies. Here, we combine tailored gene phylogeny methodology together with the state-of-the-art ancestral teleost genome reconstruction to establish the first high resolution cartography of WGD-duplicated regions across 74 teleost fish species. We show that this fine-grained paralogy map represents a unique, robust and reliable resource for fish genomics. In the wake of massive sequencing projects that will lead to an explosion in the number of available fish genomes, the paralogy map represents a valuable lever to conduct comparative studies across teleosts.

Introduction

Since the publication of the first fish genome sequence in 2002 (Aparicio et al. 2002), fish genomics has massively contributed to our understanding of vertebrate genome function and evolution. As an early-diverging vertebrate clade, fish are at an ideal phylogenetic position to study deeply rooting vertebrate features. Notably, the conservation of vertebrate regulatory circuits and developmental pathways has led zebrafish, medaka and - to a lesser extent - platyfish, to become model species for human diseases (Wittbrodt et al. 2002; Lieschke and Currie 2007; Scharl 2014). In addition, fish have become popular in the fields of evolutionary and physiological genomics, providing insight into as various processes as environmental adaptation, species diversification, social behavior or sex determination (Rittschof et al. 2014; Capel 2017; Salzburger 2018; Kim et al. 2019; Xie et al. 2019; Greenway et al. 2020). Yet, while they represent roughly half of all vertebrate species, only around two hundred fish genomes have been sequenced to date. Many more fish genomes are expected to become available in the next years (Rhie et al. 2020), bringing up the necessity to design frameworks that will effectively permit to expand functional annotations to non-model species.

The vast majority of fish species descend from an ancient round of whole genome duplication (WGD), dated 320 Mya (Jaillon et al. 2004). This dramatic evolutionary event, referred to as the teleost specific genome duplication (TGD), generated a copy for all genes present in the teleost ancestor. Conceivably, the TGD has left a significant imprint on modern teleost genomes. While most anciently duplicated genes have returned to a single-copy state, an important fraction of teleost genes remain in two copies (26% of zebrafish genes are TGD-duplicates; (Howe et al. 2013)). Diverse factors govern the fate of duplicated genes: there is evidence that TGD duplicates have been involved in the evolution of functional innovations (Moriyama et al. 2016), but it remains unclear to which extent differential gene retention has fueled the impressive phenotypic diversity of the teleost clade. From a methodological viewpoint, this redundancy in teleost gene sets complexifies the process of transferring functional annotations across species. Genes may exist in either one or two copies across teleosts, and correctly identifying orthologous 'a' and 'b' copies in multiple species is non-trivial. In fact, erroneous assignments are an acknowledged source of inconsistencies in teleost gene names. For instance, due to opposing orthology predictions with curated zebrafish genes (Ruzicka et al. 2019), the same stickleback gene has been alternatively named *cxcl12a* and *cxcl12b* in subsequent releases of the Ensembl Compara database (v90 and v91) (Herrero et al. 2016).

Importantly, the redundancy in fish genomes can be appreciated at the macrosyntenic level, where remnants of ancestrally duplicated chromosomes form runs of large duplicated regions (Postlethwait et al. 2000; Taylor et al. 2003; Jaillon et al. 2004). These pairs of sister regions have been called homeologs or double-conserved syntenic regions (DCS). The

precise identification and delimitation of fish DCS regions remains elusive, due to the challenge that WGDs pose to both gene tree and ancestral genome reconstruction methodologies. Previous DCS characterizations were therefore limited to regions of highly-conserved gene order and few species (~29% of genes identified in DCS across the largest studied dataset of 8 fish species; (Conant 2020)). A fine-grained cartography of fish duplicated regions is desirable, not only as an additional support for orthology predictions, but also as a basis for gene and genome evolution hypothesis formulation. We have previously developed SCORPiOs, a method to build improved gene trees in the presence of whole genome duplications (Parey et al. 2020). Here, we apply SCORPiOs to retrace the evolutionary history of the genes and chromosomes of 74 teleost species. Further, we take advantage of the latest ancestral reconstruction of the pre-TGD ancestral genome (Nakatani and McLysaght 2017), together with these high-quality genes trees, to build the first large-scale paralogy map of teleost genomes. This paralogy map links the location of TGD-duplicated regions within and across species and predicts their ancestral duplicated chromosome of origin. The paralogy map encompasses 83% of the 74 species genes and provides rigorous evolutionary annotations for teleost genes and genomes.

Results

Construction of the set of teleost gene trees

We collected a dataset of 101 vertebrate genomes, which includes 74 teleost fish, 2 non-teleost fish (bowfin and spotted gar, which did not undergo the TGD duplication event), 20 other vertebrates (including 6 mammals) and 5 non-vertebrate outgroups (Supplementary Figure S1; Supplementary Table S1). We used the TreeBeST pipeline, as developed by the Ensembl Compara database, to build a starting set of gene trees (Methods; (Vilella et al. 2009; Herrero et al. 2016)). We then applied SCORPiOs to improve those gene trees, using the bowfin and the spotted gar as outgroups to the TGD. Briefly, SCORPiOs leverages additional information from syntenic genes to distinguish WGD-descended orthologs from paralogs, and accurately position WGD duplication nodes in genes trees (Parey et al. 2020). Out of 17,493 teleost gene subtrees, 8,144 were identified by SCORPiOs as synteny-inconsistent (47%). After five iterative rounds of correction, a total of 5,611 subtrees (32% of total subtrees) were corrected. We note that the corrected-to-inconsistent tree ratio (69%) is comparable to our previous application of SCORPiOs to fish data. Similarly, the proportion of errors in initial sequence-based gene trees is in line with our previous application of SCORPiOs to a dataset of 47 teleost species (Parey et al. 2020). These 26,692 WGD-aware teleost gene trees predict that the ancestral genome of teleost fish contained 46,206 genes after the duplication event, which is in line with the latest Ensembl release (49,255 ancestral teleost genes in Ensembl v100; Methods).

Identification of WGD-duplicated regions across 74 teleost species

Teleost fish genomes can be pictured as mosaics of duplicated DCS regions, that were formed by rearrangements of ancestrally duplicated chromosomes (Figure 1A). Long-standing efforts have been made towards the reconstruction of the ancestral pre-duplication teleost karyotype (Jaillon et al. 2004; Kasahara et al. 2007; Nakatani and McLysaght 2017). The state-of-the-art reference predicts that the pre-duplication teleost karyotype comprised 13 chromosomes (Nakatani and McLysaght 2017). Further, the authors delineated the genomic regions that descend from these ancestral chromosomes in four modern reference species (zebrafish, tetraodon, stickleback and medaka). This reconstruction encompasses 70-90% of each genome, but does not address the fine grained evolutionary history of genes and sequences within those megabase-scale genomic blocks. Additionally, while this reconstruction identifies pairs of regions descended from a pre-duplication ancestral chromosome in each of the four species, it does not identify orthologous regions that descend from the same post-duplication chromosome across those species.

We combined this pre-TGD ancestral genome reconstruction with homology relationships from our WGD-aware teleost gene trees to define genes and regions that descend from sister duplicated chromosomes (Methods; Figure 1). First, we transformed the segmentation of the four reference species from 13 pre-duplication chromosomes (1, 2, ... ,13), to 26 post-duplication chromosomes (1a, 1b, 2a, 2b, ..., 13a, 13b). For each pre-duplication chromosome, we identified all descendant regions in each reference species. Starting from the largest region, we used an iterative process to assign regions to the `a` or `b` post-duplication chromosome copies based on their fraction of TGD paralogs shared with blocks already assigned to either `b` or `a`, until remaining blocks shared fewer than 5% paralogs with any assigned region (Figure 1B). Chromosome copies annotated as `a` or `b` were then homogenized between all four species, using orthology relationships from the gene trees (Figure 1C). This results in a four-species paralogy map, which serves as a basis for the annotation of all other teleost species.

We next propagated these paralogy annotations from reference species to all teleost species in our dataset, using gene homologies from the phylogenetic gene trees. Each gene in a non-reference species is assigned a post-duplication chromosome of origin based on its orthologous genes in the reference species (Figure 1D). For 1,560 gene trees (8%), inconsistencies arose between assignments of the reference genes, and a majority vote was used. This process results in a 74-species paralogy map with genes annotated to post-duplication chromosomes, along with fully resolved orthology and paralogy links between all included species (Figure 1E). This global paralogy map integrates 69% to 90% of each species genome (Supplementary Figure S2), and greatly improves upon the state-of-the-art both in terms of species and genomic coverage.

We further assessed discrepancies between pre-TGD ancestral chromosome assignments in our paralogy map and in the reconstruction by (Nakatani and McLysaght 2017). Because the ancestral reconstruction provides the basis for the paralogy map, both should be globally congruent, unless paralogies in our gene trees strongly disagree with ancestral chromosomes assignments. We found that zebrafish genes were the most affected by re-assignments with 18% of ancestral chromosome reassignment in our paralogy map for zebrafish genes, versus 3-4% in the other 3 reference species. This likely reflects small-scale rearrangements in zebrafish, captured in our individual gene trees but missed by the macrosyntenic approach of (Nakatani and McLysaght 2017). Alternatively, this could be due to noise in our gene trees, affecting zebrafish genes due to a possible bias in taxon sampling. To explore this possibility, we analysed SCORPiOs correction results to identify gene trees that remain synteny-inconsistent after correction, which potentially contain errors and may lead to erroneous ancestral chromosome assignments. We observed that zebrafish genes reassigned to a different ancestral chromosome in the paralogy map are not over-

represented in synteny-inconsistent trees (13% of reassigned zebrafish genes are in dubious trees, while the global proportion of zebrafish genes in dubious trees is of 18%). This suggests that zebrafish gene ancestral chromosome reassignments are globally well-supported by our gene trees.

The paralogy map is a robust and flexible resource for fish genomics

We used random noise simulations to demonstrate that the paralogy map is robust to potential uncertainty or errors in the original ancestral genome reconstruction. The delineation of genomic regions descended from each pre-duplication chromosome from (Nakatani and McLysaght 2017) is based on the definition of conserved synteny blocks and corresponding breakpoints between non-teleost and teleost genomes. Because breakpoint locations are typically challenging to predict - as also attested by lower posterior probabilities close to breakpoints in (Nakatani and McLysaght 2017) - the regions boundaries can vary in precision. We shifted the positions of these boundaries with increasingly large errors, mimicking situations where up to 25% of genes change pre-duplication chromosome assignments in each of the reference species (Methods). We found that even large errors in region boundaries did not majorly affect the final paralogy map, with only 10% of genes changing ancestral chromosome assignments at the highest noise settings (Supplementary Figure S3). This is because region boundaries correspond to chromosomal rearrangements, which have largely occurred independently in the four reference teleost species. As a result, genes assigned to an incorrect ancestral chromosome due to an imprecise boundary in one species are typically corrected by the information provided by the other reference species. Additionally, we report that using SCORPiOs to improve gene trees had a decisive impact on the establishment of the paralogy map, enabling the inclusion of a significantly larger fraction of teleost genes (61% vs 83% Supplementary Figure S4). The teleost paralogy map represents therefore a reliable, comprehensive and robust resource for fish genomics. The elevated coverage of the paralogy map in terms of species number and fine-grained orthology relations represents a substantial improvement over previous references.

The paralogy map provides insights into teleost genome evolution at different resolution levels. At the karyotype scale, the 74 teleost genomes are segmented with respect to the 26 post-duplication ancestral chromosomes (Figure 2A). This view expands previous observations regarding teleost structure evolution made by Nakatani and McLysaght, while allowing to pinpoint rearrangement affecting a specific copy of duplicated chromosomes. For instance, zebrafish chromosomes 18 and 5 result from fusions of ancestral chromosomes deriving from the same homeologous pairs: 10a and 10b, and 9a and 9b, respectively. These fusions were also previously suggested, although with a significantly lower resolution in

resolving zebrafish chromosome ancestry (Kasahara et al. 2007). We note that propagation of the evolutionary annotations to arapaima, a basally branching teleost (order Osteoglossiform), also paints large genomic duplicated blocks, which can be viewed as an independent validation of the initial ancestral reconstruction, as the arapaima genome was not used in this previous study (Nakatani and McLysaght 2017). Importantly, osteoglossiform chromosome evolution remains poorly documented. We report here that post-TGD chromosome rearrangements appear vastly independent in osteoglossiforms and clupeocephala, with no strict one-to-one orthologous chromosomes between arapaima and medaka. Further, 3 of 7 well-documented major rearrangements in the pre-clupeocephalan lineage (Kasahara et al. 2007; Nakatani and McLysaght 2017) are unambiguously absent in arapaima and the 2 other osteoglossiforms in our dataset, contrary to assumptions made by (Bian et al. 2016) (Figure 2A-B, Supplementary Table S1, Supplementary Figures S5-9).

The paralogy map can also be visualized at the chromosome level, to assess the dispersion of genes from ancestrally duplicated chromosomes into extant genomes. For instance, the ancestral homeologous pair 2a-2b has remained strikingly conserved across species (Figure 2C), while other ancestral chromosomes, such as homeologous pair 8a-8b have been extensively rearranged since the TGD event.

The paralogy map enhances teleost gene and genome annotations

Next, we investigated how the paralogy map can help improve crucial resources that are widely used in fish evolutionary and ecological genomics. The Zebrafish Information Network (ZFIN) provides manually curated, high-quality reference annotations for zebrafish genes and establishes rigorous conventions for gene naming (Ruzicka et al. 2019). These gene names are then reverberated to orthologous genes in other teleost genomes, providing the basis of the entire teleost gene nomenclature and functional annotation. In an effort to convey evolutionary meaning within the gene names, zebrafish paralogs descended from the TGD are identified with an 'a' and 'b' suffix. In addition, ZFIN guidelines recommend to name neighbor genes with the same suffix, to reflect that these genes correspond to a continuous syntenic block inherited since the TGD, sometimes called syntelogs. This aspiration is however difficult to implement in practice, in the absence of a high-resolution map of zebrafish duplicated regions and their ancestral chromosomes of origin. To assess the consistency in consecutive zebrafish gene names, we extracted and compared zebrafish 'a' and 'b' suffixes with our duplicated block annotations (Figure 3). We find that current zebrafish 'a' and 'b' suffixes are not consistent across large regions (Figure 3A). Using the paralogy map as a guide, we estimate that 43% of 'a' and 'b' gene suffixes should be reassigned in order to reflect the shared history of syntenic genes (Methods, Figure 3B). Additionally, the ZFIN nomenclature currently does not impart a suffix to singleton genes,

which correspond to TGD-duplicated genes where one of the copies was eventually lost (26% of annotated zebrafish genes in ZFIN, 83% in the paralogy map, Figure 3D). How ancient tetraploids return to a mostly diploid state is an active area of research, where distinguishing ancestral 'a' and 'b' singleton genes is essential (Inoue et al. 2015; Robertson et al. 2017; Conant 2020; Simakov et al. 2020). As the paralogy map integrates the majority of genes, it opens the possibility to systematically extend suffix annotations to singletons and formally identify which ancestral copies have been retained in zebrafish and other teleosts (Figure 3C). Altogether, the paralogy map is a key resource to include biologically meaningful, historically accurate insight into reference gene annotations and support further investigations of teleost genome evolution.

Discussion

As sequencing technologies progress, data analysis and integration is becoming the major roadblock to extract the rich information that comparative genomics can provide on species evolution and genome function. Teleost fish have a long-standing history as tractable model species for vertebrate development and human disease (Lieschke and Currie 2007; Schartl 2014), and have contributed major breakthroughs in ecological, evolutionary and functional genomics over the years (Braasch et al. 2016; Capel 2017; Xie et al. 2019). As such, they have become a high priority taxon for several large-scale projects aiming to extend phylogenetic coverage of vertebrate genome resources (Fan et al. 2020; Rhie et al. 2020). However, functional annotations of these genomes lack behind those of intensely studied model organisms such as human and mouse, where diverse datasets including tissue-specific transcriptomes, epigenomic information and knock-out data are available to dissect how the genome is deployed and how genes contribute to phenotypes. Because these datasets remain costly and time-consuming to generate, it is not expected that non-model teleost genomes will benefit from deep functional characterization in the near future. Being able to transfer functional annotations from one species to another is therefore all the more crucial to exploit their shared genomic resources. While this solution is imperfect, there is extensive evidence that orthologous genes and regions often maintain conserved functions and characteristics across long evolutionary scales, so that comparative approaches can supply a rich source of functional annotations where none exist (Altenhoff et al. 2012; Villar et al. 2015; Polychronopoulos et al. 2017). Transferring functional annotations is however particularly difficult in fish, where many regions are duplicated remnants of the teleost-specific whole genome duplication and orthologous copies are difficult to identify across species. We expect that the genome-scale, clade-wide paralogy and orthology resources we provide here will allow functional annotations to be extrapolated across species and propel studies of both evolution and functional characterization of teleost genomes.

It is important to note that the teleost paralogy map comes with a number of limitations that directly stem from the way it was built. First, the assignments to ancestral chromosomes before the TGD are only as good as the state-of-the-art ancestral genome reconstruction. There is a general consensus that the ancestral teleost genome contained 13 chromosomes (Kasahara et al. 2007; Nakatani and McLysaght 2017) – however, the precise delineation of genes descending from each of those 13 groups may be subject to modifications as the field evolves, which may lead to updates in the ancestral attributions of the regions in the paralogy map. Second, we do not know at this time whether the teleost whole-genome duplication corresponds to an ancestral autotetraploidization (a genome self-doubling typically resulting from errors during meiosis) or an allotetraploidization event (resulting from the fusion of two parental genomes from related species) (Stebbins 1947; Garsmeur et al. 2014). The annotations of ‘a’ and ‘b’ chromosomal copies in the paralogy map should not be interpreted as two distinct subgenomes where all ‘a’ chromosomes (or ‘b’) descend from a single parental genome. While this distinction is irrelevant for autotetraploidization events where there are no distinct subgenomes, further investigations into the mode of tetraploidization—possibly enabled by this paralogy map itself – may reveal that fish genomes are ancient allotetraploids. In this case, the ‘a’ and ‘b’ labels may require updates to reflect advancing knowledge on the parental subgenomes structures. Finally, while we show that the paralogy map is resilient to species-specific errors in ancestral chromosome or gene orthology groups assignments due to the redundant information provided by multiple species, the map is ultimately based on gene family tree models, which are sometimes inaccurate or incomplete (Hahn 2007; Som 2015). Inaccurate trees may result in local misspecifications of gene paralogs or ancestral assignments in the paralogy map. We have previously shown that SCORPiOs significantly improves gene tree accuracy after a WGD event, and we found only few discrepancies between megabase-scale regions predicted to descend from the same ancestral chromosome from (Nakatani and McLysaght 2017) and paralogy relationships predicted at gene-to-gene resolution by our gene trees. This suggests that the map is generally accurate. However, we note that SCORPiOs flagged 2,832 gene trees where sequence identity relationships are inconsistent with their local syntenic context, which may represent areas where the map is either less reliable or biologically less informative. There are legitimate reasons why gene sequence and gene locus may report contradictory information: notably, after whole-genome duplications, some paralogous genomic regions can maintain meiotic recombination, essentially remaining tetraploids and exchanging alleles over millions of years. In this evolutionary scenario, gene loci are duplicated at the WGD event, but gene sequences only start diverging once recombination is suppressed, so that locus and sequence histories do not match. The paralogy map attempts to reconstruct the history of loci rather than gene sequences, and may help in the future to shed light on the

poorly documented process through which teleost fish genomes returned to a largely diploid state.

Methods

Libraries and packages

Scripts to build the paralogy map were written in Python 3.6.8 and assembled together in a workflow with Snakemake (version 5.13.0). The ete3 package (version 3.1.1) was used for phylogenetic gene tree manipulation and drawing. Other python package dependencies used for plots and analyses include matplotlib (version 3.1.1), seaborn (version 0.9.0) and numpy (version 1.18.4). Chromosomes painting (Figure 2A and Figure 3) and synteny comparisons (Figure 2B, Figure 2C, Supplementary Figures S5-S9) were drawn with the RIdeogram R package (Hao et al. 2020).

Phylogenetic gene trees

Initial gene trees were built using the Ensembl Compara pipeline (Vilella et al. 2009). Briefly, starting from the sets of proteins derived from the longest transcripts, we performed an all-against-all BLASTP+ (Altschul et al. 1990), with the following parameters '-seg no -max_hsps 1 -use_sw_tback -evaluate 1e-5'. We then performed clustering with hcluster_sg to define gene families, using parameters '-m 750 -w 0 -s 0.34 -O'. We built multiple alignments using M-Coffee (Wallace et al. 2006), with the command 't-coffee -type=PROTEIN -method mafftgins_msa, muscle_msa, kalign_msa, t_coffee_msa -mode=mcoffee'. Next, we conducted phylogenetic trees construction and reconciliation with TreeBeST, using default parameters (Vilella et al. 2009). Because TreeBeST is systematically biased towards inferring gene duplication nodes that are overly old (Hahn 2007), we pre-processed gene trees to edit nodes with a very low duplication confidence score (score < 0.1), using ProfileNJ (Noutahi et al. 2016). Finally, we ran SCORPIOs (version [v1.1.0](#)) to account for several whole genome duplication events in the species phylogeny and correct gene trees accordingly: the teleost 3R WGD, using bowfin and gar as outgroups; the carps 4R WGD, using the zebrafish as outgroup; and the salmonids 4R WGD, using the Northern pike as outgroup.

Ancestral gene statistics

We calculated the predicted number of genes in the post-duplication ancestral teleost genome using our set of 26,692 gene trees, and compared to 60,447 state-of-the art gene trees stored in Ensembl Compara v100. This ancestral gene number is an indirect but accurate approximation of the quality of inferred gene trees, since the major challenge is to accurately position duplications at this ancestral node. TreeBeST phylogeny-reconciled gene trees were recursively browsed and gene copies inferred in the teleost ancestor (*Osteoglossocephalai*) were collected.

Ancestral chromosome annotations in reference species

Within the 4 reference species, we identified WGD sister regions as sharing a high fraction of ohnologs, i.e. duplicated genes descended from the WGD duplication node (Figure 1B). We grouped regions descended from the same pre-duplication ancestral chromosome, and used an iterative process to annotate pairs of regions into internally consistent sets of 'a' and 'b' post-duplication sister regions. For each ancestral chromosome, we started from the largest descendant region and arbitrarily defined it as 'a'. All regions sharing 50% ohnologs or more with this 'a' region are identified its 'b' paralogous region(s). Additional rounds of sister regions search were then conducted to spread the 'a' and 'b' annotations in a stepwise manner to all regions descended from this ancestral chromosome. We decreased the required ohnolog fraction at each new round and finally stopped the search when remaining blocks had less than 5% shared ohnologs with the already annotated regions. Since this identification of duplicated regions was performed independently in each of the 4 species, 'a' and 'b' annotations were homogenized to ensure consistency across species (Figure 1C). The homogenization step uses orthology relationships from the gene trees and stickleback as a guide species: 'a' and 'b' annotations were switched in other species if 'a' segments shared more orthologous genes with the 'b' region of stickleback.

Simulation of ancestral chromosome boundary shifts

Segmentation of the four teleost species with respect to the 13 ancestral chromosomes were extracted from (Nakatani and McLysaght 2017) and genomic interval coordinates converted to lists of genes. All genomes were then reduced to ordered lists of genes. To simulate shifts in interval boundaries, we randomly drew new boundaries in the vicinity of their original location according to a Gaussian distribution with σ varying in [5, 10, 25, 50, 75, 100] genes. These boundary shifts were independently generated for each of the 4 reference species, with $n=100$ random noise simulations for each σ value and each reference species. In total, simulations generated 600 noisy input datasets that we fed to the paralogy map pipeline in order to assess its robustness to noise. Each of the 600 produced outputs were then compared to our paralogy map, by counting the proportion of orthogroups with a reassigned ancestral chromosome (Supplementary Figure S3).

ZFIN gene names

Zebrafish ZFIN gene names were extracted using Biomart from the Ensembl database (version 95). We extracted the last letter of gene names, which represents 'a' and 'b' ohnologous copy annotations. We then computed the minimal number of 'a' and 'b' ZFIN gene name reassignments that would be necessary to be consistent with the paralogy map. In the paralogy map, 'a' and 'b' are arbitrarily assigned to homeologous chromosomes, i.e.

genes descended from chromosomes 1a and 1b could be swapped to 1b and 1a. In order to not artificially overestimate discordances, we first swapped such arbitrary 'a' and 'b' annotations to minimize differences with ZFIN. Finally, we counted the remaining number of 'a' and 'b' disagreement for zebrafish genes in the paralogy map that were also annotated in ZFIN.

Figures

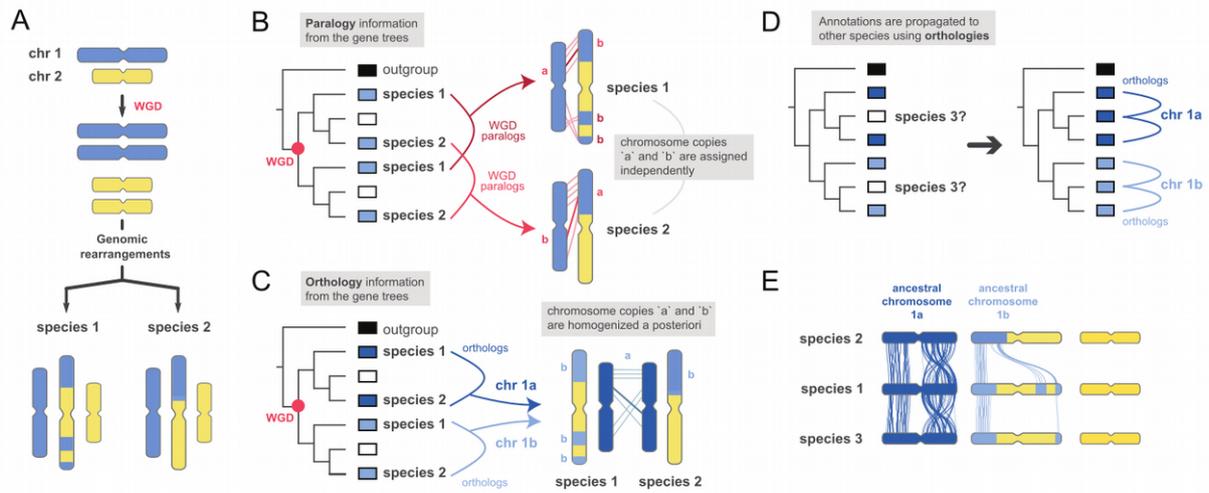


Figure 1: Illustration of the key steps in the Paralogy Map workflow. **A.** Schematic representation of karyotype histories for two species with a common whole-genome duplication followed by a chromosomal fusion event. **B.** Identification of WGD-duplicated regions using paralogy relationships inferred from gene trees (in red). This identification is performed in each species independently. **C.** Identification of orthologous regions across species inferred from gene trees (in blue). This information is used to homogenize 'a' and 'b' ancestral chromosome assignments across species. **D.** Propagation of duplicated regions annotations to a non-reference species. Here annotation from reference species 1 and 2 are propagated to species 3 through gene orthologies. **E.** Schematic representation of the paralogy map. Dark blue and light blue regions correspond to WGD-inherited paralogy regions across all species. Links represent orthologous genes.

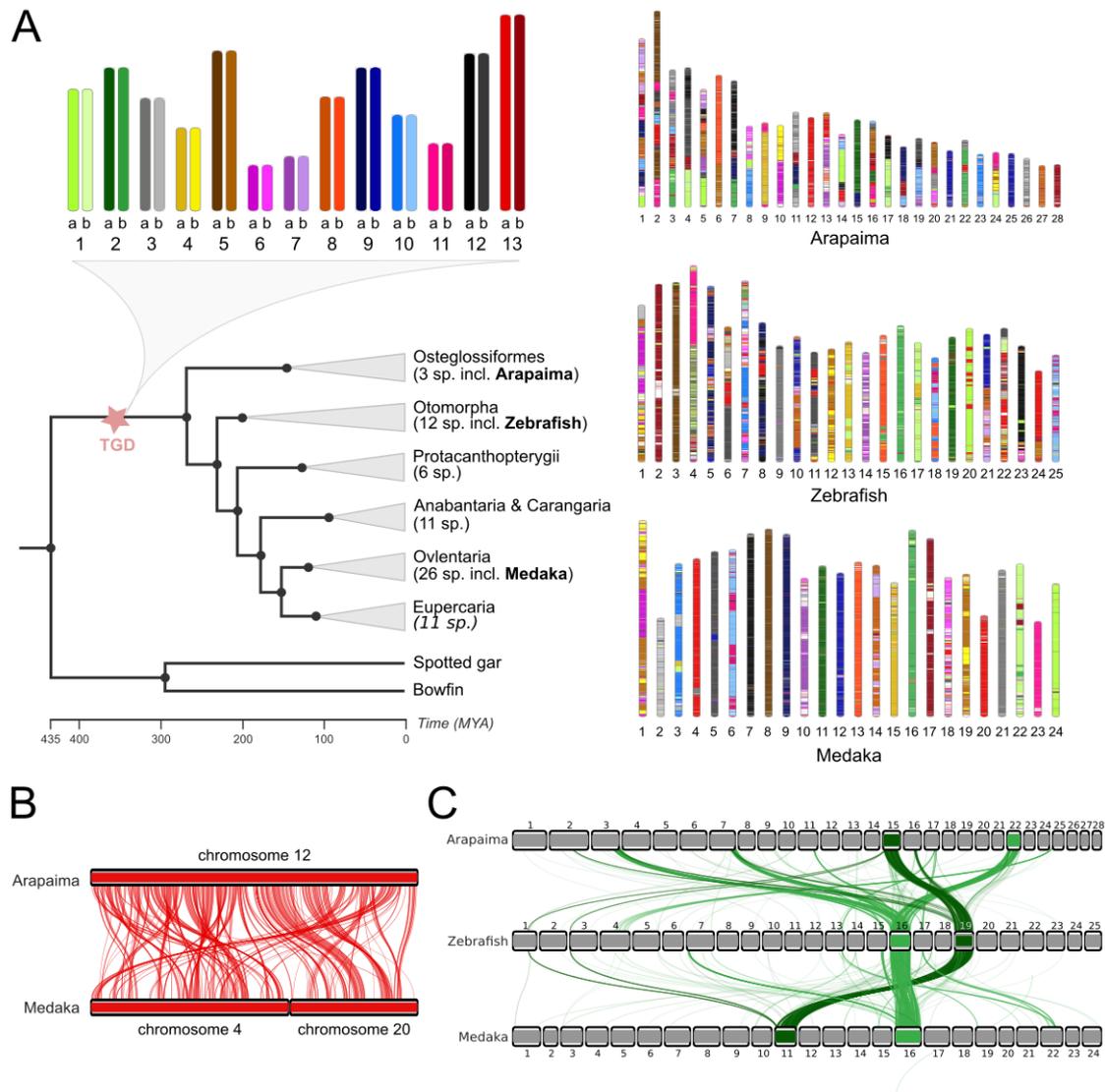


Figure 2: Ancestral chromosome annotations and paralogy map across 74 teleost species. **A.** Simplified teleost phylogeny with all major groups represented in this study. Top inset: inferred ancestral karyotype after the teleost whole-genome duplication (TGD). Right: paralogy map visualized at the karyotype level in 3 selected teleost species. Genomes are segmented according to post-duplication chromosomes (1a, 1b, ..., 13a, 13b). **B.** Comparison of arapaima chromosome 12 and medaka chromosome 4 and 20, with orthology relationships inferred from the paralogy map. Medaka chromosome 4 and 20 descend from ancestral chromosome 13a, but orthologous genes remain linked on a single chromosome in arapaima, suggesting that a fission event happened in the medaka lineage after divergence from Osteoglossiformes. **C.** Dispersion of genes from the homeologous pair 2a-2b : this ancestral chromosome pair is highly conserved, with an almost 1:1 mapping in arapaima, zebrafish and medaka.

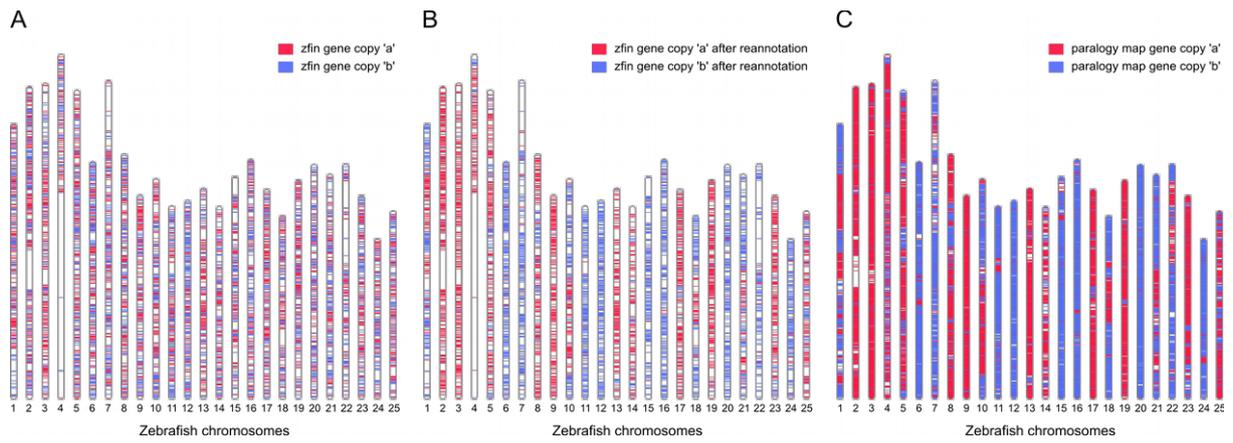


Figure 3: Zebrafish gene names are not evolutionary-consistent. **A.** Visualization of Zebrafish ZFIN 'a' and 'b' copies on the genome. **B.** Re-annotation of ZFIN zebrafish gene 'a' and 'b' copies using evolutionary information from the paralogy map. **C.** Full re-annotation of zebrafish 'a' and 'b' copies using the paralogy map (83% of zebrafish genes annotated, including genes without a WGD paralog).

References

- Altenhoff AM, Studer RA, Robinson-Rechavi M, Dessimoz C. 2012. Resolving the ortholog conjecture: orthologs tend to be weakly, but significantly, more similar in function than paralogs. *PLoS Comput. Biol.* 8:e1002514.
- Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *J. Mol. Biol.* [Internet] 215:403–410. Available from: <http://www.sciencedirect.com/science/article/pii/S0022283605803602>
- Aparicio S, Chapman J, Stupka E, Putnam N, Chia J-M, Dehal P, Christoffels A, Rash S, Hoon S, Smit A, et al. 2002. Whole-genome shotgun assembly and analysis of the genome of *Fugu rubripes*. *Science* 297:1301–1310.
- Bian C, Hu Y, Ravi V, Kuznetsova IS, Shen X, Mu X, Sun Y, You X, Li J, Li X, et al. 2016. The Asian arowana (*Scleropages formosus*) genome provides new insights into the evolution of an early lineage of teleosts. *Sci. Rep.* [Internet] 6:24501. Available from: <https://www.nature.com/articles/srep24501>
- Braasch I, Gehrke AR, Smith JJ, Kawasaki K, Manousaki T, Pasquier J, Amores A, Desvignes T, Batzel P, Catchen J, et al. 2016. The spotted gar genome illuminates vertebrate evolution and facilitates human-teleost comparisons. *Nat. Genet.* [Internet] 48:427–437. Available from: <https://www.nature.com/articles/ng.3526>
- Capel B. 2017. Vertebrate sex determination: evolutionary plasticity of a fundamental switch. *Nat. Rev. Genet.* [Internet] 18:675–689. Available from: <http://www.nature.com/articles/nrg.2017.60>
- Conant GC. 2020. The lasting after-effects of an ancient polyploidy on the genomes of teleosts. *PLOS ONE* [Internet] 15:e0231356. Available from: <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0231356>
- Fan G, Song Y, Yang L, Huang X, Zhang S, Zhang M, Yang X, Chang Y, Zhang H, Li Y, et al. 2020. Initial data release and announcement of the 10,000 Fish Genomes Project (Fish10K). *GigaScience* [Internet] 9. Available from: <https://academic.oup.com/gigascience/article/9/8/giaa080/5893976>
- Garsmeur O, Schnable JC, Almeida A, Jourda C, D'Hont A, Freeling M. 2014. Two Evolutionarily Distinct Classes of Paleopolyploidy. *Mol. Biol. Evol.* [Internet] 31:448–454. Available from: <https://academic.oup.com/mbe/article/31/2/448/1001232>

Greenway R, Barts N, Henpita C, Brown AP, Rodriguez LA, Peña CMR, Arndt S, Lau GY, Murphy MP, Wu L, et al. 2020. Convergent evolution of conserved mitochondrial pathways underlies repeated adaptation to extreme environments. *Proc. Natl. Acad. Sci.* [Internet] 117:16424–16430. Available from: <https://www.pnas.org/content/117/28/16424>

Hahn MW. 2007. Bias in phylogenetic tree reconciliation methods: implications for vertebrate genome evolution. *Genome Biol.* [Internet] 8:R141. Available from: <https://doi.org/10.1186/gb-2007-8-7-r141>

Hao Z, Lv D, Ge Y, Shi J, Weijers D, Yu G, Chen J. 2020. RIdeogram: drawing SVG graphics to visualize and map genome-wide data on the idiograms. *PeerJ Comput. Sci.* [Internet] 6:e251. Available from: <https://peerj.com/articles/cs-251>

Herrero J, Muffato M, Beal K, Fitzgerald S, Gordon L, Pignatelli M, Vilella AJ, Searle SMJ, Amode R, Brent S, et al. 2016. Ensembl comparative genomics resources. *Database J. Biol. Databases Curation* [Internet] 2016. Available from: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4761110/>

Howe K, Clark MD, Torroja CF, Torrance J, Berthelot C, Muffato M, Collins JE, Humphray S, McLaren K, Matthews L, et al. 2013. The zebrafish reference genome sequence and its relationship to the human genome. *Nature* [Internet] 496:498–503. Available from: <https://www.nature.com/articles/nature12111>

Inoue J, Sato Y, Sinclair R, Tsukamoto K, Nishida M. 2015. Rapid genome reshaping by multiple-gene loss after whole-genome duplication in teleost fish suggested by mathematical modeling. *Proc. Natl. Acad. Sci. U. S. A.* [Internet] 112:14918–14923. Available from: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4672829/>

Jaillon O, Aury J-M, Brunet F, Petit J-L, Stange-Thomann N, Mauceli E, Bouneau L, Fischer C, Ozouf-Costaz C, Bernot A, et al. 2004. Genome duplication in the teleost fish *Tetraodon nigroviridis* reveals the early vertebrate proto-karyotype. *Nature* [Internet] 431:946–957. Available from: <https://www.nature.com/articles/nature03025>

Kasahara M, Naruse K, Sasaki S, Nakatani Y, Qu W, Ahsan B, Yamada T, Nagayasu Y, Doi K, Kasai Y, et al. 2007. The medaka draft genome and insights into vertebrate genome evolution. *Nature* [Internet] 447:714–719. Available from: <https://www.nature.com/articles/nature05846>

Kim B-M, Amores A, Kang S, Ahn D-H, Kim J-H, Kim I-C, Lee JH, Lee SG, Lee H, Lee J, et al. 2019. Antarctic blackfin icefish genome reveals adaptations to extreme environments.

Nat. Ecol. Evol. [Internet] 3:469–478. Available from: <https://www.nature.com/articles/s41559-019-0812-7>

Lieschke GJ, Currie PD. 2007. Animal models of human disease: zebrafish swim into view. *Nat. Rev. Genet.* [Internet] 8:353–367. Available from: <https://www.nature.com/articles/nrg2091>

Moriyama Y, Ito F, Takeda H, Yano T, Okabe M, Kuraku S, Keeley FW, Koshiba-Takeuchi K. 2016. Evolution of the fish heart by sub/neofunctionalization of an elastin gene. *Nat. Commun.* [Internet] 7:1–10. Available from: <https://www.nature.com/articles/ncomms10397>

Nakatani Y, McLysaght A. 2017. Genomes as documents of evolutionary history: a probabilistic macrosynteny model for the reconstruction of ancestral genomes. *Bioinformatics* [Internet] 33:i369–i378. Available from: <https://academic.oup.com/bioinformatics/article/33/14/i369/3953974>

Noutahi E, Semeria M, Lafond M, Seguin J, Boussau B, Guéguen L, El-Mabrouk N, Tannier E. 2016. Efficient Gene Tree Correction Guided by Genome Evolution. *PLOS ONE* [Internet] 11:e0159559. Available from: <http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0159559>

Parey E, Louis A, Cabau C, Guiguen Y, Roest Crollius H, Berthelot C. Synteny-Guided Resolution of Gene Trees Clarifies the Functional Impact of Whole-Genome Duplications. *Mol. Biol. Evol.* [Internet]. Available from: <https://academic.oup.com/mbe/advance-article/doi/10.1093/molbev/msaa149/5859632>

Polychronopoulos D, King JWD, Nash AJ, Tan G, Lenhard B. 2017. Conserved non-coding elements: developmental gene regulation meets genome organization. *Nucleic Acids Res.* [Internet] 45:12611–12624. Available from: <https://academic.oup.com/nar/article/45/22/12611/4599184>

Postlethwait JH, Woods IG, Ngo-Hazelett P, Yan YL, Kelly PD, Chu F, Huang H, Hill-Force A, Talbot WS. 2000. Zebrafish comparative genomics and the origins of vertebrate chromosomes. *Genome Res.* 10:1890–1902.

Rhie A, McCarthy SA, Fedrigo O, Damas J, Formenti G, Koren S, Uliano-Silva M, Chow W, Fungtammasan A, Gedman GL, et al. 2020. Towards complete and error-free genome assemblies of all vertebrate species. *bioRxiv* [Internet]:2020.05.22.110833. Available from: <https://www.biorxiv.org/content/10.1101/2020.05.22.110833v1>

Rittschof CC, Bukhari SA, Sloofman LG, Troy JM, Caetano-Anollés D, Cash-Ahmed A, Kent M, Lu X, Sanogo YO, Weisner PA, et al. 2014. Neuromolecular responses to social challenge: Common mechanisms across mouse, stickleback fish, and honey bee. *Proc. Natl. Acad. Sci. U. S. A.* [Internet] 111:17929–17934. Available from: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4273386/>

Robertson FM, Gundappa MK, Grammes F, Hvidsten TR, Redmond AK, Lien S, Martin SAM, Holland PWH, Sandve SR, Macqueen DJ. 2017. Lineage-specific rediploidization is a mechanism to explain time-lags between genome duplication and evolutionary diversification. *Genome Biol.* [Internet] 18:111. Available from: <https://doi.org/10.1186/s13059-017-1241-z>

Ruzicka L, Howe DG, Ramachandran S, Toro S, Van Slyke CE, Bradford YM, Eagle A, Fashena D, Frazer K, Kalita P, et al. 2019. The Zebrafish Information Network: new support for non-coding genes, richer Gene Ontology annotations and the Alliance of Genome Resources. *Nucleic Acids Res.* 47:D867–D873.

Salzburger W. 2018. Understanding explosive diversification through cichlid fish genomics. *Nat. Rev. Genet.* [Internet] 19:705–717. Available from: <https://www.nature.com/articles/s41576-018-0043-9>

Schartl M. 2014. Beyond the zebrafish: diverse fish species for modeling human disease. *Dis. Model. Mech.* [Internet] 7:181–192. Available from: <https://dmm.biologists.org/content/7/2/181>

Simakov O, Marlétaz F, Yue J-X, O’Connell B, Jenkins J, Brandt A, Calef R, Tung C-H, Huang T-K, Schmutz J, et al. 2020. Deeply conserved synteny resolves early events in vertebrate evolution. *Nat. Ecol. Evol.* [Internet] 4:820–830. Available from: <https://www.nature.com/articles/s41559-020-1156-z>

Som A. 2015. Causes, consequences and solutions of phylogenetic incongruence. *Brief. Bioinform.* [Internet] 16:536–548. Available from: <https://academic.oup.com/bib/article/16/3/536/243419>

Stebbins GL. 1947. Types of Polyploids: Their Classification and Significance. In: Demerec M, editor. *Advances in Genetics*. Vol. 1. Academic Press. p. 403–429. Available from: <http://www.sciencedirect.com/science/article/pii/S0065266008604903>

Taylor JS, Braasch I, Frickey T, Meyer A, Peer YV de. 2003. Genome Duplication, a Trait Shared by 22,000 Species of Ray-Finned Fish. *Genome Res.* [Internet] 13:382–390. Available from: <http://genome.cshlp.org/content/13/3/382>

Vilella AJ, Severin J, Ureta-Vidal A, Heng L, Durbin R, Birney E. 2009. EnsemblCompara GeneTrees: Complete, duplication-aware phylogenetic trees in vertebrates. *Genome Res.* [Internet] 19:327–335. Available from: <http://genome.cshlp.org/content/19/2/327>

Villar D, Berthelot C, Aldridge S, Rayner TF, Lukk M, Pignatelli M, Park TJ, Deaville R, Erichsen JT, Jasinska AJ, et al. 2015. Enhancer Evolution across 20 Mammalian Species. *Cell* [Internet] 160:554–566. Available from: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4313353/>

Wallace IM, O’Sullivan O, Higgins DG, Notredame C. 2006. M-Coffee: combining multiple sequence alignment methods with T-Coffee. *Nucleic Acids Res.* [Internet] 34:1692–1699. Available from: <https://academic.oup.com/nar/article/34/6/1692/2401531>

Wittbrodt J, Shima A, Schartl M. 2002. Medaka — a model organism from the far east. *Nat. Rev. Genet.* [Internet] 3:53–64. Available from: <https://www.nature.com/articles/nrg704>

Xie KT, Wang G, Thompson AC, Wucherpfennig JI, Reimchen TE, MacColl ADC, Schluter D, Bell MA, Vasquez KM, Kingsley DM. 2019. DNA fragility in the parallel evolution of pelvic reduction in stickleback fish. *Science* [Internet] 363:81–84. Available from: <https://science.sciencemag.org/content/363/6422/81>

High-resolution map of ancient tetraploidy across 74 teleost fish genomes

Elise Parey¹, Alexandra Louis¹, Jérôme Monfort², Yann Guiguen², Hugues Roest Crolius^{1*}, Camille Berthelot^{1*}

¹ Institut de Biologie de l'Ecole normale supérieure (IBENS), Département de Biologie, Ecole normale supérieure, CNRS, INSERM, Université PSL, 75005 Paris, France.

² INRAE, LPGP, 35000, Rennes, France.

* Corresponding authors

Supplementary Material

Supplementary Tables

Supplementary Table S1 : List of the 101 genome assemblies used in this study.

Supplementary Table S2 : Description of the 7 major pre-clupeocephalan chromosomal rearrangements.

Supplementary Figures

Supplementary Figure S1 : Species tree of the 101 genomes used in this study.

Supplementary Figure S2 : Species tree of the 74 teleost species and their percentage of genes annotated in the paralogy map.

Supplementary Figure S3 : Noise robustness of the paralogy map pipeline.

Supplementary Figure S4 : Impact of SCORPIOs on the establishment of the paralogy map.

Supplementary Figures S5-9 : Synteny comparisons of arapaima genome and, respectively, medaka chromosomes 1 and 19, 1, 10,14 and 6.

Species	Source	Data	N50 size (nb genes)
Acanthochromis polyacanthus	ensembl	ensembl V95	14
Amphilophus citrinellus	ensembl	ensembl V95	616
Amphiprion ocellaris	ensembl	ensembl V95	913
Amphiprion percula	ensembl	ensembl V95	41
Anabas testudineus	ensembl	ensembl V95	8
Anolis carolinensis	ensembl	ensembl V95	868
Astatotilapia calliptera	ensembl	ensembl V95	1187
Astyanax mexicanus	ensembl	ensembl V95	671
Bos taurus	ensembl	ensembl V95	845
Caenorhabditis elegans	ensembl	ensembl V95	3305
Canis lupus familiaris	ensembl	ensembl V95	606
Chrysemys pictabellii	ensembl	ensembl V95	58
Ciona intestinalis	ensembl	ensembl V95	816
Cynoglossus semilaevis	ensembl	ensembl V95	955
Cyprinodon variegatus	ensembl	ensembl V95	22
Danio rerio	ensembl	ensembl V95	1004
Drosophila melanogaster	ensembl	ensembl V95	2731
Eptatretus burgeri	ensembl	ensembl V95	19
Esox lucius	ensembl	ensembl V95	938
Fundulus heteroclitus	ensembl	ensembl V95	35
Gadus morhua	ensembl	ensembl V95	10
Gallus gallus	ensembl	ensembl V95	546
Gambusia affinis	ensembl	ensembl V95	229
Gasterosteus aculeatus	ensembl	ensembl V95	869
Haplochromis burtoni	ensembl	ensembl V95	46
Hippocampus comes	ensembl	ensembl V95	90
Homo sapiens	ensembl	ensembl V95	1022
Ictalurus punctatus	ensembl	ensembl V95	841
Kryptolebias marmoratus	ensembl	ensembl V95	89
Labrus bergylta	ensembl	ensembl V95	36
Latimeria chalumnae	ensembl	ensembl V95	11
Lepisosteus oculatus	ensembl	ensembl V95	773
Loxodonta africana	ensembl	ensembl V95	230
Mastacembelus armatus	ensembl	ensembl V95	1019
Maylandia zebra	ensembl	ensembl V95	978
Mola mola	ensembl	ensembl V95	319
Monodelphis domestica	ensembl	ensembl V95	2870
Monopterus albus	ensembl	ensembl V95	84
Mus musculus	ensembl	ensembl V95	1109
Neolamprologus brichardi	ensembl	ensembl V95	157

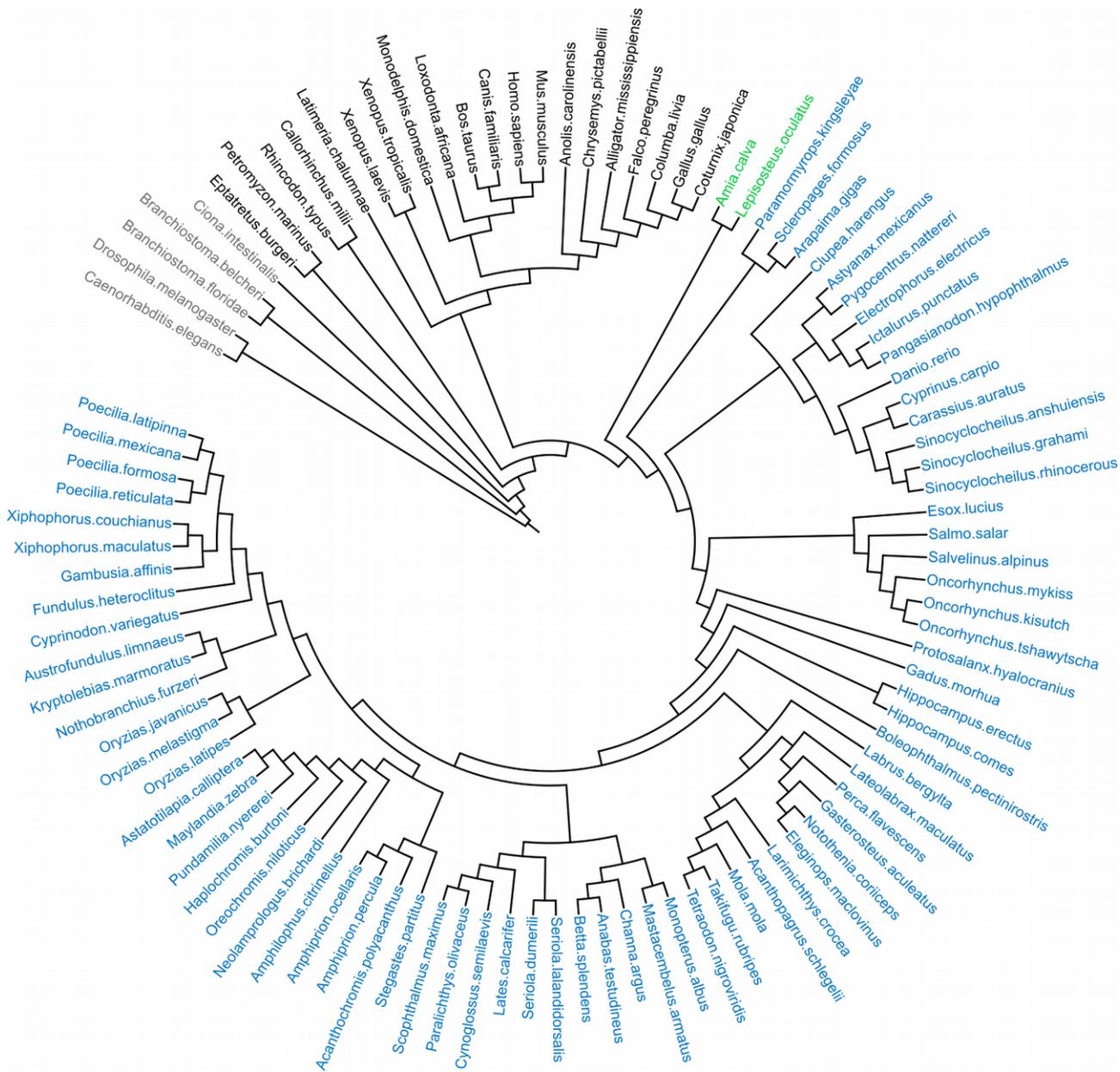
Species	Source	Data	N50 size (nb genes)
Oreochromis niloticus	ensembl	ensembl V95	88
Oryzias latipes	ensembl	ensembl V95	993
Oryzias melastigma	ensembl	ensembl V95	759
Paramormyrops kingsleyae	ensembl	ensembl V95	64
Petromyzon marinus	ensembl	ensembl V95	4
Poecilia formosa	ensembl	ensembl V95	52
Poecilia latipinna	ensembl	ensembl V95	11
Poecilia mexicana	ensembl	ensembl V95	12
Poecilia reticulata	ensembl	ensembl V95	934
Pundamilia nyererei	ensembl	ensembl V95	92
Pygocentrus nattereri	ensembl	ensembl V95	41
Scleropages formosus	ensembl	ensembl V95	172
Scophthalmus maximus	ensembl	ensembl V95	944
Seriola dumerili	ensembl	ensembl V95	208
Seriola lalandi dorsalis	ensembl	ensembl V95	57
Stegastes partitus	ensembl	ensembl V95	15
Takifugu rubripes	ensembl	ensembl V95	298
Tetraodon nigroviridis	ensembl	ensembl V95	767
Xenopus tropicalis	ensembl	ensembl V95	23
Xiphophorus couchianus	ensembl	ensembl V95	844
Xiphophorus maculatus	ensembl	ensembl V95	1020
Arapaima gigas	genofish		783
Oryzias javanicus	genofish		1009
Perca flavescens	genofish		920
Acanthopagrus schlegelii	gigaDB	http://gigadb.org/dataset/100409	238
Betta splendens	gigaDB	http://gigadb.org/dataset/100433	1093
Channa argus	gigaDB	http://gigadb.org/dataset/100279	44
Eleginops maclovinus	gigaDB	http://gigadb.org/dataset/102163	28
Hippocampus erectus	gigaDB	http://gigadb.org/dataset/100298	96
Lateolabrax maculatus	gigaDB	http://gigadb.org/dataset/100458	46
Protosalanx hyalocranius	gigaDB	http://gigadb.org/dataset/100262	49
Alligator mississippiensis	NCBI	GCF_000281125.3	97
Austrofundulus limnaeus	NCBI	GCF_001266775.1	33
Boleophthalmus pectinirostris	NCBI	GCF_000788275.1	111
Branchiostoma belcheri	NCBI	GCF_001625305.1	166
Branchiostoma floridae	NCBI	GCF_000003815.1	138
Carassius auratus	NCBI	GCF_003368295.1	743
Clupea harengus	NCBI	GCF_000966335.1	63
Columba livia	NCBI	GCF_000337935.1	50
Coturnix japonica	NCBI	GCF_001577835.1	523
Cyprinus carpio	NCBI	GCF_000951615.1	249

Species	Source	Data	N50 size (nb genes)
<i>Electrophorus electricus</i>	NCBI	GCF_003665695.1	29
<i>Falco peregrinus</i>	NCBI	GCF_000337955.1	51
<i>Larimichthys crocea</i>	NCBI	GCF_000972845.1	980
<i>Lates calcarifer</i>	NCBI	GCF_001640805.1	53
<i>Nothobranchius furzeri</i>	NCBI	GCF_001465895.1	1039
<i>Notothenia coriiceps</i>	NCBI	GCF_000735185.1	16
<i>Oncorhynchus kisutch</i>	NCBI	GCF_002021735.1	906
<i>Oncorhynchus mykiss</i>	NCBI	GCF_002163495.1	1505
<i>Oncorhynchus tshawytscha</i>	NCBI	GCF_002872995.1	861
<i>Pangasianodon hypophthalmus</i>	NCBI	GCF_003671635.1	435
<i>Paralichthys olivaceus</i>	NCBI	GCF_001970005.1	307
<i>Rhincodon typus</i>	NCBI	GCF_001642345.1	3
<i>Salmo salar</i>	NCBI	GCF_000233375.1	1814
<i>Salvelinus alpinus</i>	NCBI	GCF_002910315.1	692
<i>Sinocyclocheilus anshuiensis</i>	NCBI	GCF_001515605.1	42
<i>Sinocyclocheilus grahami</i>	NCBI	GCF_001515645.1	42
<i>Sinocyclocheilus rhinoceros</i>	NCBI	GCF_001515625.1	33
<i>Xenopus laevis</i>	NCBI	GCF_001663975.1	1603
<i>Callorhynchus milii</i>	NCBI	GCF_000165045.1	78
<i>Amia calva</i>	private	university oregon	953

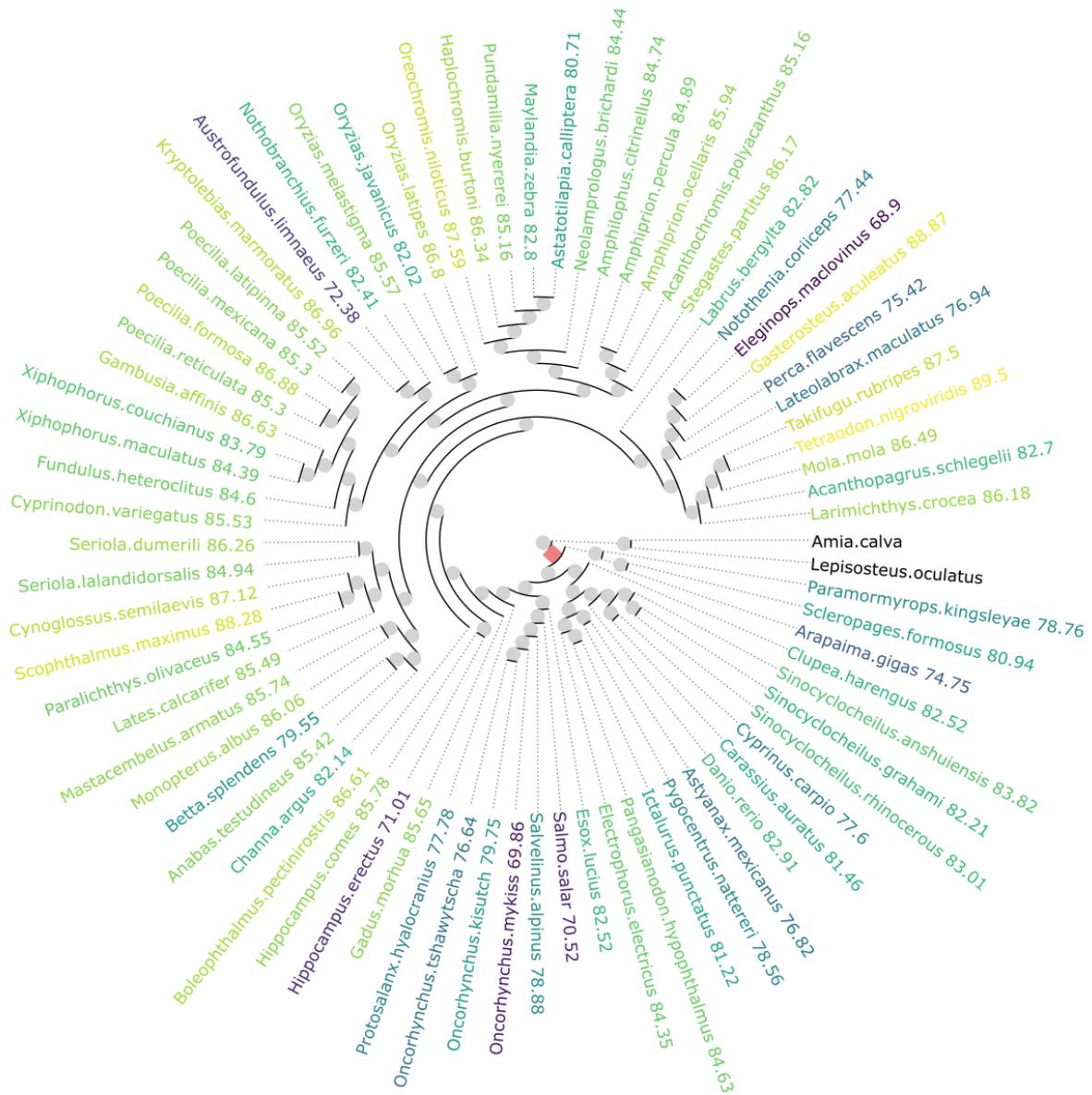
Supplementary Table S1 : List of the 101 genome assemblies used in this study. The dataset contains Ensembl, NCBI and GigaDB genomes as well as genomes from the Genofish consortium. An indication of genome assembly quality is given by the N50, a contiguity measure, which is the size (in genes) of the smallest contig amongst all contigs accounting for half the size of the genome.

Event type	Ancestral chromosomes involved	Descending medaka chromosome	Observed in arapaima
Fusion + mixing	4b-5b	chr19	chr24 (?)
Fission	4b_5b	chr19-chr1	no
Fusion	4b-5b-6a	chr1	chr10 (?)
Fusion + mixing	6b-7a	chr10	chr8 (?)
Fusion + mixing	7b-8a	chr14	chr1 (?)
Fusion + mixing	10b-11b	chr6	no
Fission	13a	chr4 & chr20	no

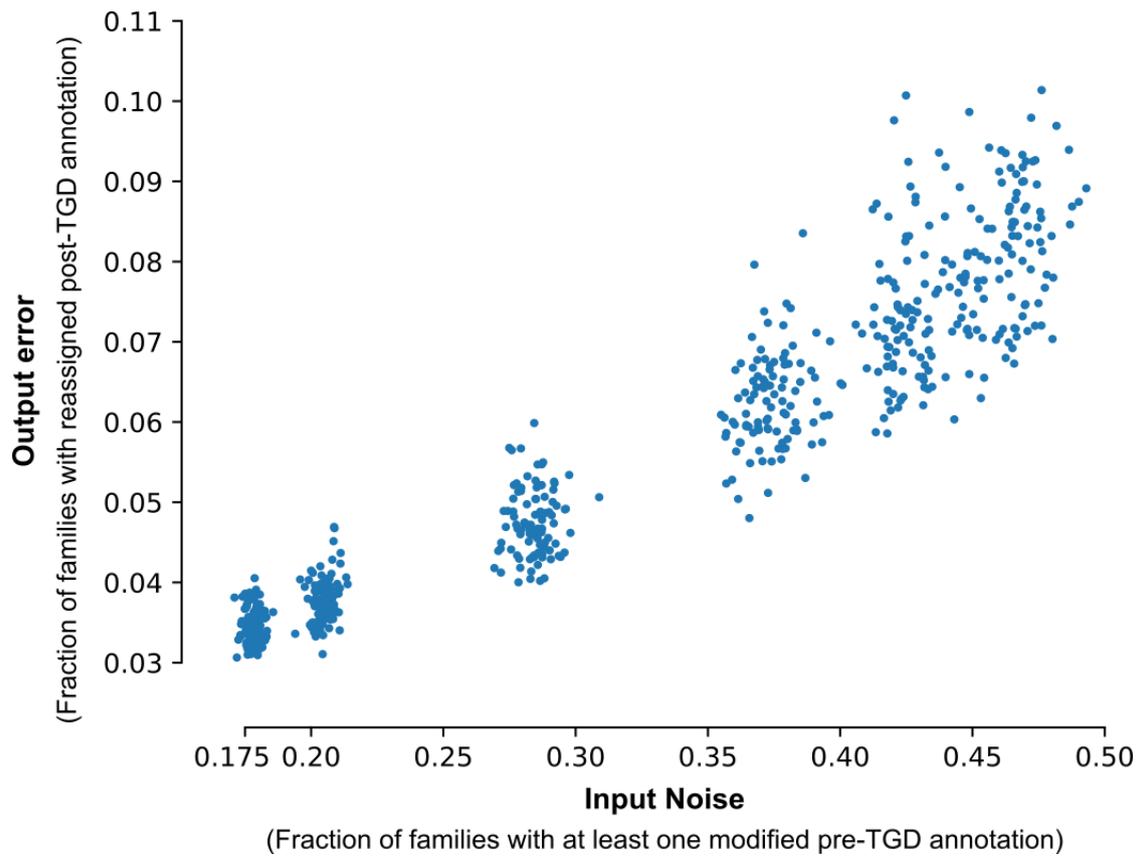
Supplementary Table S2 : Description of 7 major pre-clupeocephalan rearrangements described in (Kasahara et al. 2007) and recovered in the paralogy map. For each rearrangement, arapaima chromosomes showing the same mixing of ancestral chromosomes are indicated. Because of elevated rearrangements between arapaima and medaka genomes, and in the absence of a reconstruction of the ancestral Osteoglossiform genome, the shared or independent nature of these events can't be established with confidence, as indicated by a question mark (Supplementary Figures S5-S9).



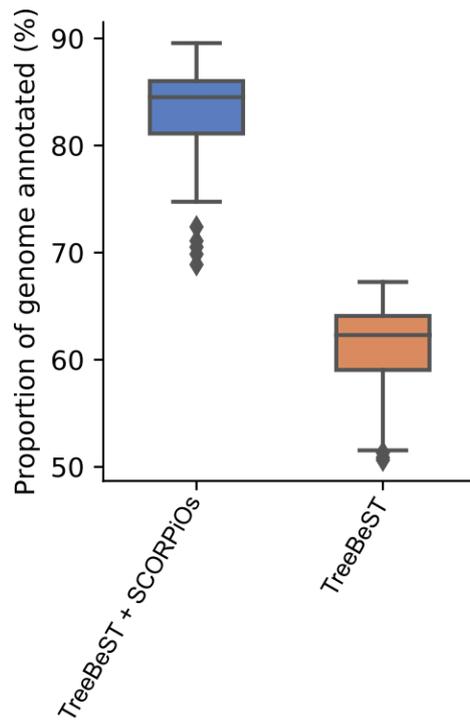
Supplementary Figure S1 : Species tree of the 101 genomes used in this study. The 74 teleost species are shown in blue, non-duplicated fish outgroups in green, other vertebrates in black and non-vertebrate outgroups in grey. Branch lengths are not to scale.



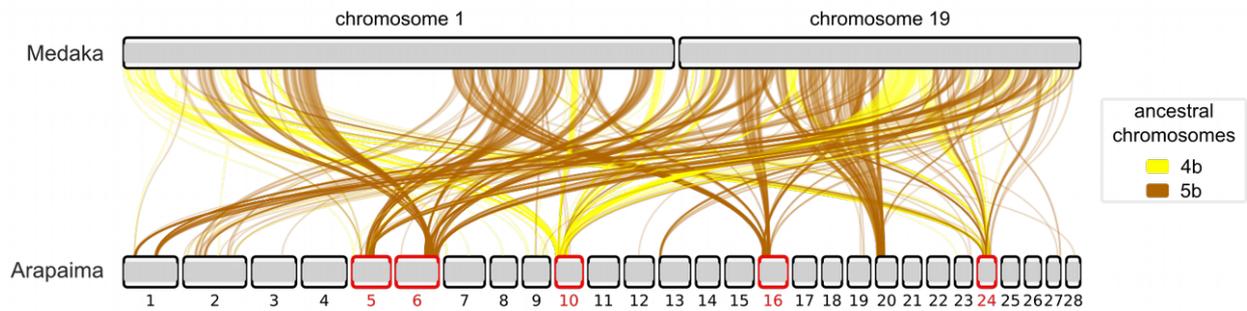
Supplementary Figure S2 : Species tree of the 74 teleost species and their percentage of genes annotated in the paralogy map. The tree is reduced to the teleost species and non-duplicated outgroups, with TGD duplication node shown as a red square. Annotated fractions of teleost genomes range from 68.9% (*Eleginops.maclovinus* or rock cod) to 89.5% (*Tetraodon.nigroviridis* or tetraodon) annotated genes.



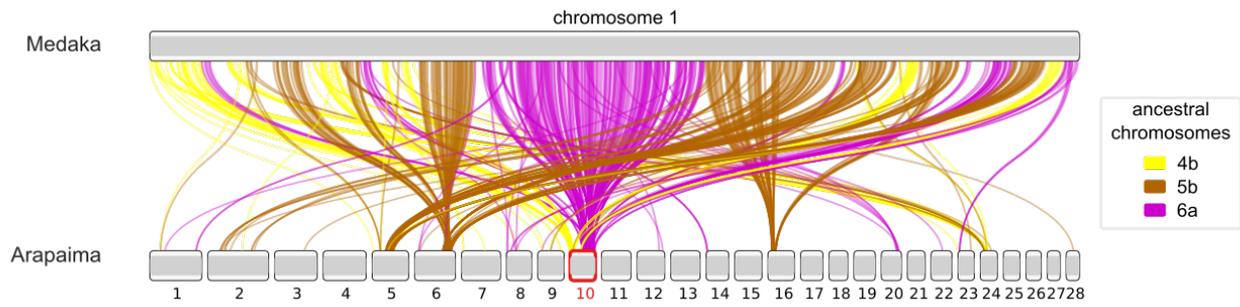
Supplementary Figure S3 : Noise robustness of the paralogy map. Proportion of gene families with post-TGD ancestral chromosome reassignments for different input noise settings (Methods). The input noise represents the proportion of families with at least one reference gene with a modified pre-TGD chromosome annotation due to an ancestral chromosome boundary shift in the simulation. The output error represents the fraction of gene families consequently assigned to a different post-TGD ancestral chromosome in the simulation. In the paralogy map workflow, the majority vote procedure ensures that the majority of these individual errors in inputs are not propagated to the output.



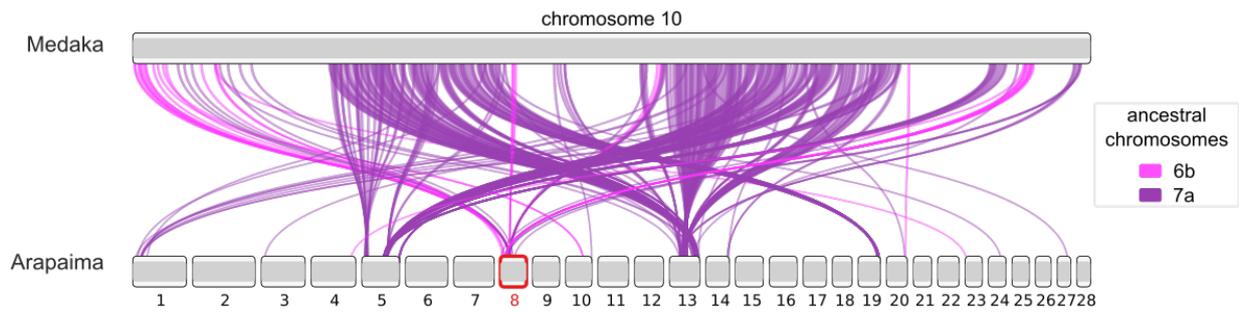
Supplementary Figure S4 : Impact of SCORPiOs on the establishment of the paralogy map. Distribution of the proportion of extant genes annotated in the paralogy before (orange) and after (blue) SCORPiOs correction across species. Each boxplot summarizes the distribution of 74 points (one per species).



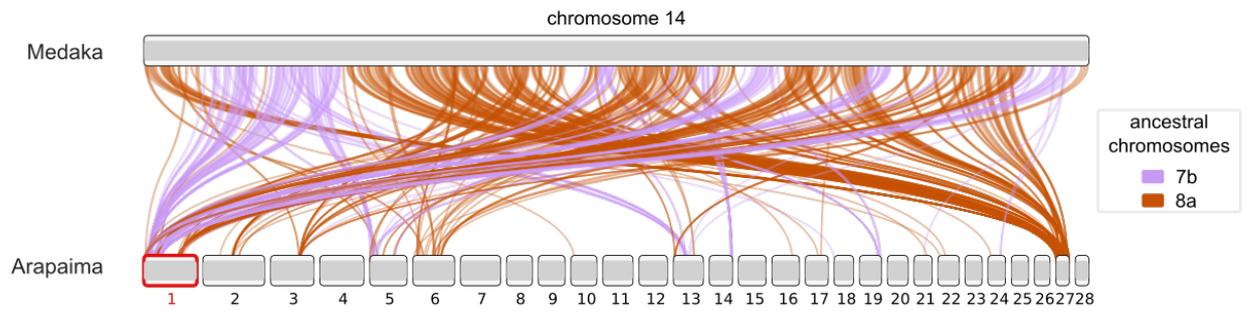
Supplementary Figure S5: Synteny comparisons of arapaima genome and medaka chromosomes 1 and 19. Orthologous genes are linked across the two genomes, with colors indicative of predicted ancestral chromosomes. Medaka chromosomes 1 and 19 descend from the fission of fused ancestral chromosomes 4b and 5b. Orthologous genes of chromosome 1 and 19 are linked together on 5 arapaima chromosome (shown in red), thus suggesting the fission to be clupeocephala-specific. Regarding the fusion event, only chromosome 24 of arapaima shows a clear mixing of 4b and 5b ancestral chromosomes, but synteny with medaka chromosomes is degraded.



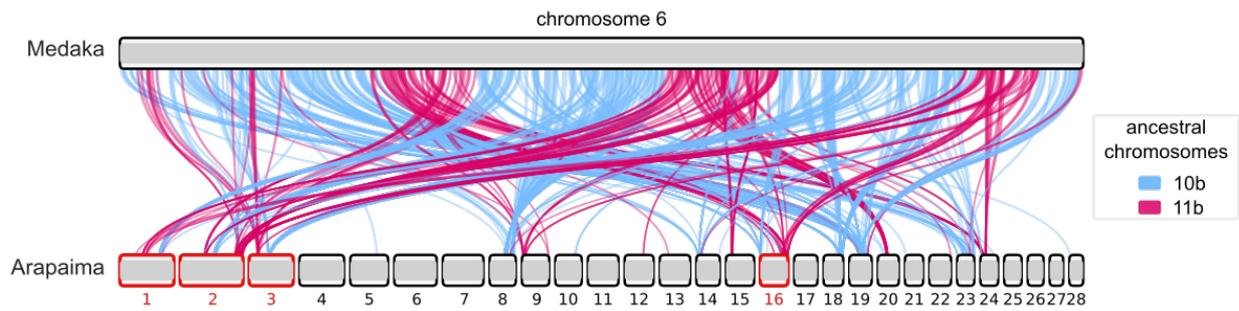
Supplementary Figures S6: Synteny comparisons of arapaima genome and medaka chromosome 1. An additional fusion event has shaped medaka chromosome 1, with the fusion of ancestral chromosome 6a with the 4b-5b fused chromosome. Ancestral chromosomes 6a and 4b are fused on arapaima chromosome 10 but 5b is absent.



Supplementary Figures S7: Synteny comparisons of arapaima genome and medaka chromosome 10. Medaka chromosome 10 was formed through the fusion and mixing of ancestral chromosomes 6b and 7a. A few orthologous genes of chromosome 10 descending from 6b and 7a are linked together on arapaima chromosome 8, but most are on distinct chromosomes.



Supplementary Figures S8: Synteny comparisons of arapaima genome and medaka chromosome 14. Medaka chromosomes 14 descend from the fusion and mixing of ancestral chromosomes 7b and 8a. A few orthologous genes of chromosome 14 descending from 7b and 8a are mixed together on arapaima chromosome 1.



Supplementary Figures S9: Synteny comparisons of arapaima genome and medaka chromosome 6. Medaka chromosome 6 descends from the fusion of ancestral chromosomes 10b and 11b. In arapaima orthologous genes of 10b and 11b are never directly fused, although they are linked together on a few chromosomes (highlighted in red). This suggest that rearrangements of ancestral chromosomes 10b and 11b have been independent in Clupeocephala and Osteoglossiforms.

Chapitre 4

Étude de la rediploïdisation suite à la duplication complète des poissons téléostéens

Je présente, dans ce chapitre, le travail ayant trait à mon dernier objectif de thèse. Cette étude concerne la caractérisation des génomes de poisson à court-terme après la duplication 3R et s'intéresse en particulier au processus de retour à l'état diploïde (ou rediploïdisation). Les résultats de ce travail sont à un stade plus préliminaire que ceux des deux chapitres précédents et ne sont pas encore mis en forme d'article, aussi, je les présente ici sous un format plus classique.

4.1 Introduction

Dans cette partie, je commence par introduire l'impact des duplications complètes sur le fonctionnement des méioses, ainsi que les mécanismes moléculaires pouvant être à l'origine d'échanges génétiques entre chromosomes dupliqués. Je présente ensuite les modèles classiques d'évolution de gènes ohnologues dans le contexte de la rediploïdisation. Enfin, je fais un point sur l'état des connaissances concernant les patrons spatio-temporels de rediploïdisation chez les Salmonidés et poissons téléostéens, et présente l'idée générale derrière la stratégie que nous avons mise en place afin de les étudier.

4.1.1 Recombinaison méiotique et échanges entre homéologues

Les duplications complètes de génome ont des conséquences dramatiques sur les processus cellulaires. La méiose est particulièrement impactée : l'appariement de chromosomes homéologues (Figure 4.1), en plus de l'appariement des chromosomes homologues, peut mener à des défauts de ségrégation des chromosomes. Sur le long terme, ces méioses instables sont généralement résolues. Dans les faits, cette résolution correspond au passage

d'un locus à 4 allèles qui interagissent génétiquement, à deux loci indépendants constitués chacun d'une paire d'allèles (WOLFE 2001).

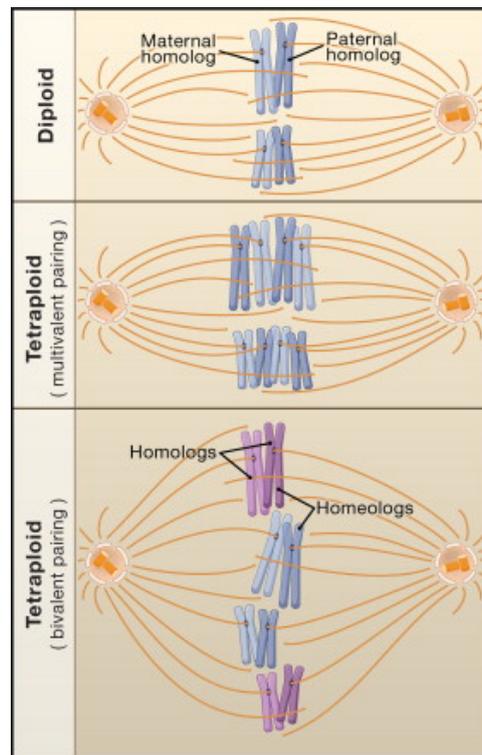


FIGURE 4.1 – Comportement méiotique de diploïdes et tétraploïdes. De haut en bas : méiose diploïde, avec appariement et séparation des chromosomes homologues ; méiose tétraploïde avec formation de tétravalents constitués de chromosomes homologues et homéologues et dont la séparation peut être inégale ; méiose tétraploïde résolue, avec appariement et séparation des chromosomes homologues. Figure tirée de (OTTO 2007).

Mécanismes de la recombinaison homologue en méiose

La méiose repose sur l'appariement correct des chromosomes homologues au stade zygotène de la prophase I, de façon à distribuer équitablement le matériel génétique dans les cellules filles. L'appariement des chromosomes homologues est stabilisé par des contacts physiques entre eux, pouvant donner lieu à des échanges génétiques. Ces contacts physiques sont initiés par un mécanisme programmé de cassures double brin (DSBs), réparées par la suite par recombinaison homologue suivant deux voies principales (voir la Figure 4.2 pour les mécanismes détaillés). Brièvement, les voies de réparation des DSBs en méiose passent par l'utilisation de la séquence homologue comme matrice à la réparation, à travers soit (i) la formation d'une double jonction de Holliday (DHJ), menant généralement à une résolution par crossing-over (CO), soit (ii) le synthesis-dependant strand annealing (SDSA), engendrant une résolution sans crossing-over (NCO). Les crossing-over sont à l'origine d'échanges de grands segments chromosomiques entre homologues, de l'ordre de 500 kb, via une translocation entre le brin matrice et le brin réparé. En comparaison, les réso-

lutions par NCOs donnent lieu à des échanges plus localisés (~ 2 kb). Dans les deux cas, ces mécanismes peuvent entraîner, localement, le remplacement d'une séquence allélique (receveuse, affectée par la cassure double-brin) par sa séquence homologue (donneuse, servant de matrice à la réparation). Puisqu'en fin de méiose, un seul des 4 produits est transmis à la descendance et peut, sur le long-terme, finir par être fixé, les COs peuvent mener à la fixation, dans toute une population, d'un chromosome où une grande région chromosomique a été remplacée par sa région homologue.

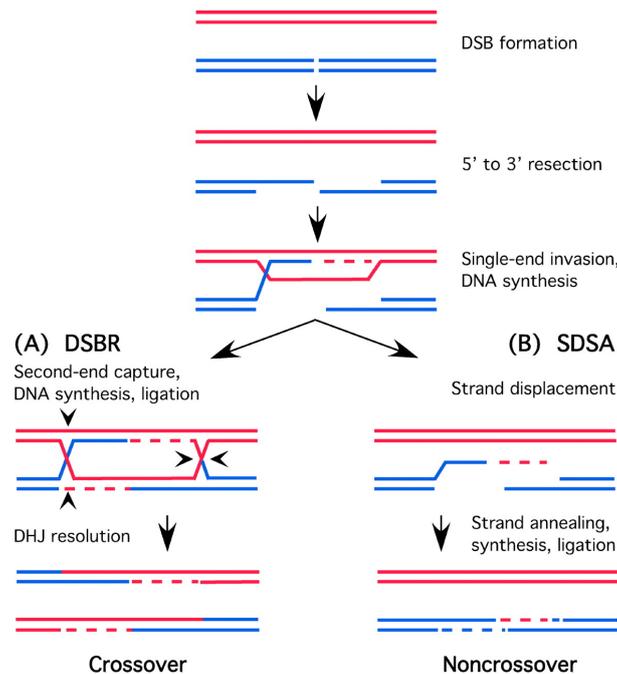


FIGURE 4.2 – Mécanismes principaux de recombinaison méiotique. La réparation de la cassure double brin affectant le chromosome bleu commence par la « résection » : la dégradation dans le sens 5' vers 3' des extrémités franches. Une des extrémités 3' simple brin exposées envahit ensuite le double brin homologue, formant un hétéroduplex. Le brin homologue sert de matrice à la synthèse d'ADN sur le brin cassé (les synthèses sont indiquées en pointillés). La résolution complète peut alors suivre deux voies principales. (A) Résolution DHJ : la seconde extrémité 3' se lie au complexe et le brin est également réparé grâce à la matrice formée par la séquence homologue. Cela aboutit à la formation d'une double jonction (DHJ) entre les 4 brins d'ADN. La DHJ est ensuite résolue par la coupure des brins internes ou externes de chaque jonction. (B) Résolution SDSA : l'hétéroduplex est désassemblé, le brin réparé se relie à sa chromatide soeur et la cassure restante est réparée. A noter que les produits après réparation, en A et en B, présentent des hétéroduplex. Ces hétéroduplex seront résolus par le système de réparation des mésappariements, pouvant mener soit à la restauration de la séquence receveuse sur les deux brins soit à sa conversion par la séquence donneuse. Figure tirée de DOONER 2002.

Mise en évidence et impact de la recombinaison homéologue

Dans le cas de méioses tétraploïdes, l'appariement des chromosomes homéologues peut amener, selon les mêmes mécanismes que ceux de la recombinaison homologue, à des

échanges de matériel génétique entre chromosomes dupliqués. Comme introduit précédemment (voir chapitre 1, paragraphe 1.3.3), les autopolyploïdes, mais également les allopolyploïdes (XIONG, GAETA et PIRES 2011 ; CHESTER et al. 2012 ; MASON et WENDEL 2020), sont confrontés au défi de la résolution de leur méiose. Chez les plantes, la formation de tétravalents en méiose était déjà proposée par Stebbins en 1947 (STEBBINS 1947) et depuis confirmée dans de nombreuses lignées naturelles et hybrides synthétiques (BOMBLIES et al. 2016). Chez les Vertébrés, l'observation de tétravalents a également été reportée chez la grenouille autotétraploïde *Odontophrynus americanus*, la loche allopolyploïde *Misgurnus anguillicaudatus* et différentes espèces de Salmonidés comme la truite mouchetée autotétraploïde *Salvelinus fontinalis* (BEÇAK, BEÇAK et RABELLO 1966 ; LEE et WRIGHT 1981 ; LI et al. 2011). La résolution des méioses peut être immédiate, comme chez les Xénopes (SESSION et al. 2016), ou s'étendre sur des millions d'années, comme chez les Salmonidés où certains homéologues continuent de recombinaison 95 millions d'années après l'événement de duplication Ss4R (GHARBI et al. 2006 ; LIEN et al. 2011 ; ALLENDORF et al. 2015 ; WAPLES, SEEB et SEEB 2016). De plus, la résolution ne se fait pas nécessairement sur la totalité du génome de manière simultanée, puisque les homéologues peuvent suivre des dynamiques de rediploïdisation différentes. Les mécanismes de résolution des méioses ne sont pas très bien caractérisés : en plus de la divergence des segments génomiques homéologues, des études chez les plantes suggèrent l'implication d'une réduction de la fréquence des contacts et crossing-over formés, de façon à les limiter à un par chromosome et ainsi favoriser la formation de bivalents (BOMBLIES et al. 2016).

Indéniablement, les échanges génétiques entre homéologues (HE) façonnent les génomes des polyploïdes et anciens polyploïdes de manière significative. Ces phénomènes, de plus en plus étudiés chez les plantes, sont encore mal caractérisés chez les Vertébrés. Chez les plantes, des exemples de recombinaison biaisée entre différents sous-génomes d'allopolyploïdes ont été mis en évidence. Cette recombinaison biaisée entraîne la conversion d'un sous-génome par le génome dominant, c'est à dire le plus exprimé (voir BIRD et al. 2018 pour une revue détaillée). Dans les cas les plus extrêmes, l'homogénéisation des sous-génomes pourrait ainsi masquer l'origine allopolyploïde des duplications sur le long-terme. De plus, les conséquences fonctionnelles des HE sont encore mal connues. Notamment, les événements de conversion génique, c'est à dire le remplacement d'un gène par sa copie paralogue, ont un effet difficile à prédire sur l'expression du gène remplacé. En effet, elle ne correspond ni à celle pré-HE ni à un doublement de l'expression de la séquence copiée (LLOYD et al. 2018). Le modèle le mieux caractérisé d'échanges entre homéologues chez les Vertébrés vient des saumons, où ils pourraient avoir joué un rôle dans leur radiation évolutive (voir le paragraphe 4.1.3).

4.1.2 Les modèles de résolution des ohnologues : LORe et AORe

Les échanges entre chromosomes homéologues peuvent mener à l'homogénéisation des génomes dupliqués et donc, en particulier, à l'homogénéisation des séquences de gènes dupliqués. Notamment, les événements de conversion génique entre gènes paralogues ont pour résultat d'homogénéiser leur séquence et ainsi retarder leur divergence. Dans un effort pour unifier les idées formulées dans la littérature depuis plusieurs années (FURLONG et HOLLAND 2002 ; MACQUEEN et JOHNSTON 2014 ; MARTIN et HOLLAND 2014) au sujet de la recombinaison homéologue, ROBERTSON et al. 2017 ont établi deux modèles d'évolution de gènes ohnologues dans le contexte de la rediploïdisation (Figure 4.3).

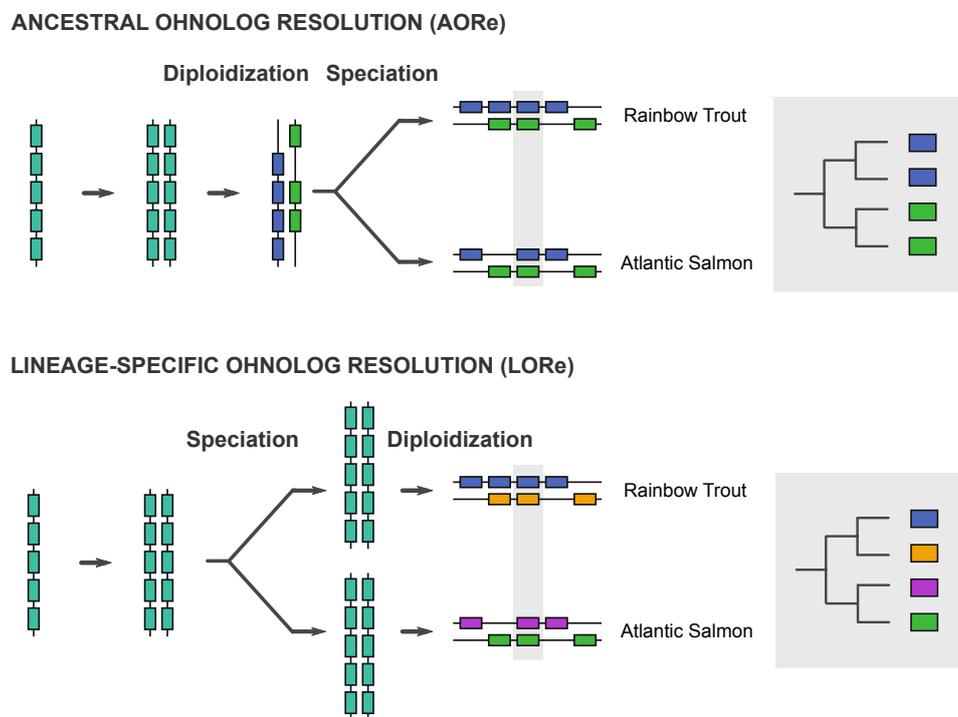


FIGURE 4.3 – Modèles de résolution des ohnologues. Exemple schématique de régions génomiques du saumon atlantique et de la truite arc-en-ciel suivant les modèles de rediploïdisation AORe et LORe, avec les topologies d'arbres correspondantes. Dans le modèle AORe (haut), la région considérée a été diploïdisée avant la spéciation des deux espèces, initiant ainsi la divergence des séquences paralogues avant la spéciation. Dans le modèle LORe (bas), la diploïdisation est retardée à après la spéciation, retardant la divergence des séquences.

Les deux modèles définissent les topologies d'arbres de gènes attendues suivant l'ordre d'occurrence des événements de rediploïdisation et de spéciation. Lorsque les gènes ohnologues sont rediploïdisés avant la spéciation des espèces considérées, la divergence des séquences s'effectue avant la spéciation et les gènes sont groupés dans les arbres par orthologues (gènes bleus et gènes verts). Ce modèle est nommé modèle AORe (« Ancestral Ohnolog Resolution ») par opposition au second modèle, LORe (« Lineage-Specific Ohnolog

Resolution ») dans lequel la rediploïdisation s’effectue après un événement de spéciation. Dans les topologies LORe, les deux ohnologues d’une même espèce (ou clade) sont groupés ensemble car les séquences n’avaient pas encore divergé au moment de la spéciation. Lorsqu’une topologie LORe est réconciliée à l’arbre des espèces, selon un modèle duplication-perte, les ohnologues apparaissent comme descendant de duplications lignées-spécifiques.

4.1.3 Mise en évidence de rediploïdisation lignée-spécifique chez les saumons

Le comportement méiotique des génomes de Salmonidés est encore caractérisé par la formation de tétravalents chez le mâle, un phénomène observé depuis les années 80 (LEE et WRIGHT 1981). L’obtention de cartes génétiques, puis le séquençage du génome de différentes espèces de Salmonidés ont depuis confirmé la présence de régions dupliquées pour lesquelles la rediploïdisation est toujours en cours, 95 millions d’années après la duplication (GHARBI et al. 2006 ; LIEN et al. 2011 ; BERTHELOT et al. 2014 ; ALLENDORF et al. 2015).

Par la suite, ROBERTSON et al. 2017 ont mis en évidence les patrons saptio-temporels de la rediploïdisation chez les Salmonidés, à travers la classification de 383 arbres de gènes en topologies AORE et LORe. La reconstruction de la chronologie des événements de rediploïdisation dans les 3 grandes familles de Salmonidés (Thymallinés, Coregoninés et Salmoninés, Figure 4.4) a permis de préciser la dynamique temporelle de la diploïdisation. Il est estimé que 25% du génome était encore non-rediploïdisé chez l’ancêtre des Salmonidés (daté à 50 millions d’années, soit 45 millions d’années après la Ss4R) et à 15% non-rediploïdisé à l’ancêtre Salmoninae (daté à 30 millions d’années, soit 65 millions d’années après la Ss4R). La projection des arbres AORE et LORe sur le génome du saumon atlantique a permis d’établir une cartographie des régions à rediploïdisation tardive dans cette espèce (Figure 4.5).

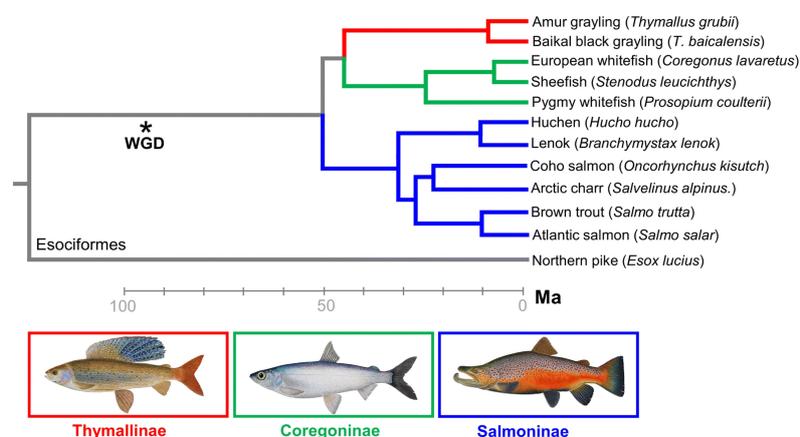


FIGURE 4.4 – Phylogénie des espèces de Salmonidés considérées dans l’étude de ROBERTSON et al. 2017. Les trois grands clades, les Thymallinés, les Coregoninés et les Salmoninés, sont représentés respectivement en rouge, vert et bleu. Figure tirée de ROBERTSON et al. 2017.

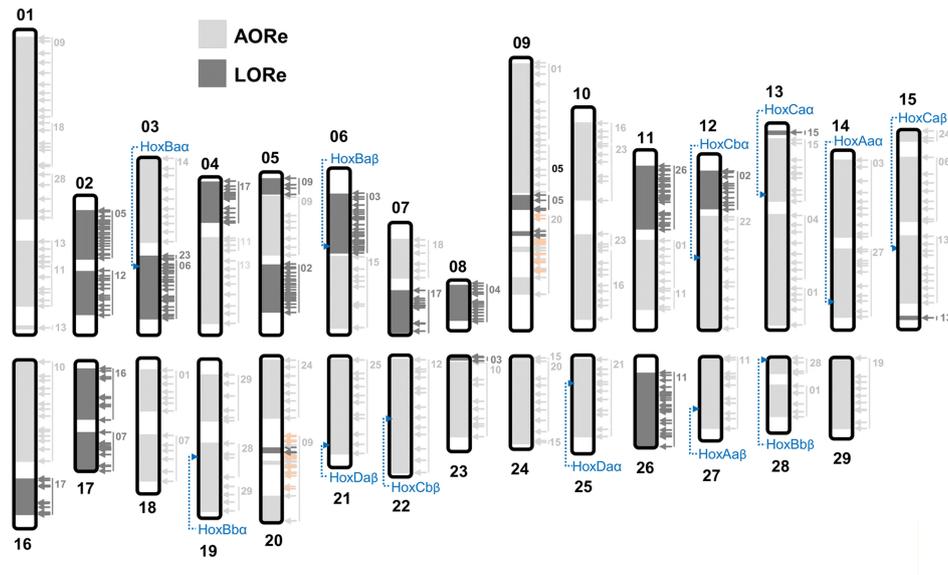


FIGURE 4.5 – Régions chromosomiques à rediploïdisation tardive chez le saumon atlantique. L'analyse de 383 arbres de gènes a permis de définir des régions à rediploïdisation ancestrale (AORE) et retardées (LORE) chez les Salmonidés. Les nombres à droite des régions chromosomiques identifient l'homologue correspondant (N.B. : la région LORE sur le chromosome 4 est homologue à celle du chromosome 8 et non le 17 comme indiqué). Les flèches oranges indiquent des topologies ambiguës, ni strictement LORE, ni strictement AORE. Figure adaptée de ROBERTSON et al. 2017.

Cette étude a également mis en lumière l'impact fonctionnel de la rediploïdisation lignée-spécifique. Au sein d'une espèce, la faible divergence de séquences des ohnologues LORE corrèle avec une faible différence d'expression. En revanche, entre lignées, elles permettent l'accumulation de substitutions spécifiques. Les auteurs rapportent une vitesse élevée de rediploïdisation avant l'épisode de diversification majeur des Salmoninés par rapport aux autres lignées : ~47 paires d'ohnologues rediploïdisées à la base des Salmoninae contre seulement 12 par millions d'années à la base des Coregoninés. Cette diversification coïncide avec une période de refroidissement climatique et l'évolution de l'anadromie (la capacité de migrer entre eau salée et eau douce) à la base des clades diversifiés (MACQUEEN et JOHNSTON 2014). L'évolution de l'anadromie est associée à l'acquisition de différentes adaptations physiologiques à l'eau salée, comme l'osmorégulation. ROBERTSON et al. 2017 ont montré un enrichissement en fonctions biologiques en lien avec l'acquisition de l'anadromie dans les ohnologues LORE, même si ces adaptations s'expriment à travers une contribution complexe de gènes à la fois AORE et LORE. De manière importante, la rediploïdisation lignée-spécifique représente un mécanisme pouvant expliquer le décalage observé entre les événements de duplications complètes et le succès évolutifs des clades descendants, exprimé en fonction de facteurs environnementaux (le « radiation time-lag model » introduit au chapitre 1, paragraphe 1.3.3).

4.1.4 Résolution des ohnologues chez les poissons téléostéens

Bien documenté chez les Salmonidés suite à la Ss4R, le patron spatio-temporel de la rediploïdisation suite à la 3R chez les poissons téléostéens est moins bien caractérisé. Contrairement aux espèces de saumons, la rediploïdisation post-3R semble désormais complète dans toutes les espèces de téléostéens actuels. Le séquençage récent de génomes du clade des Osteoglossiformes (BIAN et al. 2016 ; DU et al. 2019) représente une opportunité pour examiner le niveau de rediploïdisation de l'ancêtre Osteoglossocéphalai, qui précède la divergence des Osteoglossiformes et Clupeocephala (voir paragraphe 1.2.2). La divergence entre ces deux grands clades est estimée à ~ 60-80 millions d'années après la 3R (date de divergence estimée à 273.8 et 244.4 millions d'années par les deux études de référence actuelles BETANCUR-R et al. 2017 ; HUGHES et al. 2018) et il est suggéré que les génomes n'étaient alors pas encore complètement rediploïdisés.

En effet, MARTIN et HOLLAND 2014 ont séquencé les clusters de gènes *hox* du poisson-papillon (*Pantodon buchholzi*, Osteoglossiforme) et étudié leurs relations d'homologie avec les clusters *hox* des Clupeocephala. Les auteurs ont mis en évidence des relations d'orthologie ambiguës entre les gènes des clusters *hoxb*, *hoxc* et *hoxd* du poisson-papillon et ceux des Clupeocephala. En effet, dans les arbres de gènes, les gènes du poisson-papillon ne sont pas groupés de manière systématique avec la copie 'a' ou 'b' des gènes de Clupeocephala. Les topologies d'arbres inférées pouvaient s'expliquer de deux manières : soit par l'occurrence de duplications complètes indépendantes chez les Clupeocephala et les Osteoglossiformes, soit par une résolution lignée-spécifique des ohnologues (LORe). Deux arguments ont mené les auteurs à favoriser l'hypothèse « LORe » : d'une part elle est plus parcimonieuse, et d'autre part l'histoire évolutive des clusters *hoxa* est cohérente avec une duplication partagée (Figure 4.6). A ce jour, aucune étude n'a tiré profit des ressources de génomes complets pour explorer la question de la rediploïdisation après la 3R.

4.1.5 Potentiel de SCORPiOs pour l'étude des patrons de rediploïdisation

Confronter les deux modèles de résolution d'ohnologues (AORé et LORe) nécessite d'être capable de reconstruire correctement l'histoire évolutive des gènes dupliqués et de pouvoir définir les topologies attendues sous chacun des deux modèles. Je présente ici comment les résultats de l'application de SCORPiOs pour corriger les arbres de gènes peut permettre, directement et à grande échelle, d'identifier des cas de résolution lignée-spécifique d'ohnologues (LORe).

Nous avons conçu SCORPiOs dans le but d'intégrer l'histoire évolutive d'un gène en tant que locus, telle qu'inférée à travers les patrons de synténie conservée, à un modèle

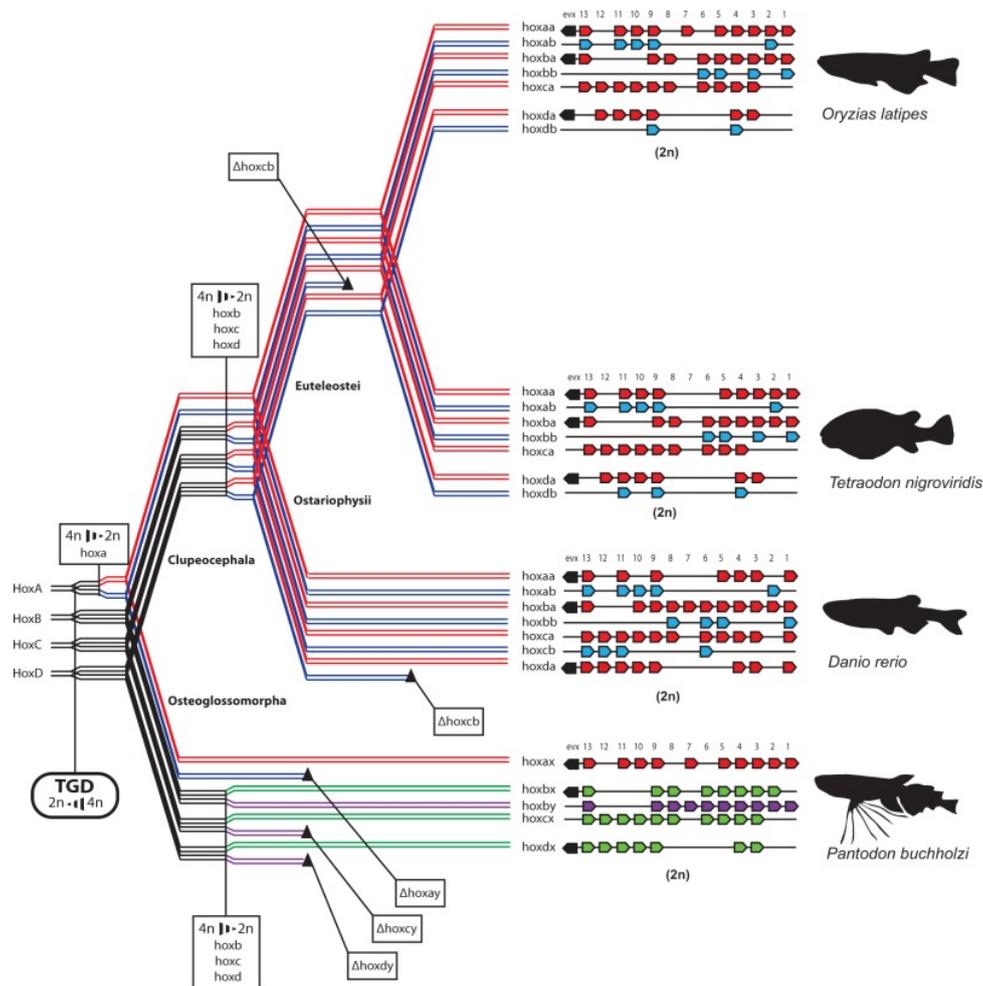


FIGURE 4.6 – Évolution des clusters *hox* chez les Osteoglossomorphes, selon MARTIN et HOLLAND 2014. Les clusters *hoxa* ont été rediploïdifiés avant la divergence des Osteoglossomorphes avec les Clupeocephala. En effet, dans les arbres de gènes, les gènes de l'unique cluster *hoxa* du poisson papillon, *hoxaa*, sont groupés de manière non ambiguë avec les gènes *hoxaa* des Clupeocephala. À l'inverse, les clusters *hoxb*, *hoxc*, et *hoxd* ont conclu leur rediploïdification indépendamment chez les Osteoglossomorphes et les Clupeocephala. Figure tirée de MARTIN et HOLLAND 2014.

d'évolution de sa séquence. Plusieurs causes peuvent donner lieu à un désaccord entre ces deux types de prédictions. D'une part, des raisons méthodologiques peuvent être à l'origine d'erreurs. Par exemple, dans le cas des méthodes de séquences, des erreurs peuvent être générées en lien avec l'utilisation de modèles d'évolution non adaptés ou à des alignements de séquences incorrects. Dans le cas de la synténie, des erreurs peuvent survenir si les hypothèses de SCORPiOs ne sont pas vérifiées. Notamment, SCORPiOs s'appuie sur l'hypothèse que les arbres de départ, sont, en majorité, corrects. De plus, des artefacts dans les données, que ce soit des erreurs dans les séquences des gènes ou dans l'assemblage des génomes peuvent constituer une deuxième source d'incongruence.

La résolution tardive des ohnologues à travers l'occurrence de conversion génique est également une source de conflit entre les arbres de gènes basés sur la séquence et les pré-

dictions de SCORPiOs basées sur la synténie. SCORPiOs décrit l'histoire du locus du gène en positionnant correctement la position de sa duplication. SCORPiOs fait l'hypothèse que la rediploïdisation était terminée dans l'ancêtre des espèces considérées et il se sert de cette histoire postulée commune de diploïdisation pour regrouper les segments les plus similaires entre eux. De fait, SCORPiOs infère une topologie d'arbre « AORe », cohérente avec l'histoire de duplication du locus lorsque l'arbre de gènes est réconcilié à l'arbre des espèces. En présence de conversion génique entre paralogues, les arbres basés sur la séquence groupent les gènes par espèces (ou clades) : la topologie reflète correctement l'histoire de sa séquence, mais sa réconciliation ne reflète pas correctement l'histoire de son locus. En conséquence, si les erreurs méthodologiques sont moins représentées que ces processus biologiques, il est possible d'analyser les arbres pour lesquels SCORPiOs infère un conflit séquence-synténie pour identifier les topologies LORe. Comme SCORPiOs ne force pas la topologie inférée à partir des patrons de synténie si celle-ci n'est pas cohérente avec l'histoire évolutive des séquences, identifier ces conflits revient à identifier les arbres pour lesquels SCORPiOs propose une correction mais qu'elle est rejetée par les tests de vraisemblance (AU-tests, voir chapitre 2, paragraphe 2.1.5). Dans la partie suivante, je présente les résultats de l'application de cette stratégie à différents jeux de données.

4.2 Résultats

Cette partie est structurée de la manière suivante : dans un premier temps, je présente la validation d'une nouvelle stratégie de caractérisation des modèles LORe et AORe basée sur les arbres prédits par SCORPiOs, chez les Salmonidés ; dans un second temps, je présente les résultats de l'application de cette méthodologie aux poissons téléostéens.

4.2.1 SCORPiOs permet de retrouver et de préciser les patrons de rediploïdisation chez les Salmoninés

Les régions génomiques à rediploïdisation tardive sont bien définies chez les Salmonidés. Précédemment, nous avons reconstruit, avec SCORPiOs, des arbres de gènes contenant 74 espèces de poissons téléostéens (Matériel et Méthodes), dont 5 espèces de Salmoninés (Figure 4.7). Avec les Coregoninés et les Thymallinés, les Salmoninés représentent l'un des trois grands clades de Salmonidés. Afin de valider la pertinence de SCORPiOs dans l'étude des modes de rediploïdisation, nous avons en premier lieu vérifié si, sur ce jeu de données, les conflits de prédiction séquence/synténie - que je désigne par la suite sous le terme d'arbres incongruents - s'expliquaient effectivement par la rediploïdisation tardive plus que pour les raisons méthodologiques évoquées plus haut.

Nous avons commencé par identifier les arbres incongruents et analysé leur distribution sur les différentes paires de chromosomes dupliqués (Matériel et Méthodes). Le raison-

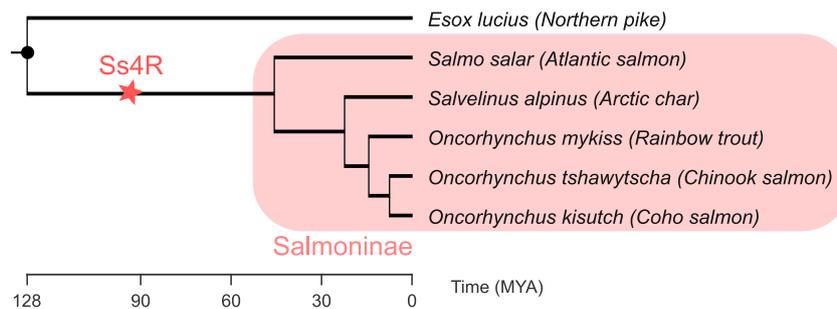


FIGURE 4.7 – Phylogénie des espèces Salmoninés utilisées. Les dates de divergences (en millions d’années) sont extraites de TimeTree (KUMAR et al. 2017). La duplication complète des Salmonidés (Ss4R) est indiquée par une étoile.

nement est le suivant : si les incongruences sont liées à des raisons méthodologiques, on s’attend à ce qu’elles soient distribuées aléatoirement sur les génomes. Or, nous observons ces arbres incongruents répartis préférentiellement sur 7 des 25 paires de chromosomes dupliqués (Figure 4.8). Ces paires de chromosomes correspondent à 7 des 8 paires identifiées dans la littérature comme présentant des traces de tétrasomie résiduelle en méiose mâle, au niveau des télomères (KODAMA et al. 2014). La huitième paire non-retrouvée ici correspond aux chromosomes dupliqués orthologues du chromosome LG06 du groupe externe *Esox lucius*. Cette paire a un comportement méiotique distinct dans différentes espèces de Salmoninés (KODAMA et al. 2014) et ne présente ni tétrasomie résiduelle ni rediploïdisation tardive chez le saumon atlantique (LIEN et al. 2011 ; ROBERTSON et al. 2017). Enfin, la projection des arbres incongruents sur le génome du saumon atlantique (*Salmo salar*) confirme qu’ils sont regroupés sur ces paires de chromosomes dupliqués, et correspondent effectivement aux régions définies par ROBERTSON et al. 2017 (Figure 4.5). La proportion d’arbres incongruents (2 279 incongruents sur 18 164 soit ~ 13%) est en accord avec celle inférée par Robertson et al. (15%) au même ancêtre Salmoninae. De fait, nous avons démontré que les résultats de la correction des arbres par SCORPiOs permettent de révéler des régions à rediploïdisation lignée-spécifique.

Dans un second temps, nous avons cherché à mettre en place une stratégie automatique et objective qui permettrait, à partir des résultats de SCORPiOs, de valider les régions à rediploïdisation tardive. En effet, dans le cas des Salmoninés, au vu de l’intersection de la position des arbres incongruents et des régions LORe connues, il apparaît évident que le modèle LORe est à l’origine des incongruences. Néanmoins, nous ne l’avons ici pas démontré explicitement. Dans cette optique, nous avons appliqué une approche de clustering d’arbres (« treeCl », GORI et al. 2016, voir Matériel et Méthodes en 4.3), afin de vérifier que les arbres incongruents étaient effectivement compatibles avec le modèle LORe et qu’ils pouvaient être groupés ensemble sur la base de leur similarité topologique. Le clustering s’appuie sur l’information de topologie et longueur de branche des arbres. Nous l’avons appliqué à un sous-échantillon de 4000 arbres (voir Matériel et Méthodes), constitués de 3285

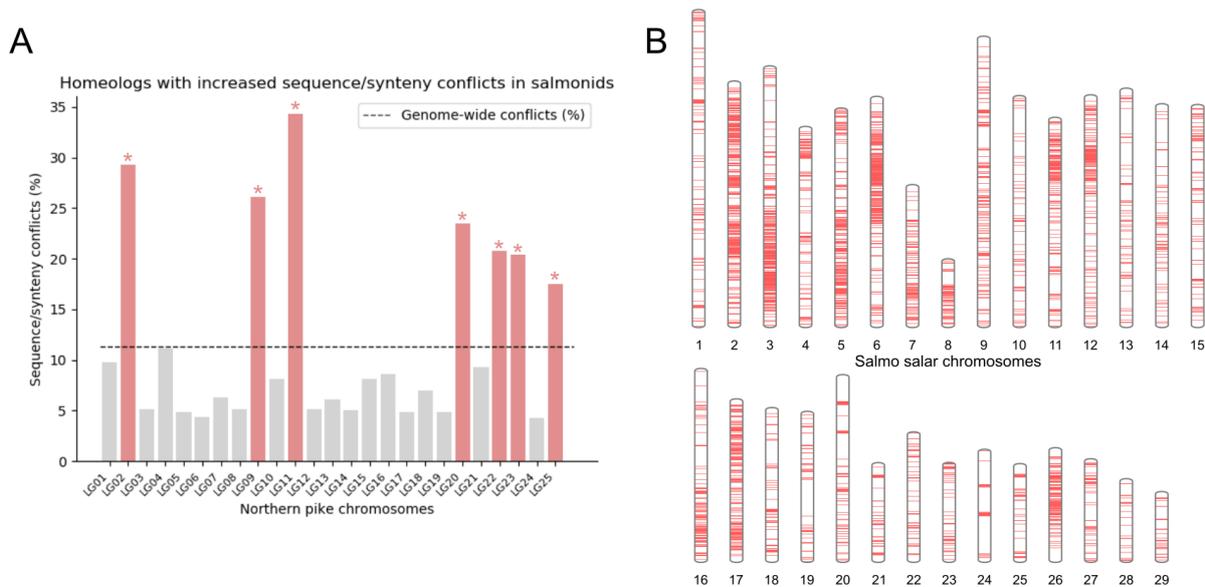


FIGURE 4.8 – Distribution des conflits séquences/synténie dans les génomes de Salmoninés. A. Proportion de conflits pour chaque paire de chromosomes dupliqués, représentée par les chromosomes du brochet *Esox lucius* (« Northern pike »), une espèce groupe externe non-dupliquée. L'utilisation du brochet facilite l'identification des paires de chromosomes homéologues, brouillées dans les espèces dupliquées par des événements de fusions chromosomiques. Le brochet est estimé proche du caryotype pré-duplication : un chromosome du brochet correspond à une paire de chromosomes dupliqués par la Ss4R. La ligne pointillée représente la proportion de conflits moyenne sur les chromosomes du brochet. Les chromosomes dupliqués avec un excès de conflits sont indiqués en rouge (* p-valeur < 0.05, tests hypergéométriques corrigés pour les tests multiples). B. Distribution des gènes appartenant aux arbres incongruents sur le génome du saumon atlantique.

suggérés corrects par SCORPiOs et 715 arbres incongruents. Le clustering a permis de grouper les arbres en trois clusters (1), (2), (3), contenant respectivement 1 615, 1 185 et 1 200 arbres (Matériel et Méthodes, Tableau 4.1). Les arbres synténie-cohérents sont principalement (47%) retrouvés dans le cluster 1, qui regroupe effectivement des topologies de type AORe. A l'inverse, les arbres incongruents sont retrouvés en majorité (60 %) dans le cluster 2, qui regroupe des topologies de type LORe. Enfin, 30% des arbres synténie-cohérents et 30% des arbres synténie incohérents sont retrouvés dans le cluster 3, qui contient un ensemble de topologies assez distantes de toutes les autres, dû à la présence d'une ou plusieurs longues branches.

	Total	Cluster 1 (AORe)	Cluster 2 (LORe)	Cluster 3
Congruent	3285	1544 (47%)	755 (23%)	986 (30%)
Incongruent	715	71 (10%)	429 (60%)	215 (30%)

TABLE 4.1 – Résultats du clustering des arbres de gènes Salmoninés. Le tableau indique la répartition des arbres définis par SCORPiOs comme synténie congruents et incongruents dans les différents clusters.

Les gènes des arbres du cluster 3 semblent répartis aléatoirement sur le génome du saumon atlantique, tandis que les arbres des clusters 1 et 2 sont, chacun, co-localisés. Nous avons ajusté un modèle de Markov caché afin de lisser le signal obtenu et ainsi délimiter, sur le génome du saumon atlantique, les régions correspondant à différents types de topologies d'arbres (Figure 4.9).

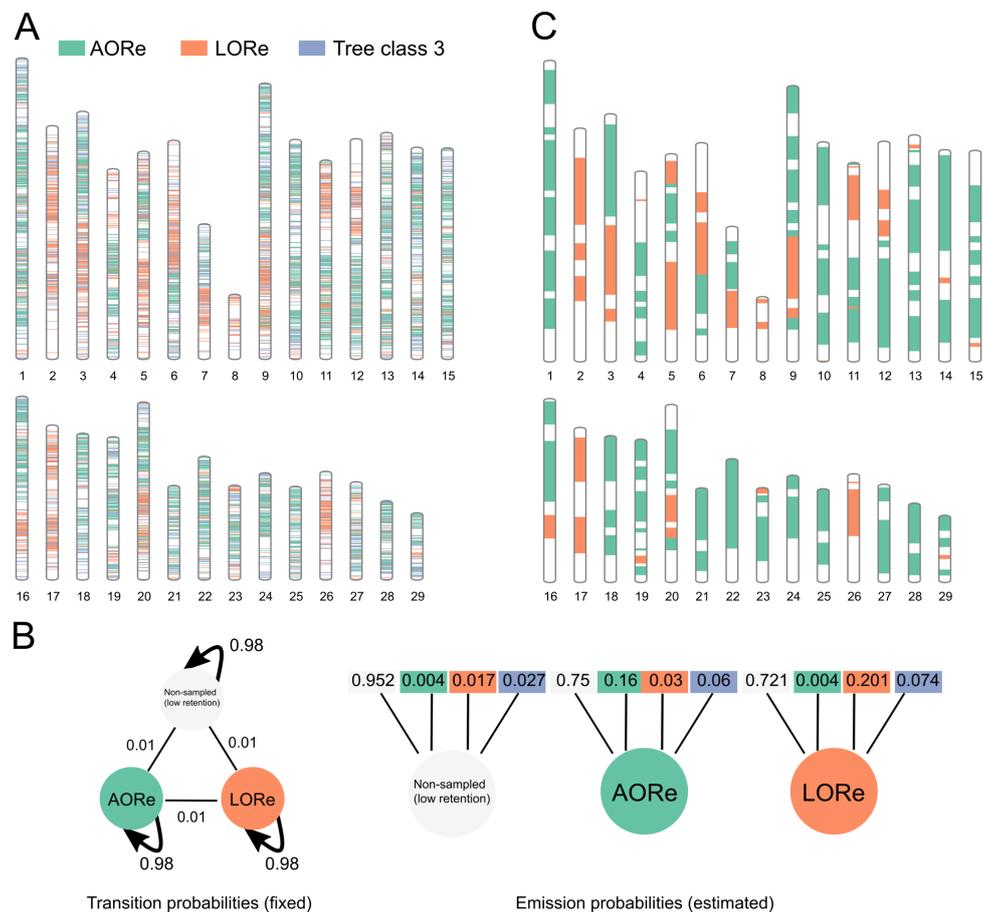


FIGURE 4.9 – Validation des régions AORE et LORe chez les Salmoninés. A. Distribution des 3 types de topologies d'arbres issues du clustering sur le génome du saumon atlantique. Les topologies AORE et LORe sont regroupées sur le génome. Les topologies du cluster 3 (indigo), réparties plus aléatoirement, correspondent à des topologies très distantes les unes des autres, généralement dû à une longue branche pour un ou plusieurs ohnologue(s). B. Paramètres du modèle de Markov caché ajusté pour inférer les limites des régions AORE et LORe. Un troisième état caché modélise les régions à faible rétention d'ohnologues, non-échantillonnées pour le clustering (Matériel et Méthodes). C. Régions AORE et LORe inférées sur le génome du saumon atlantique.

Ainsi, nous avons pu obtenir une cartographie des régions AORE et LORe, très similaire à l'état de l'art (ROBERTSON et al. 2017) et aux clusters d'arbres incongruents visualisés en Figure 4.8. La différence majeure avec ROBERTSON et al. 2017 vient de la paire de chromosomes dupliqués 9-20, pour laquelle les auteurs trouvent un certain nombre de topologies d'arbres « ambiguës » (en orange sur la Figure 4.5). Ici, le clustering groupe ces topologies d'arbres dans le groupe LORe malgré leur topologie AORE et l'absence d'incongruence rele-

vée par SCORPiOs (Figure 4.8). L'inspection de ces arbres révèle la présence d'une longue branche commune avant la divergence des ohnologues, indiquant un retard potentiel de la diploïdisation après la Ss4R, mais résolue avant la spéciation des Salmoninae. De plus, LIEN et al. 2016 avaient également relevé une similarité de séquence élevée pour cette paire d'homéologues. Enfin, la région LORe impliquant la paire de chromosomes 4-8 du saumon atlantique est beaucoup plus petite que dans ROBERTSON et al. 2017, en lien avec le fait que peu d'arbres de gènes ont été échantillonnés dans ces régions pour le clustering (Matériel et Méthodes).

Ces résultats précisent les connaissances précédentes, en permettant une analyse à plus grande résolution des patrons de rediploïdisation chez les Salmoninés (383 arbres inspectés par ROBERTSON et al. 2017, alors que SCORPiOs intègre tous les arbres de gènes Salmoninés n=18 164, Figure 4.8). En effet, nous avons validé les régions de rediploïdisation tardive suggérées par les résultats de SCORPiOs grâce à l'analyse plus approfondie d'un sous-jeu de 4 000 arbres de gènes. Cette application montre que SCORPiOs est suffisamment robuste pour être informatif à l'étude des patrons de rediploïdisation et permet d'envisager son utilisation pour les préciser, à travers le détournement de l'outil en un « chasseur de LORe ¹ ».

4.2.2 Caractérisation des patrons de rediploïdisation après la 3R

Échantillonnage des espèces

Pour étudier l'état de rediploïdisation de l'ancêtre Osteoglossocephalai, nous avons extrait un sous-jeu d'espèces du jeu de données de 74 poissons téléostéens afin de sélectionner un jeu de génomes de bonne qualité et représentant de manière équilibrée les espèces d'Osteoglossiformes et de Clupeocephala (Figure 4.10). Nous avons ensuite reconstruit un jeu d'arbres de gènes initiaux avec TreeBeST, corrigés par la suite avec SCORPiOs.

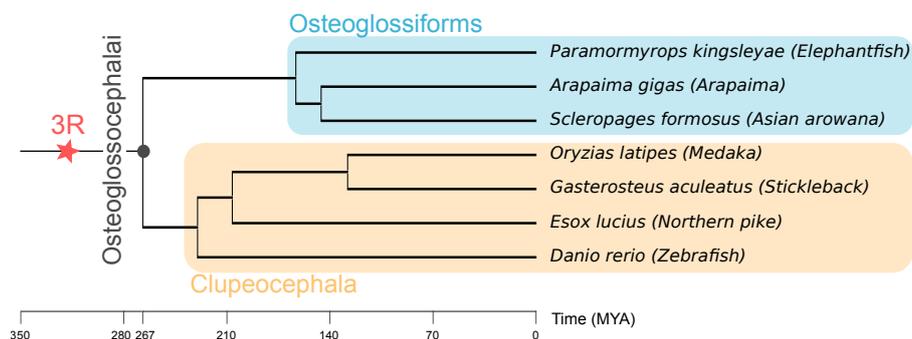


FIGURE 4.10 – Phylogénie des espèces Osteoglossocephalai utilisées. Les dates de divergences (en millions d'années) sont extraites de TimeTree (KUMAR et al. 2017). La position de la duplication 3R, datée à 320 millions d'années (VANDEPOELE et al. 2004), est indiquée par une étoile.

1. Merci à Daniel Macqueen pour l'expression ("SCORPiOs the LORe hunter").

Analyse des conflits séquence/synténie et chromosomes homéologues impliqués

De la même façon que sur le jeu de données d'arbres de gènes Salmonidés, nous avons analysé la distribution des conflits séquence/synténie sur les paires de chromosomes dupliqués. Ici, la répartition des conflits est représentée sur la base des chromosomes ancestraux pré-duplication (Figure 4.11), décrits dans le manuscrit présenté au chapitre précédent. En effet, dans le cas de la 3R, le génome du groupe externe non-dupliqué le plus proche, le lépisosté tacheté, possède un caryotype composé de micro-chromosomes. Ces micro-chromosomes ont été fusionnés dans la lignée des téléostes (BRAASCH et al. 2016). De ce fait, les caryotypes n'ont pas une correspondance 1 à 2 suite à la duplication, contrairement aux Salmonidés où l'espèce groupe externe, le brochet, a un caryotype proche du caryotype pré-duplication.

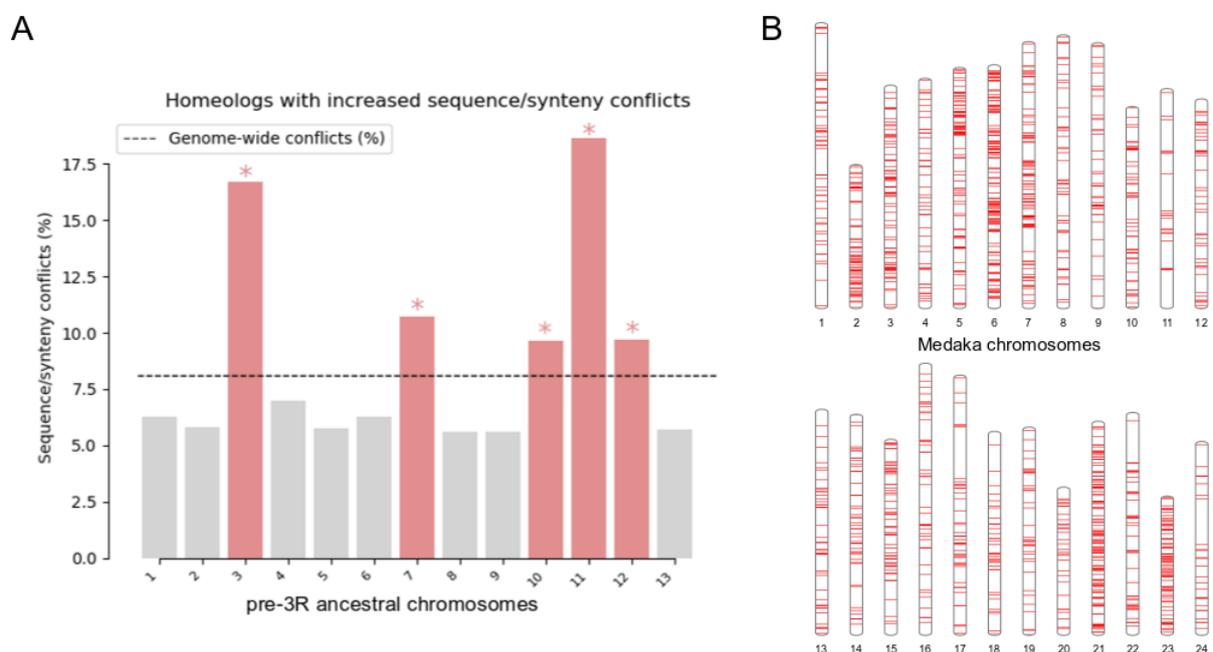


FIGURE 4.11 – Distribution des conflits séquences/synténie dans les génomes d’Osteoglossoccephalai. A. Chromosomes homéologues enrichis en conflits de prédictions entre séquence et synténie. Représentation comme en Figure 4.8, à l’exception de l’axe des abscisses qui correspond aux 13 chromosomes pré-3R (voir le chapitre précédent). Les paires d’homéologues qui descendent des chromosomes pré-duplication 3, 7, 10, 11 et 12 sont enrichis en conflits séquence/synténie (* p-valeur < 0.05, tests hypergéométriques corrigés pour les tests multiples). B. Distribution des gènes appartenant aux 1 121 arbres incongruents sur le génome du medaka, comme en Figure 4.8.

Cinq chromosomes pré-duplication, correspondant à des paires de chromosomes dupliqués dans les espèces modernes, sont sur-représentés dans les conflits de prédiction séquence/synténie : les chromosomes descendant des chromosomes pré-duplication 3, 7, 10, 11 et 12 (tests hypergéométriques corrigés pour les tests multiples avec la méthode de Benjamin-Hocheberg, p-valeur < 0.05, Figure 4.11). De plus, les gènes des 1 121 arbres incongruents forment à nouveau des clusters lorsque visualisés sur les génomes d’une espèce dupliquée peu réarrangée après la 3R, comme le medaka (Figure 4.11). Ce résultat

suggère que les incongruences observées sont effectivement reliées à la présence de régions à rediploïdisation tardive. Dans la prochaine sous-partie, nous testons rigoureusement si les incongruences peuvent s'expliquer par des topologies d'arbres LORe, mieux soutenues que leur contrepartie AORE.

Confrontation directe des modèles LORe et AORE

Pour confirmer que les incongruences reportées par SCORPiOs s'expliquent effectivement par l'existence de régions à rediploïdisation tardive, j'ai, dans un premier temps, tenté d'appliquer l'approche de clustering comme présentée pour le jeu de données de Salmonidés (Matériel et Méthodes). Cependant, cette approche n'a pas permis d'identifier des régions aux topologies d'arbres distinctes sur le génome du medaka. Je propose deux explications à ce résultat. Le taux de rétention des ohnologues chez les téléostéens est beaucoup plus bas que chez les Salmonidés, ce qui induit une plus grande quantité de données manquantes, auxquelles le clustering est sensible (GORI et al. 2016). De plus, le signal majoritaire exploité par le clustering vient des longueurs de branches (GORI et al. 2016) et la saturation des substitutions dans les séquences est nécessairement plus importante ici que chez les Salmoninés, puisque les temps de divergences étudiées sont situés bien plus dans le passé.

En conséquence, nous avons validé le modèle LORe par une autre approche. Nous avons récolté l'ensemble des topologies d'arbres inférées par SCORPiOs (AORE) et les avons confrontées, via des tests de vraisemblance, aux topologies LORe correspondantes, reconstruites sous l'hypothèse d'événements de rediploïdisation indépendantes chez les Clupeocephala et les Osteoglossiformes (Matériel et Méthodes, Figure 4.12).

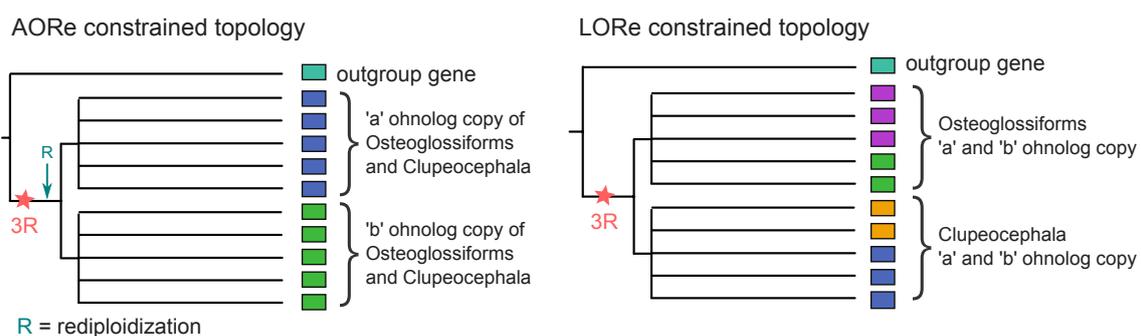


FIGURE 4.12 – Topologies contraintes AORE et LORe. La topologie AORE groupe les copies 'a' ensemble et les 'b' ensemble, sous l'hypothèse d'un seul événement de rediploïdisation ancestrale. La topologie LORe groupe les ohnologues des Osteoglossiformes dans un groupe et les ohnologues Clupeocephala dans un second groupe. A noter que la topologie LORe implique au moins un épisode de rediploïdisation indépendant chez les Osteoglossiformes et Clupeocephala mais ne fait pas explicitement d'hypothèse sur le nombre d'événements.

Ces tests de vraisemblance sont différents des tests effectués par SCORPiOs : SCORPiOs compare la topologie de l'arbre AORE prédite par l'analyse de synténie à celle de l'arbre

de gènes initial, or cette dernière n'est pas forcément une topologie LORe. Les tests ont été effectués sur un total de 5 005 arbres de gènes, soit le sous-jeu d'arbres pour lesquels la rétention d'ohnologues est suffisante pour construire des topologies AORe et LORe distinctes (Matériel et Méthodes). Pour 1 698 arbres de gènes, les tests révèlent une différence significative entre les deux modèles (AU-test à $\alpha=0.05$, SHIMODAIRA 2002, Matériel et Méthodes), permettant de prédire avec confiance 1 125 topologies LORe et 573 AORe. Ces résultats permettent effectivement de dessiner des régions LORe et AORe sur les génomes dupliqués (Figure 4.13) et confirment la rediploïdisation tardives des paires descendants des chromosomes ancestraux 3 et 11 sur toute leur longueur, comme suggéré par les taux élevés d'incongruence observés sur la Figure 4.11. De manière importante, les régions LORe qui ressortent de cette analyse (Figure 4.13) correspondent bien aux conflits séquence - synténie inférés par SCORPiOs (Figure 4.11), permettant une validation de l'occurrence de rediploïdisation retardée.

Enfin, ces résultats sont globalement en accord avec les résultats proposés par MARTIN et HOLLAND 2014, concernant la rediploïdisation des clusters de gènes *hoxa*, *hoxc* et *hoxd* (4.6). En effet, les clusters *hoxc* et *hoxd* sont compris dans des régions caractérisées par des topologies LORe tandis que *hoxa* est effectivement entouré de topologies AORe. En revanche, nous trouvons ici les clusters *hoxb* dans des régions AORe alors que MARTIN et HOLLAND 2014 avaient prédit leur rediploïdisation indépendante. Il est à noter cependant que nos résultats n'incluent pas directement les arbres des clusters *hox* (à l'exception de quelques gènes du cluster *hoxa*) : nous décrivons les régions dans lesquels se trouvent les clusters *hox*, régions définies par les arbres des gènes retrouvés dans leur voisinage. Les deux grandes différences méthodologiques avec MARTIN et HOLLAND 2014 sont les suivantes : (i) les arbres de gènes *hoxb*, *hoxc* et *hoxd* ne font pas partie de nos topologies identifiées AORe et LORe car le support statistique des topologies ne permet pas de favoriser une hypothèse sur l'autre, (ce faible support était également retrouvé dans MARTIN et HOLLAND 2014) et (ii) MARTIN et HOLLAND 2014 ne testent pas strictement de topologie LORe mais une topologie où les gènes d'Osteoglossiformes groupent avec les gènes d'espèces non-dupliquées. De fait, comme MARTIN et HOLLAND 2014, nous ne pouvons pas démontrer définitivement la rediploïdisation ancestrale ou lignée-spécifique des clusters *hoxb*. En revanche, nous montrons qu'ils se trouvent dans le voisinage de gènes effectivement ancestralement rediploïdisés.

Ce résultat représente la première mise en évidence, sur génome complet, de rediploïdisation tardive chez les poissons téléostéens après la 3R. Ces occurrences de rediploïdisation indépendantes ne peuvent être ignorées dans le contexte d'études fonctionnelles et évolutive de la génomique comparative des poissons.

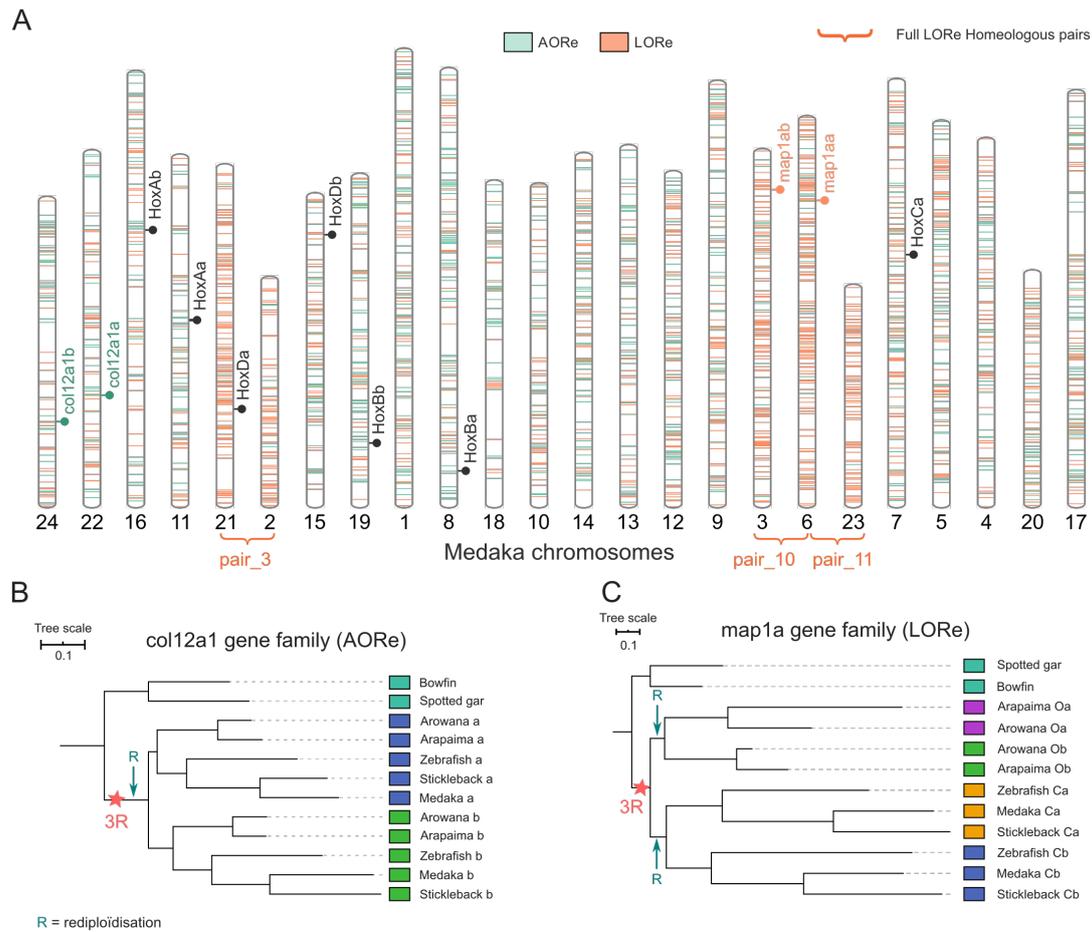


FIGURE 4.13 – Régions AORe et LORe sur le génome du medaka. A. Les gènes du médaka appartenant aux deux classes d’arbres sont représentés de couleur différentes (AORe en vert, LORe en orange). Les chromosomes 2, 3, 6, 21, 23 sont très enrichis en topologies LORe. Ils forment les paires d’homéologues suivantes : 2 - 21 (paire descendant du chromosome ancestral pré-duplication 3), 3 - 6 (paire 10) et 3 - 23 (paire 11). Les positions des 7 clusters *hox*, ainsi que celles des paires d’ohnologues des familles *map1a* et *col12a1* sont indiquées. B. Topologie AORe inférée pour la famille *col12a1*. La rediploïdisation ancestrale permet de grouper les ohnologues par copie ‘a’ et copie ‘b’. C. Topologie LORe inférée pour la famille *map1a*. Les deux épisodes de rediploïdisation sont indiqués par une flèche bleue, les suffixes ‘Ca’/‘Cb’ et ‘Oa’/‘Ob’ indiquant les copies ‘a’ et ‘b’ des Clupeocephala et Osteoglossiformes, respectivement.

4.3 Matériel et Méthodes

4.3.1 Arbres de gènes

Les arbres de gènes de Salmonidés et de Téléostéens ont été obtenus à partir des familles de gènes établies pour le jeu total de 101 espèces et 74 poissons téléostéens, présenté dans le chapitre précédent (voir Matériel et Méthodes du manuscrit présenté). Pour les Salmonidés, les sous-arbres contenant les 5 espèces et leur groupe externe non-dupliqué, le brochet (*Esox lucius*), ont été directement extraits des arbres complets contenant les 101 espèces. Pour les Osteoglossoscephalai, les familles de gènes ont été réduites aux 6 espèces Osteo-

glossocephalai sélectionnées (Figure 4.10) et les 27 espèces non téléostes, pour recalculer des arbres réduits. Ces arbres réduits ont été reconstruits par la même procédure que les arbres contenant 101 espèces présentés au chapitre précédent (inférence des arbres avec TreeBeST suivie d'une correction itérative avec SCORPiOs, voir Matériel et Méthodes du manuscrit présenté).

4.3.2 Identification des arbres incongruents entre séquence et synténie.

Brièvement, la correction des arbres par le processus itératif de SCORPiOs permet de corriger les arbres en plusieurs itérations, au fur et à mesure que les patrons de synténie sont précisés par des arbres de plus en plus corrects. Dans ce mode itératif, SCORPiOs ne considère que les régions pour lesquelles un arbre a été amélioré à l'itération précédente. Afin d'obtenir un jeu complet des topologies incongruentes à la fin de la correction, nous avons relancé SCORPiOs une deuxième fois sur les arbres et répertorié toutes les topologies identifiées par SCORPiOs comme restant incongruentes avec la synténie après la correction itérative effectuée. Ce jeu correspond aux arbres appelés « incongruents » dans la partie Résultats, où l'information de séquence et de synténie n'a pas pu être réconciliée dans un même modèle : la topologie inférée par la synténie induit une baisse significative de la vraisemblance du modèle d'évolution des séquences.

4.3.3 Clustering des arbres de gènes

Les arbres de gènes de Salmonidés ont été clusterisés à l'aide de l'outil treeCl (GORI et al. 2016). Afin de réduire le temps de calcul d'une part et de minimiser le déséquilibre de taille attendu entre clusters d'autre part, 4 000 arbres de gènes ont été échantillonnés parmi le sous-jeu de 8 186 arbres avec le taux de rétention d'orthologues le plus élevé. En effet, le clustering est effectué sur la base d'une matrice de distance entre toutes les paires d'arbres, la complexité en temps de l'établissement de cette matrice est donc de l'ordre $O(n^2)$ avec n le nombre d'arbres. Pour limiter le déséquilibre de taille entre partitions, nous avons sur-échantillonné les arbres incongruents ($n=715$, soit 95% des arbres incongruents parmi le jeu des 8 186 arbres) par rapport aux arbres congruents ($n=3 285$, soit 50% des arbres congruents parmi le jeu des 8 186 arbres), sous l'hypothèse qu'ils représentaient des clusters topologiques distincts. En accord avec les recommandations, les distances entre arbres ont été calculées en utilisant la distance euclidienne, qui prend en compte les différences en terme de longueur de branche et de topologie, suivi d'un clustering spectral. Pour les calculs de distances entre arbres au contenu en gènes différent, nous avons utilisé l'approche décrite dans (GORI et al. 2016) : les arbres sont réduits à leur ensemble de feuilles communes avant comparaison, en prenant soin de conserver les longueurs de branche. Enfin, pour tenir compte de la présence de la duplication 3R dans les arbres de gènes, nous

avons légèrement adapté le calcul de distance entre arbres. En pratique, il s'agit d'intégrer, pour chaque espèce dupliquée, l'existence potentielle de deux feuilles ('1' et '2'). Lors de la comparaison de deux arbres, toutes les combinaisons possibles d'assignation de '1' et '2' sont essayées et la distance la plus faible est retenue.

Différentes valeurs de k ont été testées (2 à 7), et la valeur de $k = 3$ a été sélectionnée sur la base du regroupement des clusters de topologies inférées sur le génome du saumon atlantique. En effet, les auteurs ont montré que les techniques classiques (silhouette, inertie) d'estimation du nombre de clusters n'étaient pas adaptées au clustering des arbres. La procédure recommandée consiste à concaténer les séquences des mêmes clusters et inférer un arbre par cluster afin d'estimer le gain en terme de vraisemblance ajouté par l'addition de nouveaux clusters. Dans notre cas, le nombre de familles rendrait cette étape computationnellement rédhibitoire. Une façon de procéder aurait été de sous-échantillonner les clusters. Pour l'étude de la Ss4R, la sélection manuelle du nombre de clusters apparaît justifiable au vue des connaissances déjà accumulées dans la littérature et retrouvée avec $k = 3$. Nous avons choisi de réserver cette approche de sous-échantillonnage pour une étude rigoureuse de la 3R, mais le clustering des topologies n'a pas fonctionné, comme évoqué dans la partie Résultats.

4.3.4 Lissage par modèles de Markov cachés

En nous appuyant sur les clusters d'arbres de gènes inférés par le clustering, nous avons ajusté des modèles de Markov cachés le long du génome du saumon atlantique avec la librairie python 'hmmlearn', en fixant la matrice de transitions de façon à favoriser les transitions vers le même état (Figure 4.9), afin réaliser un lissage des données discrètes observées. Le nombre d'états cachés a été sélectionné sur la base du critère d'information bayésien (CELEUX et DURAND 2008), permettant d'obtenir le modèle de meilleur fit présenté en Figure 4.9.

4.3.5 Tests de vraisemblance

Pour les 5 005 arbres de gènes pour lesquels une topologie AORe et LORe différente pouvait être inférée (rétention d'ohnologues dans au moins une espèce, ou pertes différentielles), des tests de vraisemblance ont été effectués afin de comparer ces deux modèles. Trois arbres, l'arbre de maximum de vraisemblance non-contraint (ML), la topologie AORe et la topologie LORe ont été comparés à travers le test de vraisemblance AU (SHIMODAIRA 2002), à $\alpha = 0.05$. Les résultats des tests ont été classés en 3 catégories : (i) aucune différence significative, dans le cas où les modèles LORe et AORe sont tous les deux inclus ou tous les deux exclus de l'intervalle de confiance autour de l'arbre ML, (ii) rejet de la topologie LORe au profit de l'AORe quand cette dernière est incluse dans l'intervalle alors

que la LORe est exclue et (iii) réciproquement pour la rejet de l'AORe. Seules les topologies des classes AORe (ii) et LORe (iii) sont représentées en Figure 4.13.

4.4 Discussion

Les résultats présentés dans ce chapitre représentent la première preuve, à l'échelle de génomes complets, de l'existence de recombinaison entre homéologues dans l'ancêtre Osteoglossocephalai, soit au moins jusqu'à ~ 60 millions d'années après la 3R. La seule étude préalable abordant ce sujet avait restreint son champ d'étude aux gènes *hox*, soit une seule famille de gènes. Nous confirmons ici que les clusters de gènes *hoxc* et *hoxd* ont été soumis à une rediploïdisation tardive, persistant dans la lignée Osteoglossocephalai.

Cependant, il faut noter que quelques limitations colorent nos observations. En extrapolant naïvement les résultats obtenus sur les 1 698 arbres testés à l'ensemble des familles, les régions non-rediploïdisées représenteraient 2/3 du génome de l'ancêtre Osteoglossocephalai (1 125 sur 1 698 familles). Ce chiffre est nécessairement sur-estimé car l'échantillon des topologies testées exclut les familles présentant une perte de la même copie de gène dans toutes les espèces. Or, on peut s'attendre, selon le principe de parcimonie, à ce que la majorité des ces pertes soient communes et représentent des cas de résolution ancestrale des ohnologues concernés. Nos résultats suggèrent également 3 paires de chromosomes homéologues sur les 13 qui recombinaient encore sur toute leur longueur dans l'ancêtre Osteoglossocephalai, ce qui constitue $\sim 20\%$ des gènes de ce génome ancestral (3 640 familles sur ces chromosomes pour 19 349 au total). Dans le futur, une délimitation plus précise des régions LORe et AORe devra permettre de préciser ce large intervalle [20% - 66%] décrivant la proportion du génome Osteoglossocephalai non-rediploïdisé. De plus, nous avons ici uniquement testé une topologie contrainte LORe générale, en faisant l'hypothèse de rediploïdisations indépendantes post-Osteoglossocephalai : une ou plusieurs dans le clade des Osteoglossiformes et une ou plusieurs dans le clade des Clupeocephala. A l'avenir, une analyse plus fine des topologies d'arbres devrait permettre de préciser les patrons spatio-temporels de la rediploïdisation dans chacun de ces clades. De plus, l'inclusion de génomes d'Elopomorphes, divergeant avant les Osteoglossiformes à une date estimée à 50 millions d'années après la 3R, sera cruciale à la précision et validation des patrons observés. Dans le cadre du projet « Genofish », dans lequel s'inscrit mon projet de thèse, des génomes d'Elopomorphes sont séquencés *de novo* et prochainement intégrés à notre jeu de données afin de préciser ces résultats.

Le retard de la rediploïdisation après la duplication 3R pose à nouveau la question de l'origine de cette dernière par allo- ou auto-tétraploïdisation. L'argument principal pour l'auto-polyploïdisation des Salmonidés repose sur ce même retard de la rediploïdisation jusqu'à 95 millions d'années après la Ss4R. Des exemples venant de la génomique des plantes

révèlent cependant des exemples de rediploïdisation tardive dans le cas d'allopolyploïdie (dites « segmentales »), même si le comportement méiotique attendu à aussi long terme reste difficile à prévoir (XIONG, GAETA et PIRES 2011 ; CHESTER et al. 2012 ; MASON et WENDEL 2020). Les études chez les plantes ont également mené à dériver deux grandes classes de conséquences génomiques attendues selon l'origine par allo- ou autopolyploïdie (GARSMEUR et al. 2014). Dans le cadre d'une allo-tétraploïdie, la différence entre les sous-génomes parentaux induit l'établissement d'un sous-génome dominant en terme d'expression et de rétention des gènes dupliqués. Dans le cas d'auto-tétraploïdie, l'absence de différence entre les génomes parentaux induit des pertes de gènes stochastiques sur l'un ou l'autre chromosome dupliqué. De manière générale, les autopolyploïdes sont beaucoup moins étudiés et aucun cas d'autopolyploïdie avec une rétention biaisée n'a été reporté, même si l'on peut imaginer qu'un biais de rétention pourrait survenir en cas de forte hétérozygotie des génomes parentaux.

Chez les Vertébrés, les génomes allotétraploïdes issus de la 4R des carpes représentent un exemple qui semble suivre ces prédictions : on observe un biais d'expression des deux sous-génomes parentaux (XU et al. 2019). Néanmoins, il reste difficile d'estimer comment ces prédictions se généralisent aux Vertébrés et comment elles se manifestent sur le long terme. Le génome de l'esturgeon, ancien polyploïde, est un second exemple vertébré suivant ces prédictions : l'absence de dominance de sous-génome et les phylogénies des populations de TE des chromosomes dupliqués indiquent son origine auto-tétraploïde (DU et al. 2020). Dans le cas de la 3R, la connaissance des régions à rediploïdisation tardive est cruciale pour examiner la question de son origine par allo- ou autotétraploïdisation. En effet, dans le cas d'une allopolyploïdie, il n'est pas évident qu'il serait possible d'observer une différenciation génomique entre les sous-génomes après 320 millions d'années d'évolution (populations de TE, contenu en GC, biais de rétention ou biais d'expression). De plus, la recombinaison entre homéologues et le remplacement de segments génomiques d'un sous-génome par un autre aura également masqué cette différenciation initiale des génomes parentaux (MASON et WENDEL 2020). Dans le futur, les analyses visant à résoudre la question de l'allo ou autopolyploïdie de la 3R devront se concentrer sur les régions rediploïdisées tôt après la 3R pour minimiser cet effet. L'objectif ultime de la poursuite de notre étude présentée ici serait de développer un pipeline automatique de détection de régions LORe, possible à appliquer à n'importe quel événement de duplication complète : chez les plantes, carpes ou encore les poissons à ventouse. Tirer profit de la diversité de la biologie des organismes concernés et de leurs dates de duplications et modes d'origine devrait permettre d'observer si des principes globaux se dégagent concernant les dynamiques de la rediploïdisation.

Comparer des génomes paleopolyploïdes et identifier leurs gènes orthologues nécessite de prendre en compte leur histoire de rediploïdisation. La caractérisation précise de ces événements biologiques significatifs aura un impact décisif sur le succès de la génomique

comparative des clades regroupant d'anciens polyploïdes, de manière générale, et donc des poissons en particulier. En effet, puisque le modèle LORe implique des relations d'homologie d'un nouveau type, il sera nécessaire de développer de nouveaux modèles permettant d'appréhender ces relations de 'tétralogie', comme proposé par MARTIN et HOLLAND 2014. En particulier, il est crucial de prendre en compte ces relations de tétralogie lors du transfert d'annotations fonctionnelles, puisque leur rediploïdisation indépendante peut avoir donné lieu à l'évolution de fonctions distinctes. En résumé, cette caractérisation préliminaire des dynamiques de rediploïdisation chez les téléostéens est vouée à contribuer à une meilleure compréhension de la structure, fonction et évolution des génomes de poissons.

Chapitre 5

Discussion

5.1 Résumé des résultats principaux

J'ai présenté, dans les chapitres précédents, trois études complémentaires visant à améliorer notre compréhension de l'évolution des gènes et génomes de poissons après leur événement commun de duplication complète. La première étude s'est concentrée sur la résolution du problème méthodologique de la reconstruction d'arbres de gènes en présence de duplications complètes. L'application de cette nouvelle méthode permet une meilleure caractérisation de l'évolution du génome codant des téléostéens et sert d'ancrage pour étendre ces connaissances à l'évolution des chromosomes issus de la duplication. Enfin, la mise en relation des arbres de gènes avec les chromosomes homéologues permet de caractériser leur comportement méiotique ~60 millions d'années après la duplication complète. En résumé, mon travail s'articule autour du développement d'une nouvelle méthodologie et l'établissement d'une nouvelle ressource, qui peuvent être combinées pour mieux comprendre la structure et la fonction des génomes. Je décris dans la suite les liens qui unissent ces différentes études et discute des perspectives qu'elles ouvrent concernant l'évolution du génome non-codant après duplication complète.

5.2 Validation de la pertinence biologique de SCORPiOs

J'ai présenté, dans le chapitre 2, SCORPiOs, un nouvel outil développé dans le but de reconstruire des arbres de gènes en présence de duplication(s) complète(s) de génome. Le fonctionnement de SCORPiOs est fondé sur deux grandes composantes : (i) une analyse des patrons de synténie doublement conservée, par paires d'espèces, afin de prédire des relations d'orthologie et de paralogie et (ii) l'intégration des relations prédites dans le modèle de l'arbre de gènes. SCORPiOs prend en entrée la phylogénie des espèces avec la position de la duplication complète, un jeu d'arbres de gènes initiaux, les alignements multiples correspondants, ainsi que la position des gènes dans les génomes. En l'absence d'arbres de gènes pré-calculés, SCORPiOs peut également reconstruire un jeu d'arbres de départ avec Tree-

BeST (VILELLA et al. 2009). En sortie, SCORPiOs renvoie un jeu d'arbres de gènes corrigés, cohérents avec la position connue d'une duplication complète et les patrons de synténie doublement conservée observés. Ainsi SCORPiOs intègre deux types d'information phylogénétique : l'évolution moléculaire des séquences et l'organisation génomique des gènes.

Une limite de SCORPiOs est son absence de validation par simulation, qui aurait permis de mieux caractériser les taux de vrais et faux positifs dans les groupes d'orthologie dérivés de l'analyse des patrons de synténie. Ces simulations permettraient également d'évaluer le comportement de SCORPiOs au regard de différents paramètres (par exemple, la fragmentation des génomes, la proportion d'erreurs dans les arbres initiaux, le nombre d'espèces, les taux de réarrangements des génomes...), dans le cadre parfait des simulations. Cependant, l'absence de méthode bien établie pour simuler de manière réaliste l'évolution de l'ordre des gènes et de leur séquence après duplication complète rend cette tâche compliquée. Il aurait été possible de compléter un simulateur existant en implémentant les duplications complètes : doubler les chromosomes est une opération simple. En revanche, il s'agirait également de simuler le processus de pertes de copies de gènes dupliqués. Pour simuler les pertes, on peut s'appuyer sur les résultats de INOUE et al. 2015, qui ont montré, chez les poissons téléostéens, que la dynamique des pertes suit une décroissance exponentielle en deux-phases : une phase de pertes massives suivie d'une stabilisation. Récemment, Davin et al. ont développé « Zombi », le premier simulateur d'évolution multi-échelle, pour simuler itérativement un arbre d'espèces, sur lequel évoluent les génomes (par réarrangements génomique et évolution du contenu en gènes) et définissant les arbres des familles de gènes à partir desquels des séquences peuvent également être simulées (DAVIN et al. 2020). Il serait possible de s'inspirer de cette méthode pour réaliser nos simulations : il s'agirait tout de même d'introduire les duplications complètes de génome et de transformer les génomes circulaires de Zombi en génomes linéaires multi-chromosomes. Les alignements de séquences, les génomes et la phylogénie d'espèces générés en sortie de Zombi serviraient d'entrée à SCORPiOs, permettant donc de comparer les orthologues ou arbres prédits aux vrais arbres utilisés pour simuler les séquences. Ainsi, bien que de nouveaux développements laissent entrevoir la possibilité d'une stratégie de simulation applicable à la validation de SCORPiOs, la mise en place d'une telle méthode apparaît non-triviale.

En dépit de cela, les différentes analyses que j'ai présentées valident amplement SCORPiOs sur données réelles, du point de vue de sa pertinence pour l'étude de processus biologiques. Nous avons notamment montré que SCORPiOs permettait de :

- prédire des orthologues fonctionnellement plus similaires, à travers le proxy de leur expression, comme attendu du fait de leur divergence plus récente que les ohnologues ;
- identifier des enrichissements fonctionnels des gènes ohnologues masqués par les erreurs dans les arbres et proposer une contribution de la rétention de gènes dupliqués

- à des innovations évolutives du clade téléostéen ;
- transférer les relations d'orthologie et de paralogie des gènes au niveau de segments génomiques afin de caractériser l'évolution des chromosomes dupliqués ;
- identifier des événements de conversion génique entre gènes ohnologues et ainsi explorer les patrons spatio-temporels de la rediploïdisation.

5.3 Unification des résultats dans la nomenclature des gènes

Au chapitre 3, j'ai présenté l'application de SCORPiOs à un jeu de 74 poissons téléostéens, et l'établissement d'une ressource répertoriant les relations d'orthologie et de paralogie des gènes et segments génomiques de ces espèces. Au chapitre 4, j'ai décrit les résultats d'une étude préliminaire suggérant l'existence de rediploïdisation indépendante de certains chromosomes dupliqués dans deux grands clades de poissons téléostéens (Osteoglossiformes et Clupeocephala). Ces résultats représentent une connaissance précieuse pour les études comparatives chez les poissons, en particulier pour le transfert des annotations et ontologies des gènes du poisson-zèbre aux espèces non-modèles. Actuellement, en dehors des gènes du poisson-zèbre régis par ZFIN (RUZICKA et al. 2019), aucune règle ni ressource ne gouverne la nomenclature des gènes de poissons téléostéens de manière générale. Nous proposons, à partir de nos résultats, d'enrichir les annotations de gènes présentes dans ZFIN en les étendant, de manière stable, aux autres espèces (Figure 5.1).

Dans un premier temps, il est nécessaire que l'information d'orthologie et de paralogie soit reflétée dans cette nomenclature : les gènes descendant d'une même copie ancestrale, ou gènes orthologues, doivent porter le même suffixe (« a » ou « b » si on suit la nomenclature ZFIN, ou « TGD_a », « TGD_b » pour "Teleost Genome Duplication"). Pour cela, nous proposons de nous appuyer sur les arbres de gènes inférés par SCORPiOs. De plus, il est également désirable que l'information de l'origine synténique commune des gènes soit apparente : au sein d'une même espèce, les gènes voisins descendant d'un même chromosome dupliqué ancestral doivent être facilement identifiables, c'est-à-dire porter le même suffixe « a » ou « b ». Cette information peut également être complétée par une annotation évolutive désignant le chromosome ancestral d'origine. Éventuellement, si des analyses subséquentes montrent que la duplication complète des téléostéens est une allopolyploïdie, ces « a » et « b » pourraient identifier les deux sous-génomés parentaux. Enfin, il est important de prendre en compte les occurrences de rediploïdisation indépendante, puisque les fonctions de ces gènes « tétralogues » ne seront pas nécessairement conservées entre espèces. Ceci pourrait être indiqué par l'ajout d'une lettre représentant le nom des clades rediploïdisés indépendamment : par exemple, dans le cas d'une rediploïdisation indépendante chez les Clupeocephala et les Osteoglossiformes, les gènes pourraient porter un suffixe « Ca », « Cb »

et « Oa », « Ob ». Nous réfléchissons actuellement à la mise en place d'un tel système, en collaboration avec John Postlethwait (University of Oregon) et Monte Westerfield (ZFIN, University of Oregon). Ce système pourra également être étendu pour annoter les gènes des espèces ayant subi des duplications complète supplémentaires, comme les carpes et les saumons.

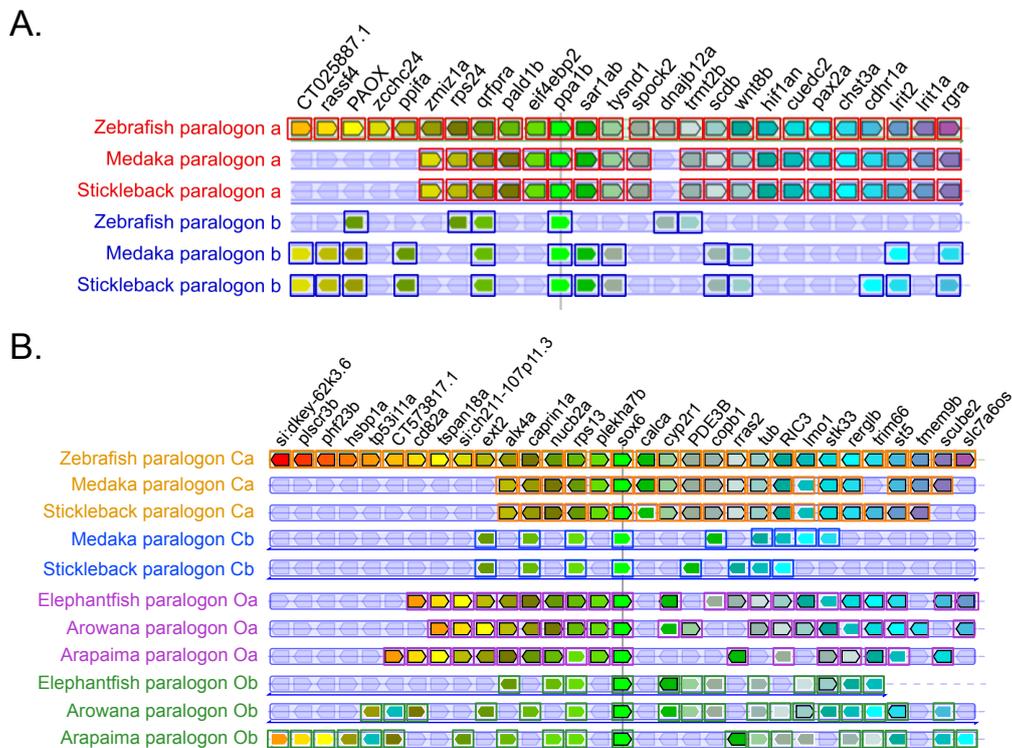


FIGURE 5.1 – Annotation évolutive des gènes de poissons téléostéens. A. Annotation proposée avec l'appui de la carte de paralogie, visant à annoter les gènes descendant d'un même segment génomique post-duplication (ou paralogon) avec le même suffixe. Les noms de gènes au-dessus des pistes correspond aux noms ZFIN actuels pour les gènes du paralogon A du poisson-zèbre : il apparaît clairement un mélange des suffixes 'a' et 'b'. B. Annotation proposée pour le cas particulier d'une région rediploïdisée indépendamment chez les Clupeocephala et les Osteoglossiformes. La famille du gène *sox6* (gène vert clair, au centre) correspond à une topologie d'arbre « LORE » identifiée dans le chapitre 4.

5.4 Perspectives pour l'étude du génome non-codant

Mon travail s'est principalement concentré sur la caractérisation de l'évolution du génome codant après duplication complète. Néanmoins, de nombreuses questions restent ouvertes concernant la manière dont les duplications complètes façonnent l'expression des gènes dupliqués. En particulier, la contribution des modèles de sub et néo-fonctionnalisation à la rétention des gènes ohnologues chez les poissons téléostéens est encore floue. De plus, les forces sélectives qui agissent sur les niveaux et patrons d'expression des gènes ainsi que

leurs conséquences sur le génome régulateur sont mal connues. Je propose des pistes pour l'exploration de ces questions.

Évolution de l'expression des gènes après duplication complète

Il est généralement admis que la conséquence principale des duplications complètes est une réduction asymétrique du niveau d'expression : un ohnologue voit son expression diminuée, ce qui le dirige vers la route de la non-fonctionnalisation (PALA et al. 2010 ; LIEN et al. 2016 ; SANDVE, ROHLFS et HVIDSTEN 2018 ; XU et al. 2019).

Récemment, de nouvelles méthodes ont été développées afin de fournir un cadre statistique pour tester des hypothèses sur les régimes de sélection agissant sur l'expression des gènes (notamment « EVE », Expression Variance Evolution, ROHLFS et NIELSEN 2015). Ces méthodes modélisent l'évolution de l'expression comme un trait continu, changeant au cours du temps sur les branches d'un arbre phylogénétique. Les modèles proposés permettent de confronter l'hypothèse nulle d'une évolution neutre du niveau d'expression (modèle « BM », mouvement brownien) contre celle d'une évolution contrainte vers un niveau optimal (modèle « OU », Ornstein-Uhlenbeck). De plus, l'enrichissement du modèle OU par l'ajout de différents optimaux permet de tester l'hypothèse d'un glissement (ou « shift ») adaptatif du niveau d'expression optimal sur une branche spécifique. Une étude récente a appliqué pour la première fois EVE (ROHLFS et NIELSEN 2015) à l'étude de l'évolution de l'expression des gènes après duplication complète, sur la Ss4r des Salmonidés (GILLARD et al. 2020, preprint). Les auteurs démontrent rigoureusement que l'expression d'au moins un des deux ohnologues est abaissée dans la grande majorité des familles (75%). Le plus souvent, l'évolution est asymétrique : une seule copie voit son expression baissée et celle-ci est plus susceptible d'être pseudogénisée.

La qualité des arbres de gènes reconstruits par SCORPiOs permet d'envisager l'exploration de l'évolution de l'expression des gènes à la suite de la duplication 3R. De plus, en combinaison avec les informations topologiques données par la carte de paralogie, il serait également possible d'évaluer si, dans le cas d'une évolution asymétrique des ohnologues, celle-ci est corrélée à un des chromosomes d'une paire d'homéologues. Ces analyses nécessitent néanmoins des ressources transcriptomiques (RNA-seq) contenant plusieurs réplicats biologiques, afin de limiter les faux positifs causés par des gènes à expression très variable (ROHLFS et NIELSEN 2015). Les ressources transcriptomiques chez les téléostéens sont pour le moment limitées à quelques espèces et sans réplicat (Phylofish, PASQUIER et al. 2016).

Évolution du paysage régulateur après duplication complète

Les questions de l'évolution du paysage régulateur après duplication complète restent largement inexplorées chez les Vertébrés. Quels sont les taux de rétention et pertes des

séquences régulatrices après la duplication, quelle est la contribution des enhancers et promoteurs aux patrons de néo et sub-fonctionnalisations et par quels mécanismes passent-ils ? Quelles forces évolutives gouvernent l'évolution de l'activité des séquences régulatrices ?

Chez les poissons téléostéens, la difficulté d'explorer ces questions est en partie due à la faible quantité de ressources fonctionnelles disponibles, et à la complexité du transfert des annotations des génomes humain et de souris. Comme détaillé au chapitre 1 (voir paragraphe 1.1.2), l'annotation du génome fonctionnel non-codant passe par la combinaison d'approches de génomique comparative, permettant d'identifier des éléments conservés non-codant (CNE), et de génomique fonctionnelle permettant d'identifier les séquences régulatrices actives (régions de chromatine ouvertes et marques d'histone caractéristiques de régions régulatrices actives). Les annotations fonctionnelles sont très variables entre tissus, stade du développement et espèces, et ne peuvent donc pas être directement extrapolées entre espèces. De fait, les analyses comparatives reposent sur l'identification de CNEs et, si elles sont disponibles, la comparaison de leur annotation fonctionnelle entre espèces. Cependant, l'identification de CNEs entre mammifères et poissons téléostéens est également compliquée. En effet, elle repose sur l'alignement complet de génomes, plus correct à faible distance phylogénétique et en absence de régions dupliquées (ARMSTRONG et al. 2019b). Pour répondre à cette difficulté, plusieurs études utilisent le génome du lépisosté tacheté comme intermédiaire aux comparaisons homme - poisson-zèbre, car il s'agit d'une espèce non-dupliquée, caractérisée par un taux d'évolution lent (BRAASCH et al. 2016 ; YUAN et al. 2018 ; CLÉMENT et al. 2020).

Une quantité croissante de ressources fonctionnelles (marques d'histone, méthylation, chromatine ouverte) existent chez le poisson-zèbre, collectées et standardisées par le consortium DanioCode (TAN, ONICHTCHOUK et WINATA 2016). Une solution pour l'étude de l'évolution du génome non-codant au sein des téléostéens consisterait à identifier un jeu d'éléments conservés non-codant entre espèces, considérer le sous-jeu fonctionnellement caractérisé chez le poisson-zèbre comme des séquences régulatrices actives et étudier leur patron de pertes et rétention après la duplication. Néanmoins, comme pour les gènes, l'existence de plusieurs copies des CNEs suite à la duplication complète pose le problème de l'identification correcte des orthologues et paralogues. La carte de paralogie, qui décrit les régions orthologues entre génome de poissons, pourrait servir de guide et/ou de validation pour l'annotation de CNEs entre génomes de poissons.

Dans le futur, il est attendu que les analyses comparatives du génome non-codant se multiplient et contribuent significativement à la mise en lumière des bases génomiques des phénotypes. Un exemple parmi les phénotypes d'intérêt du poisson-zèbre est sa capacité de régénération de certains tissus. Les éléments régulateurs des réseaux de régulation impliqués dans la régénération ont été récemment caractérisés chez le poisson-zèbre et chez

le killi africain (WANG et al. 2020). Ces deux espèces sont capables de régénérer leurs nageoires après amputation, ainsi que leur tissu cardiaque après gelure. D'un point de vue évolutif, il n'est pas clair si la capacité de régénération du cœur chez les Téléostéens a été acquise ancestralement ou si elle est le résultat d'évolution convergente : le medaka, phylogénétiquement proche du killi africain, est incapable de régénérer son cœur à la suite d'une blessure (LAI et al. 2017 ; JAŻWIŃSKA et BLANCHOU 2020). Nous avons observé que les gènes ohnologues retenus en deux copies chez le poisson-zèbre et retournés à l'état de copie unique chez le medaka étaient enrichis en fonctions cardiaques et voies de signalisation actives lors du processus de régénération. Ces analyses préliminaires suggèrent que ce phénotype aurait été acquis suite à la duplication complète, à travers la rétention et la néofonctionnalisation de gènes ohnologues. Par la suite, l'intégration et l'analyse comparative de données fonctionnelles (transcriptomiques et épigénomiques) de différentes espèces de téléostéens pourraient permettre de préciser les réseaux de gènes mis en jeu et d'évaluer leur conservation chez les poissons téléostéens.

Chapitre 6

Perspectives

Dans ce dernier chapitre, je resitue mon travail dans le contexte plus général de la génomique comparative des Vertébrés. Dans un premier temps, je décris l'état des connaissances concernant les duplications complètes survenues à la base de l'arbre des Vertébrés. Je discute de la faisabilité et pertinence de l'analyse de synténie effectuée par SCORPiOs pour mieux comprendre ces duplications. De manière évidente, les patrons d'organisation des gènes dans les génomes, ou synténie, ne renseignent pas uniquement sur les duplications complètes. Par la suite, je discute plus généralement de l'évolution future des méthodes de la génomique comparative, où il apparaît une nécessité croissante d'utiliser des informations complémentaires aux séquences. La synténie, encore sous-exploitée, offre un point de vue supplémentaire pertinent aux différents niveaux des inférences phylogénomiques : la résolution des phylogénies d'espèces, l'explication des désaccords entre arbres d'espèces et arbres de gènes (ou réconciliation), et l'amélioration des arbres de gènes. Je détaille ici individuellement chacun de ces trois grands défis et comment ils peuvent être enrichis par l'apport de la synténie.

6.1 Contribution à la compréhension des génomes de Vertébrés

Les duplications complètes de la lignée Vertébré : état des connaissances

Les Vertébrés descendent de deux événements de duplication complètes successifs (1R-2R), qui ont significativement contribué à la complexification des génomes via l'amplification du répertoire de gènes et la sophistication de leurs interactions régulatrices (MARLÉTAZ et al. 2018). De nombreuses familles de gènes se sont diversifiées à travers leurs duplications par les 1R-2R. L'exemple le plus marquant est celui des clusters de gènes *Hox*, passés de 1 à 4 copies, et reliés à la complexification du plan d'organisation morphologique des Vertébrés (WAGNER, AMEMIYA et RUDDLE 2003). De plus, ces duplications auraient contribué à la mise en place du système immunitaire adaptatif, en permettant la diversification

des gènes du complexe majeur d'histocompatibilité (KAUFMAN 2018). Enfin, l'intégration de copies surnuméraires de facteurs de transcription dans les réseaux de gènes du développement, et en particulier les paralogues du gène *FoxD*, aurait participé à l'établissement du système nerveux (HOLLAND 2015). De manière importante, dans le génome humain, ces événements ont également mené à l'expansion des familles d'oncogènes (SINGH et al. 2012, présenté au chapitre 1, paragraphe 1.3.2).

Au cours des dernières années, différentes études se sont attelées à caractériser la datation, le mécanisme d'origine, et les conséquences fonctionnelles des duplications 1R-2R. Ces études sont rendues difficiles par l'ancienneté des duplications : beaucoup de paralogues ont été perdus et les séquences des gènes retenus ont grandement divergé. De plus, la double duplication rend compliquée l'identification des séquences descendant de chaque événement. Un autre obstacle est que le nombre d'espèces des groupes externes non-dupliqués, phylogénétiquement proches des duplications, est assez restreint. Enfin, Furlong et Holland ont proposé que la difficulté d'identifier les gènes issus des 1R-2R à partir des arbres de gènes était potentiellement amplifiée par la présence de recombinaison homéologue prolongée après la 1R et la 2R (FURLONG et HOLLAND 2002). Comme cela a été démontré chez les Salmonidés et comme nous l'avons également observé dans le cas de la 3R des poissons téléostéens (voir le chapitre 4), la rediploïdisation peut s'étendre sur des dizaines de millions d'années après duplication complète et ainsi obscurcir les relations d'orthologie entre gènes dupliqués de différentes espèces.

Longtemps remise en cause, l'occurrence des duplications 1R-2R est aujourd'hui bien supportée (VAN DE PEER, MAERE et MEYER 2010). Cependant, leur positionnement dans la phylogénie des Vertébrés reste débattu. La 1R est estimée commune à toutes les espèces de Vertébrés, tandis que la 2R a été prédite alternativement dans la lignée précédant la divergence des Agnathes (lamproies et myxines) (SACERDOT et al. 2018), ou après leur divergence mais commune à l'ensemble des Gnathostomes (SIMAKOV et al. 2020) (Figure 6.1). A travers la reconstruction du génome de l'ancêtre Amniota et son partitionnement en tétrades descendant des chromosomes Vertébrés pre-duplication, Sacerdot et al. mettent en évidence une correspondance 1 à 4 entre les chromosomes ancestraux Vertébrés et le génome de la lamproie. L'explication la plus parcimonieuse face à cette observation est que les deux événements 1R-2R sont partagés avec la lamproie. Simakov et al., quant à eux, tirent profit d'un nouvel assemblage chromosomique du génome de l'Amphioxus, un groupe externe du clade des Vertébrés. A travers la comparaison du génome de l'Amphioxus aux génomes de Vertébrés, les auteurs retracent les événements de fusions chromosomiques survenus avant la 2R. Ils observent que ces fusions ne sont pas retrouvées chez les lamproies, ce qui implique que leur divergence est antérieure à la duplication. Dans le futur, la génération d'assemblages chromosomiques des génomes de lamproies, d'autres Agnathes comme les myxines, ainsi que de poissons cartilagineux, devrait permettre de préciser les

premières étapes de l'histoire évolutive des chromosomes vertébrés.

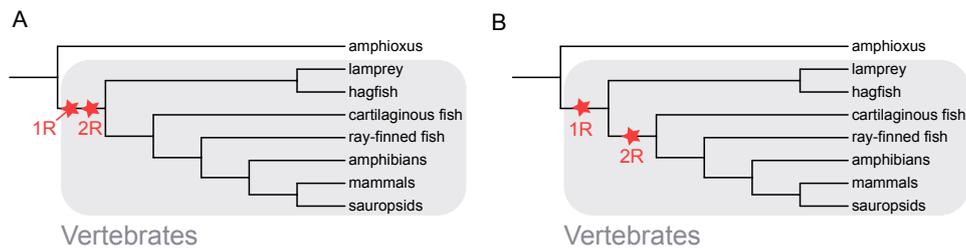


FIGURE 6.1 – Hypothèses de positionnement des duplications 1R-2R. A. Hypothèse de deux événements successifs, communs à tous les Vertébrés. B. Hypothèse d'une première duplication complète, 1R, commune aux Vertébrés et d'une seconde, 2R, commune aux Gnathostomes.

De même, l'origine par allo ou auto-polyploïdisation des 1R-2R est difficilement caractérisable. Simakov et al. proposent un scénario d'autotétraploïdisation (1R) suivi d'une allotétraploïdisation (2R). Les auteurs s'appuient sur l'observation d'un biais de rétention de gènes sur les différents segments dupliqués, biais retrouvé uniquement entre segments issus de la 2R. Chez les plantes, un tel biais de rétention entre chromosomes dupliqués est considéré être une indication d'allopolyploïdisation (GARSMEUR et al. 2014). La précision des conséquences des mécanismes d'allo- et autopolyploïdisation sur les génomes de Vertébrés de manière générale, et chez les poissons téléostéens en particulier, contribuera à l'évaluation de la validité de ces différentes hypothèses.

Perspectives pour l'application de SCORPiOs aux 1R-2R

En dehors de son application chez les poissons, nous n'avons pas évalué la pertinence de SCORPiOs pour l'étude d'autres événements de duplications complètes. Dans l'implémentation actuelle de SCORPiOs, l'appliquer à la 1R-2R nécessite de disposer d'un génome non-dupliqué pour chaque duplication, dont la position pré-duplication est bien supportée et dont le génome est bien assemblé. En effet, SCORPiOs ne permet pas directement de tester des hypothèses quant à la position des duplications complètes sur une phylogénie. Un second point à prendre en considération est la dégradation de la synténie, attendue beaucoup plus importante après les 1R-2R que dans le cas de la 3R.

Dans l'optique d'évaluer la robustesse de SCORPiOs face à des génomes plus réarrangés, nous avons utilisé le poulet et la souris comme groupes externes pour corriger les nœuds de duplication 3R. Ces expériences ont révélé un accord global des corrections effectuées par rapport aux résultats publiés dans PAREY et al. 2020, qui utilisent le lepisosté tacheté comme groupe externe (72% de corrections identiques avec le poulet comme groupe externe, 68% avec la souris). L'effet majeur de l'utilisation d'un groupe externe plus distant est une diminution du nombre des corrections effectuées (1 643 avec le poulet, 1 802 avec la souris contre 2 387 avec le lepisosté tacheté). Deux raisons principales peuvent expliquer

ce résultat. Tout d'abord, plus les génomes sont distants, moins ils partagent de gènes orthologues, soit parce que les gènes ont effectivement été perdus dans une lignée soit parce que leur divergence les rend plus difficilement identifiables. Cela induit une diminution du nombre d'arbres pris en compte pour les corrections, puisque SCORPiOs ne considère que les familles contenant un gène chez l'espèce utilisée comme groupe externe. De plus, la comparaison des patrons de synténie réalisée par SCORPiOs repose sur une étape dite de « threading », compliquée par la présence de nombreux réarrangements. Cette étape consiste à grouper un jeu de gènes dupliqués d'une paire d'espèces en paires de segments orthologues, selon un critère de parcimonie : on s'attend à ce que les segments orthologues soient plus similaires en terme d'évolution moléculaire des séquences, ainsi qu'en terme de pertes et rétentions de gènes. La conservation de la synténie permet de placer directement les gènes voisins dans un même segment, sous l'hypothèse de leur origine évolutive commune. En revanche, plus les génomes sont réarrangés, plus l'étape de « threading » de SCORPiOs est guidée par l'information d'évolution moléculaire donnée par les arbres initiaux, car les gènes dupliqués se retrouvent sur des segments non-contiguës. Le plus grand poids ainsi donné aux arbres initiaux durant le threading pourrait être responsable de leur sous-correction.

Ce résultat préliminaire suggère néanmoins que SCORPiOs pourrait potentiellement être appliqué à la 1R-2R, à condition de résoudre leur positionnement phylogénétique. Corriger les erreurs dans les arbres de gènes pourrait permettre de mieux caractériser l'évolution des gènes dupliqués après les 1R-2R et améliorer notre compréhension des conséquences génomiques et fonctionnelles de ces anciens événements de duplication.

6.2 Pertinence de l'information de synténie pour la reconstruction de phylogénies d'espèces

Mon travail de thèse s'est concentré sur l'utilisation des patrons de synténie dans le but d'améliorer les modèles d'arbres de gènes, dans le contexte spécifique des duplications complètes des poissons téléostéens. Dans le paragraphe précédent, j'ai discuté de l'élargissement de ce travail à l'échelle des Vertébrés. De manière plus générale que dans le cadre des duplications complètes de génome, la synténie est employée de manière croissante en phylogénomique, où elle permet de caractériser l'évolution des espèces, génomes et gènes, en complément des méthodes basées sur les séquences. Dans cette partie, je présente les stratégies basées sur la synténie dans le contexte de la résolution des phylogénies d'espèces.

En effet, la conservation des adjacences de gènes est une information utilisée pour reconstruire l'histoire évolutive des espèces. Ici, c'est le signal phylogénétique inscrit dans les réarrangements affectant les génomes au cours du temps qui est mis à profit afin de prédire

leurs relations évolutives. Dans une étude parallèle à mon travail de thèse (THOMPSON et al. 2020, preprint), en collaboration avec Ingo Braasch (Michigan State University), nous avons appliqué une méthode inspirée des stratégies décrites dans la littérature pour confirmer la monophylie du poisson-castor (*Amia calva*) et du lépisosté tacheté (*Lepisosteus oculatus*) dans le groupe des Holostei (Figure 6.2). Le clade Holostei est bien supporté par les phylogénies moléculaires mais contredit par les études basées sur des critères morphologiques, qui placent le poisson-castor en groupe frère des poissons téléostéens (SALLAN 2014). Nous avons reconstruit un arbre Neighbor-Joining à partir d'une matrice de distance entre génomes, sur la base de la proportion d'adjacences non-conservées entre paires d'espèces (Breakpoint Distance, BD). Un seul logiciel était alors disponible pour reconstruire des phylogénies d'espèces à partir de données d'adjacences de gènes : « MLGO » (LIN et al. 2013). Néanmoins, contrairement à notre approche, l'application de MLGO à nos données n'a pas permis de reconstruire correctement les nœuds bien résolus de la phylogénie des poissons. L'absence d'outils bien établis pour reconstruire les phylogénies d'espèces à partir d'information de synténie, ainsi que le succès de l'application de notre stratégie pour reconstruire la phylogénie des poissons, soulignent à la fois le besoin et la pertinence de nouveaux développements méthodologiques dans ce sens.

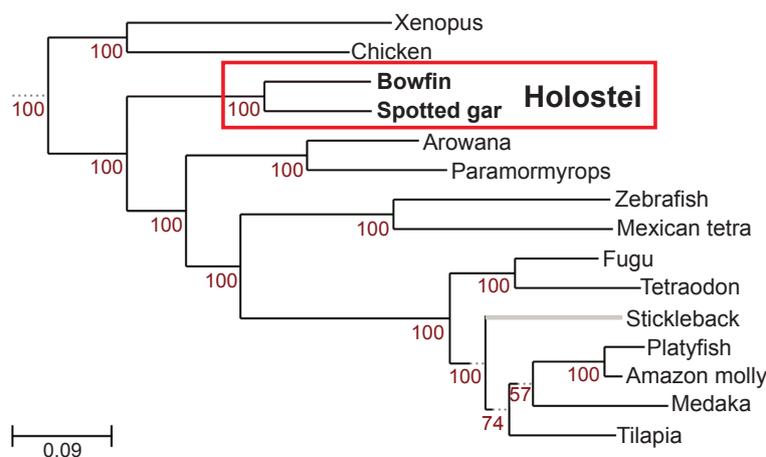


FIGURE 6.2 – Phylogénie des poissons reconstruite à partir de données de synténie. Arbre Neighbor-joining basé sur une mesure de distance reflétant la proportion de points de cassure entre paires de génomes. Les scores de bootstrap sont indiqués en rouge. Le clade des Holostei, regroupant le poisson-castor et le lépisosté tacheté est fortement soutenu, en accord avec les phylogénies moléculaires. A noter que la position relative de la branche de l'épinoche (Stickleback, en gris) est en désaccord avec les phylogénies consensus (BETANCUR-R et al. 2017 ; HUGHES et al. 2018), même si son placement est débattu (NEGRISOLO et al. 2010).

Récemment, Drillon et al. ont développé PhyChro, une méthode pour inférer une phylogénie à partir de données d'ordre de gènes (DRILLON et al. 2020). La métrique de distance de PhyChro est innovante par rapport à la BD, dans le sens où elle ne considère pas uniquement des comparaisons isolées entre paires d'espèces. PhyChro est basé sur la « Partial Split

Distance », qui répertorie les adjacences incompatibles entre paires d'espèces, c'est à dire celles qui permettent de les séparer parmi un jeu de génomes. PhyChro reconstruit correctement la phylogénie de 13 vertébrés et 21 levures et est largement supérieure à MLGO sur des simulations de génomes de vertébrés (3% des arbres MLGO corrects contre 79% pour PhyChro).

Preuve de l'intérêt pour l'intégration de la synténie dans la reconstruction de phylogénies, une seconde étude propose une nouvelle méthodologie pour répondre à ce problème (ZHAO et al. 2020, preprint). Les auteurs mettent en avant l'apport de la synténie pour la reconstruction de phylogénies chez les plantes, dont les familles de gènes ont des histoires évolutives complexes, caractérisées par un taux élevé de duplications. Un résultat intéressant de cette application est sa confrontation aux phylogénies moléculaires. Par exemple, un point de désaccord concerne la position relative de différents clades de Brassicacées : *Arabidopsis* (Clade A), *Boechera* (Clade B), et *Capsella* et *Camelina* (Clade C). Toutes les méthodes de phylogénie moléculaire regroupent le clade B avec le clade C, de la manière suivante : (A,(B,C)). La phylogénie de Zhao et al. groupe le clade A avec le clade C : (B,(A,C)). Ce résultat corrobore l'observation de Forsythe et al. qui suggère que les phylogénies moléculaires sont induites en erreur par la présence d'introgression massive entre les espèces des clades B et C (FORSYTHE, NELSON et BEILSTEIN 2020), alors que la « vraie » topologie est effectivement (B,(A,C)). L'introgression, par l'occurrence de croisements inter-spécifiques rares suivis de rétrocroisements successifs, aurait entraîné le remplacement d'une grande partie des gènes des espèces C par ceux des espèces B, réduisant de fait la divergence entre leurs séquences.

Un des avantages des méthodes basées sur la synténie est qu'elles sont computationnellement moins coûteuses que les méthodes de séquences. Pour cette raison, elles devraient devenir de plus en plus utilisées dans les prochaines années, aux vues de l'augmentation exponentielle du nombre de génomes de haute qualité disponibles. Comme discuté par Zhao et al., elles apportent un point de vue complémentaire à celui basé sur les séquences, parce qu'elles sont affectées par des mécanismes évolutifs différents. Ainsi, la synténie est informative pour mettre en lumière les événements biologiques mal considérés par les modèles actuels d'évolution des familles gènes, comme par exemple l'introgression. Je présente ces différents processus biologiques de manière plus détaillée dans la prochaine sous-partie.

6.3 Vers un enrichissement du modèle de l'arbre de gènes réconcilié

Nous avons vu, dans le paragraphe précédent, que l'information de synténie peut être utilisée afin de reconstruire des phylogénies d'espèces. Ici, je discute de l'intégration de

la synténie dans l'optique d'enrichir les modèles d'arbres de gènes réconciliés. La réconciliation d'un arbre de gènes consiste à expliquer, par des événements biologiques, les incongruences entre l'arbre de gènes et l'arbre des espèces. La plupart des méthodes de réconciliation considèrent les opérations de duplications et de pertes de gènes (modèle DL), ou duplications, pertes et transferts horizontaux (modèle DTL), pour expliquer ces désaccords (VILELLA et al. 2009 ; BOUSSAU et al. 2013 ; SZÖLLŐSI et al. 2013 ; HUERTA-CEPAS et al. 2014 ; SCORNAVACCA, JACOX et SZÖLLŐSI 2015 ; MOREL et al. 2020). La réconciliation suppose que la phylogénie des espèces est connue et strictement binaire. Cependant, les interactions génétiques entre espèces (introgression, hybridation) introduisent des réticulations dans les phylogénies et brisent cette structure binaire. L'introgression a été présentée au paragraphe précédent, elle désigne le transfert et remplacement de gènes entre espèces à travers l'intermédiaire d'hybrides. L'hybridation conduit à l'établissement d'une espèce hybride descendant du croisement de deux espèces parentales. Les génomes des hybrides ainsi produits sont typiquement constitués d'une mosaïque des gènes de chaque espèce parentale. Comme évoqué au paragraphe précédent, la confrontation des méthodes de synténie et de séquence peut permettre de mettre en lumière ces événements de réticulations dans les phylogénies d'espèces. Il s'agirait ensuite de les prendre en compte au moment de la réconciliation des arbres de gènes : l'introgression, par exemple, peut être modélisée comme un événement de transfert horizontal avec remplacement du gène « receveur ».

D'autres événements biologiques peuvent générer des discordances entre l'histoire évolutive d'un gène et celles des espèces. En particulier, le tri de lignées incomplet ou encore les conversions géniques, introduisent également un conflit entre l'histoire d'un gène en tant que séquence et en tant que locus. Dans le cas du tri de lignées incomplet, c'est la fixation indépendante, à la suite de spéciations rapides, des polymorphismes d'une population ancestrale dans les espèces descendantes qui introduit une discordance entre l'arbre des séquences et l'arbre des espèces. Parmi les méthodes capables de générer des arbres réconciliés pour des jeux de données de plusieurs dizaines d'espèces, seul OrthoFinder modélise le tri incomplet des lignées, à travers une adaptation moins computationnellement intensive du modèle « DLCpar » (EMMS et KELLY 2019, voir chapitre 2, paragraphe 2.1.2).

Enfin, les conversions géniques correspondent au remplacement de la séquence d'un gène par une séquence paralogue, par recombinaison méiotique. Les conversions géniques, lorsque non prises en compte lors de la réconciliation, peuvent mener à une sur-estimation du nombre de duplications dans les arbres de gènes. De plus, négliger ces événements peut avoir un impact significatif sur les études de datations : puisqu'elles retardent la divergence de séquences paralogues, elles entraînent une sous-estimation de la date de duplication (la duplication sera prédite plus récente qu'elle ne l'est). Pourtant, les événements de conversion génique sont actuellement complètement absents de toutes les implémentations de réconciliation d'arbre de gènes. La raison principale est qu'elle brise les principes sur les-

quels ces méthodes sont bâties, en autorisant des interactions entre gènes d’une même espèce. La reconnaissance croissante de leur importance biologique a récemment suscité une première exploration des possibilités algorithmiques pour l’inclusion d’événements de conversion génique dans une réconciliation avec duplications, pertes et conversions (HASIĆ et TANNIER 2019). Puisque l’événement de conversion génique correspond au remplacement d’une séquence paralogue « receveuse » par sa séquence paralogue « donneuse », les auteurs proposent de la modéliser par un événement simultané de duplication (de la séquence donneuse) et de perte (séquence receveuse). Ils décrivent un algorithme polynomial permettant d’obtenir les solutions de maximum de parcimonie d’une réconciliation DLConversion, dans le cas où tous les événements ont un coût égal. Cependant, dans le cas de topologies LORe dans le contexte de duplications complètes, une telle approche mènerait à une réconciliation erronée, identique à la solution des méthodes DL évoquées dans le chapitre 4 (Figure 6.3). Dans ce contexte, considérer l’information de synténie et la connaissance *a priori* du mécanisme de duplication apparaît à nouveau comme une information précieuse pour guider l’inférence d’événements de conversion génique.

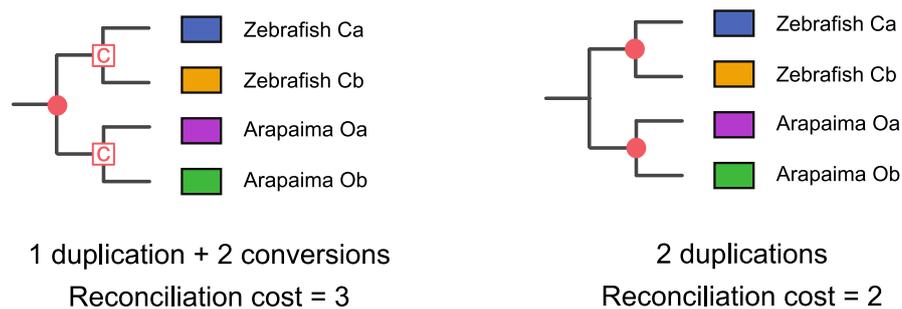


FIGURE 6.3 – Réconciliation de topologie d’arbre de type LORe. A gauche, l’arbre LORe est réconcilié correctement : la duplication est positionné au niveau de la 3R et les séquences des gènes ohnologues ont divergé plus tardivement, à la suite d’événements de conversion génique. A droite, le nœud à la racine est un événement de spéciation et les gènes dupliqués sont issus de duplications lignée-spécifiques. A travers une réconciliation par parcimonie allouant le même coût à chaque événement, le scénario de réconciliation correct a un coût plus élevé.

Dans le paragraphe suivant, je développe une autre difficulté rencontrée lors de la réconciliation, liée aux incertitudes de la topologie de l’arbre de gènes. En effet, la réconciliation repose sur la connaissance des topologies de l’arbre des espèces et de gènes. L’augmentation rapide de la taille des jeux de données considérés ainsi que la quantité insuffisante de signal phylogénétique dans les séquences de gènes sont les principales causes d’erreurs dans les arbres de gènes. Minimiser ces erreurs est indispensable à une réconciliation correcte : la réconciliation d’une topologie d’arbre de gènes erronée mènera nécessairement à un scénario de réconciliation non-réaliste.

6.4 Méthodologies de la génomique comparative et quantité de données

6.4.1 Évolution du processus d'inférence d'arbres de gènes

Au delà de ce besoin d'enrichir les modèles d'arbre de gènes, apparaît également, de part la quantité croissante de données, la nécessité de réduire le coût computationnel de leur inférence tout en minimisant les erreurs de prédictions. En effet, la quantité d'espèces corrèle avec le nombre d'erreurs introduites à différentes étapes du pipeline de reconstruction d'arbres, que ce soit au niveau de l'inférence des alignements multiples (DEOROWICZ, DEBUDAJ-GRABYSZ et GUDYŚ 2016) ou celle des arbres en eux-même (WU et al. 2013 ; PAREY et al. 2020). Pour pallier ce problème d'augmentation du taux d'erreurs, la base de données d'Ensembl (YATES et al. 2020) ne reconstruit depuis sa version 94 que des sous-arbres pour les plus grandes familles de gènes. L'information de l'appartenance des sous-arbres à une famille est disponible, mais le branchement des sous-arbres entre eux n'est pas calculé. En conséquence, les paralogues de nombreux gènes ne sont plus directement liés entre eux. De fait, les méthodes actuelles montrent leur limite pour reconstruire des arbres de gènes contenant ~200 espèces. Dans ce contexte, il est envisageable que les stratégies de reconstruction d'arbres soient réorientées vers le calcul d'un jeu d'arbres initiaux « approximatifs » prédits par des méthodes rapides et corrigés par la suite à travers différentes stratégies.

De plus, il apparaîtrait désirable de ne pas recalculer tout le processus de reconstruction d'arbres à chaque addition d'un nouveau génome, et plutôt de développer des méthodes capables d'insérer les gènes de nouvelles espèces dans les arbres, quitte à, à nouveau, améliorer ces arbres *a posteriori*. C'est la direction prise par les méthodes d'inférence d'alignement de génomes complets : l'aligneur « ProgressiveCactus » permet de rajouter des génomes à un alignement complet pré-calculé (ARMSTRONG et al. 2019a), et devient la méthode de référence pour construire des alignements multiples de génomes avec plus de 600 espèces.

Je propose, dans la prochaine partie, une piste à laquelle nous réfléchissons afin d'étendre la philosophie de SCORPiOs à une correction plus générale des arbres de gènes et qui pourrait s'insérer dans ce contexte d'adaptation des procédures de reconstruction d'arbres.

6.4.2 Extension de SCORPiOs à la correction de duplications dans le cas général

Le positionnement correct des nœuds de duplication dans les arbres de gènes est un problème général, qui ne se restreint pas à celui des duplications complètes de génome. Notamment, le génome humain est composé d'au moins 40% de gènes dupliqués et la grande majorité des gènes de maladies humaines sont des gènes possédant au moins un paralogue

(ZHANG 2003 ; MAKINO et MCLYSAGHT 2010 ; DICKERSON et ROBERTSON 2012 ; SACERDOT et al. 2018). En plus de la prise en compte insuffisante d'autres événements biologiques lors de la réconciliation à l'arbre d'espèces (voir la partie 6.3), les incertitudes dans les topologies d'arbre de gènes en elles-mêmes sont une source majeure d'erreurs de positionnement des duplications. Dans cette optique, il est intéressant de réfléchir à une extension de la philosophie de SCORPiOs à une correction plus générale, qui pourrait servir de support pour améliorer un jeu d'arbres initiaux.

SCORPiOs se base sur les patrons de synténie attendus après duplication complète pour identifier, dans les arbres de gènes, les duplications mal placées. Dans le cas général, nous n'avons pas *a priori* sur la position des duplications, ni sur les patrons de synténie attendus, ce qui rend le problème plus difficile. En revanche, les méthodes de reconstruction ancestrale de génome contiennent potentiellement l'information des nœuds incohérents avec la synténie. En effet, certaines de ces méthodes, comme par exemple AGORA (MUFFATO 2010 ; BERTHELOT et al. 2015), reconstruisent les génomes ancestraux à partir de graphes enregistrant les adjacences de gènes observées dans les génomes modernes. La résolution des conflits à travers la linéarisation du graphe permet ensuite de prédire l'ordre des gènes dans leur ancêtre. Ces incohérences, sous forme de bifurcations ou cycles dans les graphes, peuvent être la conséquence d'erreurs présentes dans les arbres de gènes (Figure 6.4).

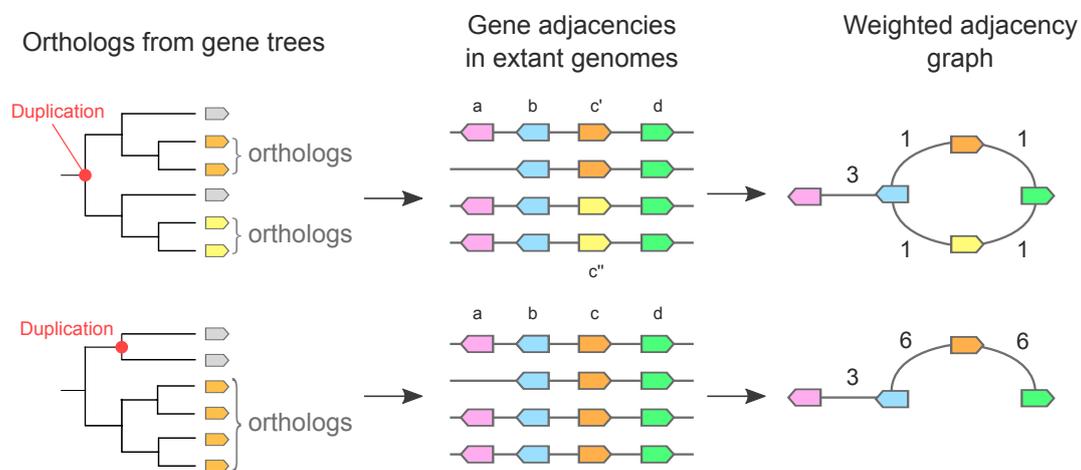


FIGURE 6.4 – Identification d'erreurs dans les arbres de gènes à travers les graphes d'adjacences conservées. Représentation simplifiée du fonctionnement d'AGORA : pour un ancêtre donné, les gènes orthologues des espèces descendantes sont collectés à partir des arbres de gènes (les gènes grisés représentent les gènes d'une espèce qui ne descend pas de l'ancêtre considéré). Les adjacences conservées dans au moins une paire d'espèces sont stockées dans le graphe pondéré, où les poids correspondent au nombre de paires d'espèces supportant l'adjacence. Par exemple, en haut, l'adjacence entre les gènes bleu et orange n'est conservée qu'entre une seule paire d'espèces. Dans l'exemple du haut, le graphe contient un cycle. Dans l'exemple du bas, la correction de la position du nœud de duplication dans l'arbre de la famille du gène c permet d'obtenir un graphe linéaire.

Cette observation a également été faite par les auteurs de la méthode de reconstruction ancestrale DeCoStar (DUCHEMIN et al. 2017), qui proposent qu'un score de non-linéarité pourrait effectivement permettre d'identifier les arbres de gènes contenant des erreurs (TANNIER et al. 2020). On peut envisager la correction des nœuds erronés ainsi identifiés selon le principe de SCORPiOs, en calculant une solution d'arbre alternative qui résoudrait le conflit identifié dans le graphe (Figure 6.4). Cette stratégie reste encore largement à explorer afin de mieux appréhender les difficultés de sa mise en place, mais elle représente une piste prometteuse pour générer de meilleurs arbres de gènes, en un temps réduit. En effet, AGORA est très efficace pour calculer les graphes d'adjacences, qui représentent une information de synténie très complète : ils répertorient l'ensemble des informations d'adjacences des espèces modernes, à tous les nœuds ancestraux d'une phylogénie.

6.5 Conclusion

La sophistication des méthodes de reconstruction d'arbres de gènes laisse entrevoir la possibilité prochaine d'enrichir les phylogénies de gènes en intégrant de manière plus complète les différents événements qui expliquent leur histoire évolutive, à la fois de leur locus et de leur séquence, avec toujours plus d'espèces. Notamment, un cadre de réconciliation des arbres de gènes aux arbres d'espèces considérant simultanément les pertes, duplications, transferts, tri incomplet des lignées et conversions géniques reste à établir. Pour résoudre le problème de signal phylogénétique insuffisant des séquences, d'autres informations devront être mises à profit : synténie, coût de réconciliation ou encore utilisation des séquences introniques (SCORNAVACCA et GALTIER 2017). L'utilisation d'information complémentaire pour répondre aux questions de la phylogénomique permettrait également de mieux qualifier l'impact de la circularité des inférences actuelles : les séquences des gènes permettent d'établir les phylogénies d'espèces, phylogénies d'espèces utilisées à leur tour pour la réconciliation d'arbres de gènes, eux-même inférés à travers l'information de séquence. L'exemple des Brassicacées présenté au paragraphe 6.2 montre que cette circularité peut effectivement masquer des événements biologiques significatifs.

Enfin, satisfaire les promesses de l'ère post-génomique passe non seulement à travers la reconstruction de l'histoire des gènes mais aussi celles des réarrangements génomiques et des éléments régulateurs, et la mise en évidence des forces qui gouvernent leur évolution. L'explosion de nouvelles données représente à la fois de nouveaux défis et de nouvelles opportunités pour explorer ces questions et décrire le lien génotype-phénotype avec plus de précision et à plus grande échelle.

Bibliographie

- ACHARYA, Debarun et Tapash C. GHOSH (2016). “Global analysis of human duplicated genes reveals the relative importance of whole-genome duplicates originated in the early vertebrate evolution”. In : *BMC Genomics* 17.1, p. 71.
- ADAMS, M. D. et al. (1995). “Initial assessment of human gene diversity and expression patterns based upon 83 million nucleotides of cDNA sequence”. In : *Nature* 377.6547 Suppl, p. 3-174.
- AKEN, Bronwen L. et al. (2016). “The Ensembl gene annotation system”. In : *Database: The Journal of Biological Databases and Curation* 2016.
- ALBALAT, Ricard et Cristian CAÑESTRO (2016). “Evolution by gene loss”. In : *Nature Reviews Genetics* 17.7, p. 379-391.
- ALBERTSON, R. Craig et al. (2009). “Evolutionary Mutant Models for Human Disease”. In : *Trends in genetics : TIG* 25.2, p. 74-81.
- ALLENDORF, Fred W. et al. (2015). “Effects of Crossovers Between Homeologs on Inheritance and Population Genomics in Polyploid-Derived Salmonid Fishes”. In : *Journal of Heredity* 106.3, p. 217-227.
- ALTENHOFF, Adrian M. et Christophe DESSIMOZ (2012). “Inferring Orthology and Paralogy”. In : *Evolutionary Genomics*. Sous la dir. de Maria ANISIMOVA. T. 855. Totowa, NJ : Humana Press, p. 259-279.
- ALTENHOFF, Adrian M. et al. (2012). “Resolving the Ortholog Conjecture: Orthologs Tend to Be Weakly, but Significantly, More Similar in Function than Paralogs”. In : *PLoS Computational Biology* 8.5. Sous la dir. de Jonathan A. EISEN, e1002514.
- ALTENHOFF, Adrian M. et al. (2016). “Standardized benchmarking in the quest for orthologs”. In : *Nature Methods* 13.5, p. 425-430.
- ALTENHOFF, Adrian M. et al. (2018). “The OMA orthology database in 2018: retrieving evolutionary relationships among all domains of life through richer web and programmatic interfaces”. In : *Nucleic Acids Research* 46.D1, p. D477-D485.
- AMORES, Angel et al. (1998). “Zebrafish hox Clusters and Vertebrate Genome Evolution”. In : *Science* 282.5394, p. 1711-1714.
- APARICIO, Samuel et al. (2002). “Whole-genome shotgun assembly and analysis of the genome of *Fugu rubripes*”. In : *Science (New York, N.Y.)* 297.5585, p. 1301-1310.
- ARCILA, Dahiana et al. (2017). “Genome-wide interrogation advances resolution of recalcitrant groups in the tree of life”. In : *Nature Ecology & Evolution* 1.2, p. 0020.

- ARMSTRONG, Joel et al. (2019a). “Progressive alignment with Cactus: a multiple-genome aligner for the thousand-genome era”. In : *bioRxiv*, p. 730531.
- ARMSTRONG, Joel et al. (2019b). “Whole-Genome Alignment and Comparative Annotation”. In : *Annual review of animal biosciences* 7, p. 41-64.
- ARRATIA, Gloria (1998). “Basal Teleosts and Teleostean Phylogeny: Response to C. Patterson”. In : *Copeia* 1998.4, p. 1109.
- BEÇAK, Maria Luiza, Willy BEÇAK et Maria Nazareth RABELLO (1966). “Cytological evidence of constant tetraploidy in the bisexual South American frog *Odontophrynus americanus*”. In : *Chromosoma* 19.2, p. 188-193.
- BERTHELOT, Camille et al. (2014). “The rainbow trout genome provides novel insights into evolution after whole-genome duplication in vertebrates”. In : *Nature Communications* 5.1, p. 3657.
- BERTHELOT, Camille et al. (2015). “The 3D organization of chromatin explains evolutionary fragile genomic regions”. In : *Cell Reports* 10.11, p. 1913-1924.
- BETANCUR-R, Ricardo et al. (2017). “Phylogenetic classification of bony fishes”. In : *BMC Evolutionary Biology* 17.1, p. 162.
- BIAN, Chao et al. (2016). “The Asian arowana (*Scleropages formosus*) genome provides new insights into the evolution of an early lineage of teleosts”. In : *Scientific Reports* 6.1, p. 24501.
- BIRD, Kevin A. et al. (2018). “The causes and consequences of subgenome dominance in hybrids and recent polyploids”. In : *New Phytologist* 220.1, p. 87-93.
- BLEI, David M. (2012). “Probabilistic topic models”. In : *Communications of the ACM* 55.4, p. 77-84.
- BOFFELLI, Dario, Marcelo A. NOBREGA et Edward M. RUBIN (2004). “Comparative genomics at the vertebrate extremes”. In : *Nature Reviews Genetics* 5.6, p. 456-465.
- BOMBLIES, Kirsten et al. (2016). “The challenge of evolving stable polyploidy: could an increase in “crossover interference distance” play a central role?” In : *Chromosoma* 125, p. 287-300.
- BOUSSAU, Bastien et al. (2013). “Genome-scale coestimation of species and gene trees”. In : *Genome Research* 23.2, p. 323-330.
- BRAASCH, Ingo et John H. POSTLETHWAIT (2012). “Polyploidy in Fish and the Teleost Genome Duplication”. In : *Polyploidy and Genome Evolution*. Sous la dir. de Pamela S. SOLTIS et Douglas E. SOLTIS. Berlin, Heidelberg : Springer, p. 341-383.
- BRAASCH, Ingo et al. (2016). “The spotted gar genome illuminates vertebrate evolution and facilitates human-teleost comparisons”. In : *Nature Genetics* 48.4, p. 427-437.
- BRADFORD, Yvonne M. et al. (2017). “Zebrafish Models of Human Disease: Gaining Insight into Human Disease at ZFIN”. In : *ILAR Journal* 58.1, p. 4-16.
- BRUNET, Frédéric G. et al. (2006). “Gene Loss and Evolutionary Rates Following Whole-Genome Duplication in Teleost Fishes”. In : *Molecular Biology and Evolution* 23.9, p. 1808-1816.

- BUGGS, Richard J. A. et al. (2012). “Rapid, Repeated, and Clustered Loss of Duplicate Genes in Allopolyploid Plant Populations of Independent Origin”. In : *Current Biology* 22.3, p. 248-252.
- BYRNE, Kevin P. et Kenneth H. WOLFE (2005). “The Yeast Gene Order Browser: Combining curated homology and syntenic context reveals gene fate in polyploid species”. In : *Genome Research* 15.10, p. 1456-1461.
- CATCHEN, Julian M., John S. CONERY et John H. POSTLETHWAIT (2009). “Automated identification of conserved synteny after whole-genome duplication”. In : *Genome Research* 19.8, p. 1497-1505.
- CELEUX, Gilles et Jean-Baptiste DURAND (2008). “Selecting hidden Markov model state number with cross-validated likelihood”. In : *Computational Statistics* 23.4, p. 541-564.
- CHALOPIN, Domitille et al. (2015). “Comparative Analysis of Transposable Elements Highlights Mobilome Diversity and Evolution in Vertebrates”. In : *Genome Biology and Evolution* 7.2, p. 567-580.
- CHEN, Feng et al. (2006). “OrthoMCL-DB: querying a comprehensive multi-species collection of ortholog groups”. In : *Nucleic Acids Research* 34.suppl_1, p. D363-D368.
- CHEN, K., D. DURAND et M. FARACH-COLTON (2000). “NOTUNG: a program for dating gene duplications and optimizing gene family trees”. In : *Journal of Computational Biology: A Journal of Computational Molecular Cell Biology* 7.3-4, p. 429-447.
- CHEN, Wei-Hua et al. (2017). “OGEE v2: an update of the online gene essentiality database with special focus on differentially essential genes in human cancer cell lines”. In : *Nucleic Acids Research* 45.D1, p. D940-D944.
- CHEN, Zelin et al. (2019). “De novo assembly of the goldfish (*Carassius auratus*) genome and the evolution of genes after whole-genome duplication”. In : *Science Advances* 5.6, eaav0547.
- CHESTER, Michael et al. (2012). “Extensive chromosomal variation in a recently formed natural allopolyploid species, *Tragopogon miscellus* (Asteraceae)”. In : *Proceedings of the National Academy of Sciences* 109.4, p. 1176-1181.
- CLÉMENT, Yves et al. (2020). “Enhancer–gene maps in the human and zebrafish genomes using evolutionary linkage conservation”. In : *Nucleic Acids Research* 48.5, p. 2357-2371.
- COMPLEX TRAIT CONSORTIUM, Members of the (2003). “The nature and identification of quantitative trait loci”. In : *Nature reviews. Genetics* 4.11, p. 911-916.
- COMTE, Nicolas et al. (2020). “Treerecs: an integrated phylogenetic tool, from sequences to reconciliations”. In : *Bioinformatics* ().
- CONANT, Gavin C. (2020). “The lasting after-effects of an ancient polyploidy on the genomes of teleosts”. In : *PLOS ONE* 15.4, e0231356.
- CONANT, Gavin C. et Kenneth H. WOLFE (2008). “Probabilistic Cross-Species Inference of Orthologous Genomic Regions Created by Whole-Genome Duplication in Yeast”. In : *Genetics* 179.3, p. 1681-1692.

- CORTESI, Fabio et al. (2015). “Ancestral duplications and highly dynamic opsin gene evolution in percomorph fishes”. In : *Proceedings of the National Academy of Sciences* 112.5, p. 1493-1498.
- CURRIE, Peter D. (1996). “Zebrafish genetics: Mutant cornucopia”. In : *Current Biology* 6.12, p. 1548-1552.
- DAVÍN, Adrián A. et al. (2020). “Zombi: a phylogenetic simulator of trees, genomes and sequences that accounts for dead lineages”. In : *Bioinformatics* 36.4, p. 1286-1288.
- DE SMET, Riet et Yves VAN DE PEER (2012). “Redundancy and rewiring of genetic networks following genome-wide duplication events”. In : *Current Opinion in Plant Biology. Genome studies molecular genetics* 15.2, p. 168-176.
- DE SMET, Riet et al. (2013). “Convergent gene loss following gene and genome duplications creates single-copy families in flowering plants”. In : *Proceedings of the National Academy of Sciences of the United States of America* 110.8, p. 2898-2903.
- DEHAL, Paramvir et Jeffrey L BOORE (2005). “Two Rounds of Whole Genome Duplication in the Ancestral Vertebrate”. In : *PLoS Biology* 3.10. Sous la dir. de Peter HOLLAND, e314.
- DELUCA, Todd F. et al. (2012). “Roundup 2.0: enabling comparative genomics for over 1800 genomes”. In : *Bioinformatics* 28.5, p. 715-716.
- DEOROWICZ, Sebastian, Agnieszka DEBUDAJ-GRABYSZ et Adam GUDYŚ (2016). “FAMSA: Fast and accurate multiple sequence alignment of huge protein families”. In : *Scientific Reports* 6.1, p. 33964.
- DICKERSON, Jonathan E. et David L. ROBERTSON (2012). “On the Origins of Mendelian Disease Genes in Man: The Impact of Gene Duplication”. In : *Molecular Biology and Evolution* 29.1, p. 61-69.
- DOONER, Hugo K. (2002). “Extensive Interallelic Polymorphisms Drive Meiotic Recombination into a Crossover Pathway”. In : *The Plant Cell* 14.5, p. 1173-1183.
- DRIEVER, W. et al. (1996). “A genetic screen for mutations affecting embryogenesis in zebrafish”. In : *Development* 123.1, p. 37-46.
- DRILLON, Guénola et al. (2020). “Phylogenetic Reconstruction Based on Synteny Block and Gene Adjacencies”. In : *Molecular Biology and Evolution* 37.9, p. 2747-2762.
- DU, Kang et al. (2019). “The genome of the arapaima (*Arapaima gigas*) provides insights into gigantism, fast growth and chromosomal sex determination system”. In : *Scientific Reports* 9.1, p. 5293.
- DU, Kang et al. (2020). “The sterlet sturgeon genome sequence and the mechanisms of segmental rediploidization”. In : *Nature Ecology & Evolution* 4.6, p. 841-852.
- DUCHEMIN, Wandrille et al. (2017). “DeCoSTAR: Reconstructing the Ancestral Organization of Genes or Genomes Using Reconciled Phylogenies”. In : *Genome Biology and Evolution* 9.5, p. 1312-1319.
- EDGER, Patrick P. et al. (2017). “Subgenome Dominance in an Interspecific Hybrid, Synthetic Allopolyploid, and a 140-Year-Old Naturally Established Neo-Allopolyploid Monkeyflower”. In : *The Plant Cell* 29.9, p. 2150-2167.

- EFRON, Bradley (1979). “Computers and the Theory of Statistics: Thinking the Unthinkable”. In : *SIAM Review* 21.4, p. 460-480.
- EMMS, David M. et Steven KELLY (2015). “OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy”. In : *Genome Biology* 16.1, p. 157.
- (2019). “OrthoFinder: phylogenetic orthology inference for comparative genomics”. In : *Genome Biology* 20.1, p. 238.
- ERNST, Jason et Manolis KELLIS (2017). “Chromatin-state discovery and genome annotation with ChromHMM”. In : *Nature Protocols* 12.12, p. 2478-2492.
- EVANS, Ben J. et al. (2017). “Evolution of the Largest Mammalian Genome”. In : *Genome Biology and Evolution* 9.6, p. 1711-1724.
- FAN, Guangyi et al. (2020). “Initial data release and announcement of the 10,000 Fish Genomes Project (Fish10K)”. In : *GigaScience* 9.8.
- FARRIS, James S. (1977). “Phylogenetic Analysis Under Dollo’s Law”. In : *Systematic Biology* 26.1, p. 77-88.
- FELSENSTEIN, Joseph (1985). “Confidence Limits on Phylogenies: An Approach Using the Bootstrap”. In : *Evolution* 39.4, p. 783-791.
- FENG, Bing, Lingxi ZHOU et Jijun TANG (2017). “Ancestral Genome Reconstruction on Whole Genome Level”. In : *Current Genomics* 18.4, p. 306-315.
- FENG, Shaohong et al. (2020). “Dense sampling of bird diversity increases power of comparative genomics”. In : *Nature* 587.7833, p. 252-257.
- FESCHOTTE, Cédric et Ellen J. PRITHAM (2007). “DNA Transposons and the Evolution of Eukaryotic Genomes”. In : *Annual Review of Genetics* 41.1, p. 331-368.
- FIELDS, Chris et al. (1994). “How many genes in the human genome?” In : *Nature Genetics* 7.3, p. 345-346.
- FLEISCHMANN, R. D. et al. (1995). “Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd”. In : *Science* 269.5223, p. 496-512.
- FORCE, A et al. (1999). “Preservation of duplicate genes by complementary, degenerative mutations.” In : *Genetics* 151.4, p. 1531-1545.
- FORSYTHE, Evan S., Andrew D. L. NELSON et Mark A. BEILSTEIN (2020). “Biased Gene Retention in the Face of Introgression Obscures Species Relationships”. In : *Genome Biology and Evolution* 12.9, p. 1646-1663.
- FREELING, Michael et al. (2012). “Fractionation mutagenesis and similar consequences of mechanisms removing dispensable or less-expressed DNA in plants”. In : *Current Opinion in Plant Biology*. Genome studies molecular genetics 15.2, p. 131-139.
- FURLONG, Rebecca F. et Peter W. H. HOLLAND (2002). “Were vertebrates octoploid?” In : *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences* 357.1420, p. 531-544.

- GAGNON, Yves, Mathieu BLANCHETTE et Nadia EL-MABROUK (2012). “A flexible ancestral genome reconstruction method based on gapped adjacencies”. In : *BMC Bioinformatics* 13.19, S4.
- GALLARDO, Milton H. et al. (2004). “Whole-genome duplications in South American desert rodents (Octodontidae)”. In : *Biological Journal of the Linnean Society* 82.4, p. 443-451.
- GARSMEUR, Olivier et al. (2014). “Two Evolutionarily Distinct Classes of Paleopolyploidy”. In : *Molecular Biology and Evolution* 31.2, p. 448-454.
- GAUT, Brandon S. et John F. DOEBLEY (1997). “DNA sequence evidence for the segmental allotetraploid origin of maize”. In : *Proceedings of the National Academy of Sciences* 94.13, p. 6809-6814.
- GENEREUX, Diane P. et al. (2020). “A comparative genomics multitool for scientific discovery and conservation”. In : *Nature* 587.7833, p. 240-245.
- GHARBI, Karim et al. (2006). “A Linkage Map for Brown Trout (*Salmo trutta*): Chromosome Homeologies and Comparative Genome Organization With Other Salmonid Fish”. In : *Genetics* 172.4, p. 2405-2419.
- GIBSON, Toby J et Jürg SPRING (1998). “Genetic redundancy in vertebrates: polyploidy and persistence of genes encoding multidomain proteins”. In : *Trends in Genetics* 14.2, p. 46-49.
- GILLARD, Gareth B. et al. (2020). “Comparative regulomics reveals pervasive selection on gene dosage following whole genome duplication”. In : *bioRxiv*, p. 2020.07.20.212316.
- GLASAUER, Stella M. K. et Stephan C. F. NEUHAUSS (2014). “Whole-genome duplication in teleost fishes and its evolutionary consequences”. In : *Molecular Genetics and Genomics* 289.6, p. 1045-1060.
- GOFFEAU, A. et al. (1996). “Life with 6000 Genes”. In : *Science* 274.5287, p. 546-567.
- GOLDMAN, N., J. P. ANDERSON et A. G. RODRIGO (2000). “Likelihood-based tests of topologies in phylogenetics”. In : *Systematic Biology* 49.4, p. 652-670.
- GORI, Kevin et al. (2016). “Clustering Genes of Common Evolutionary History”. In : *Molecular Biology and Evolution* 33.6, p. 1590-1605.
- GOSLINE, William A. (1965). “Teleostean Phylogeny”. In : *Copeia* 1965.2, p. 186.
- GOUT, Jean-François et al. (2010). “The relationship among gene expression, the evolution of gene dosage, and the rate of protein evolution”. In : *PLoS genetics* 6.5, e1000944.
- GUINDON, Stéphane et Olivier GASCUEL (2003). “A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood”. In : *Systematic Biology* 52.5, p. 696-704.
- GUINDON, Stéphane et al. (2010). “New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0”. In : *Systematic Biology* 59.3, p. 307-321.
- GUSCHANSKI, Katerina, Maria WARNEFORS et Henrik KAESSMANN (2017). “The evolution of duplicate gene expression in mammalian organs”. In : *Genome Research*, gr.215566.116.

- HAFFTER, Pascal et al. (1996). “The identification of genes with unique and essential functions in the development of the zebrafish, *Danio rerio*”. In : *Development* 123.1, p. 1-36.
- HAHN, Matthew W. (2007). “Bias in phylogenetic tree reconciliation methods: implications for vertebrate genome evolution”. In : *Genome Biology* 8.7, R141.
- HASIĆ, Damir et Eric TANNIER (2019). “Gene tree species tree reconciliation with gene conversion”. In : *Journal of Mathematical Biology* 78.6, p. 1981-2014.
- HAVELKA, M. et al. (2013). “Extensive genome duplications in sturgeons: new evidence from microsatellite data”. In : *Journal of Applied Ichthyology* 29.4, p. 704-708.
- HOFFMAN, Michael M. et al. (2012). “Unsupervised pattern discovery in human chromatin structure through genomic segmentation”. In : *Nature Methods* 9.5, p. 473-476.
- HOLLAND, Linda Z. (2015). “The origin and evolution of chordate nervous systems”. In : *Philosophical Transactions of the Royal Society B: Biological Sciences* 370.1684.
- HOWE, Kerstin et al. (2013). “The zebrafish reference genome sequence and its relationship to the human genome”. In : *Nature* 496.7446, p. 498-503.
- HUELSMANN, Matthias et al. (2019). “Genes lost during the transition from land to water in cetaceans highlight genomic changes associated with aquatic adaptations”. In : *Science Advances* 5.9, eaaw6671.
- HUERTA-CEPAS, Jaime et al. (2014). “PhylomeDB v4: zooming into the plurality of evolutionary histories of a genome”. In : *Nucleic Acids Research* 42.Database issue, p. D897-D902.
- HUFTON, Andrew L et Georgia PANOPOULOU (2009). “Polyploidy and genome restructuring: a variety of outcomes”. In : *Current Opinion in Genetics & Development*. Genomes and evolution 19.6, p. 600-606.
- HUGHES, Lily C. et al. (2018). “Comprehensive phylogeny of ray-finned fishes (Actinopterygii) based on transcriptomic and genomic data”. In : *Proceedings of the National Academy of Sciences* 115.24, p. 6249-6254.
- HUSBAND, Brian C. (2004). “The role of triploid hybrids in the evolutionary dynamics of mixed-ploidy populations: TRIPLOIDS IN MIXED-PLOIDY POPULATIONS”. In : *Biological Journal of the Linnean Society* 82.4, p. 537-546.
- INOUE, Jun et al. (2015). “Rapid genome reshaping by multiple-gene loss after whole-genome duplication in teleost fish suggested by mathematical modeling”. In : *Proceedings of the National Academy of Sciences of the United States of America* 112.48, p. 14918-14923.
- JAILLON, Olivier, Jean-Marc AURY et Patrick WINCKER (2009). ““Changing by doubling”, the impact of Whole Genome Duplications in the evolution of eukaryotes”. In : *Comptes Rendus Biologies*. La théorie de Darwin revisitée par la biologie d’aujourd’hui / Darwin’s theory revisited by today’s biology 332.2, p. 241-253.
- JAILLON, Olivier et al. (2004). “Genome duplication in the teleost fish *Tetraodon nigroviridis* reveals the early vertebrate proto-karyotype”. In : *Nature* 431.7011, p. 946-957.
- JAŻWIŃSKA, Anna et Simon BLANCHOU (2020). “Towards deciphering variations of heart regeneration in fish”. In : *Current Opinion in Physiology*. Regeneration 14, p. 21-26.

- JONES, Bradley R. et al. (2012a). “ANGES: reconstructing ANcestral GENomeS maps”. In : *Bioinformatics* 28.18, p. 2388-2390.
- JONES, Felicity C. et al. (2012b). “The genomic basis of adaptive evolution in threespine sticklebacks”. In : *Nature* 484.7392, p. 55-61.
- KADUK, Mateusz et al. (2017). “HieranoiDB: a database of orthologs inferred by Hieranoid”. In : *Nucleic Acids Research* 45.D1, p. D687-D690.
- KAPUSTA, Aurélie, Alexander SUH et Cédric FESCHOTTE (2017). “Dynamics of genome size evolution in birds and mammals”. In : *Proceedings of the National Academy of Sciences of the United States of America* 114.8, E1460-E1469.
- KARLSSON, Elinor K. et Kerstin LINDBLAD-TOH (2008). “Leader of the pack: gene mapping in dogs and other model organisms”. In : *Nature Reviews Genetics* 9.9, p. 713-725.
- KASAHARA, Masahiro et al. (2007). “The medaka draft genome and insights into vertebrate genome evolution”. In : *Nature* 447.7145, p. 714-719.
- KASSAHN, Karin S. et al. (2009). “Evolution of gene function and regulatory control after whole-genome duplication: Comparative analyses in vertebrates”. In : *Genome Research* 19.8, p. 1404-1418.
- KAUFMAN, Jim (2018). “Unfinished Business: Evolution of the MHC and the Adaptive Immune System of Jawed Vertebrates”. In : *Annual Review of Immunology* 36.1, p. 383-409.
- KELLEY, Joanna L. et al. (2016). “Mechanisms Underlying Adaptation to Life in Hydrogen Sulfide-Rich Environments”. In : *Molecular Biology and Evolution* 33.6, p. 1419-1434.
- KELLIS, M. et al. (2014). “Defining functional DNA elements in the human genome”. In : *Proceedings of the National Academy of Sciences* 111.17, p. 6131-6138.
- KELLIS, Manolis, Bruce W. BIRREN et Eric S. LANDER (2004). “Proof and evolutionary analysis of ancient genome duplication in the yeast *Saccharomyces cerevisiae*”. In : *Nature* 428.6983, p. 617-624.
- KENNY, N. J. et al. (2016). “Ancestral whole-genome duplication in the marine chelicerate horseshoe crabs”. In : *Heredity* 116.2, p. 190-199.
- KIEFER, Christiane et al. (2019). “Interspecies association mapping links reduced CG to TG substitution rates to the loss of gene-body methylation”. In : *Nature Plants* 5.8, p. 846-855.
- KIM, Jaebum et al. (2017). “Reconstruction and evolutionary history of eutherian chromosomes”. In : *Proceedings of the National Academy of Sciences* 114.27, E5379-E5388.
- KING, M. C. et A. C. WILSON (1975). “Evolution at two levels in humans and chimpanzees”. In : *Science* 188.4184, p. 107-116.
- KISHINO, H. et M. HASEGAWA (1989). “Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in hominoidea”. In : *Journal of Molecular Evolution* 29.2, p. 170-179.
- KODAMA, Miyako et al. (2014). “Comparative Mapping Between Coho Salmon (*Oncorhynchus kisutch*) and Three Other Salmonids Suggests a Role for Chromosomal Rearrange-

- ments in the Retention of Duplicated Regions Following a Whole Genome Duplication Event”. In : *G3: Genes, Genomes, Genetics* 4.9, p. 1717-1730.
- KÖSTER, Johannes et Sven RAHMANN (2012). “Snakemake—a scalable bioinformatics workflow engine”. In : *Bioinformatics* 28.19, p. 2520-2522.
- KRIVENTSEVA, Evgenia V. et al. (2019). “OrthoDB v10: sampling the diversity of animal, plant, fungal, protist, bacterial and viral genomes for evolutionary and functional annotations of orthologs”. In : *Nucleic Acids Research* 47.D1, p. D807-D811.
- KUMAR, Sudhir et al. (2017). “TimeTree: A Resource for Timelines, Timetrees, and Divergence Times”. In : *Molecular Biology and Evolution* 34.7, p. 1812-1819.
- LAALE, Hans W. (1977). “The biology and use of zebrafish, *Brachydanio rerio* in fisheries research.. A literature review”. In : *Journal of Fish Biology* 10.2, p. 121-173.
- LAFOND, Manuel et Nadia EL-MABROUK (2014). “Orthology and paralogy constraints: satisfiability and consistency”. In : *BMC Genomics* 15.6, S12.
- LAFOND, Manuel et al. (2013). “Gene tree correction guided by orthology”. In : *BMC Bioinformatics* 14.15, S5.
- LAI, Shih-Lei et al. (2017). “Reciprocal analyses in zebrafish and medaka reveal that harnessing the immune response promotes cardiac regeneration”. In : *eLife* 6. Sous la dir. de Marianne BRONNER, e25605.
- LAMICHHANEY, Sangeet et al. (2017). “Parallel adaptive evolution of geographically distant herring populations on both sides of the North Atlantic Ocean”. In : *Proceedings of the National Academy of Sciences* 114.17, E3452-E3461.
- LANDER, Eric S. et al. (2001). “Initial sequencing and analysis of the human genome”. In : *Nature* 409.6822, p. 860-921.
- LAURENT, Sacha, Nicolas SALAMIN et Marc ROBINSON-RECHAVI (2017). “No evidence for the radiation time lag model after whole genome duplications in Teleostei”. In : *PLOS ONE* 12.4, e0176384.
- LEE, Geraldine M. et James E. WRIGHT (1981). “Mitotic and meiotic analyses of brook trout, *Salvelinus fontinalis*”. In : *Journal of Heredity* 72.5, p. 321-327.
- LEMOINE, F. et al. (2018). “Renewing Felsenstein’s Phylogenetic Bootstrap in the Era of Big Data”. In : *Nature* 556.7702, p. 452-456.
- LEVY, Avraham A. et Moshe FELDMAN (2004). “Genetic and epigenetic reprogramming of the wheat genome upon allopolyploidization”. In : *Biological Journal of the Linnean Society* 82.4, p. 607-613.
- LI, Hao et Johan AUWERX (2020). “Mouse Systems Genetics as a Prelude to Precision Medicine”. In : *Trends in Genetics* 36.4, p. 259-272.
- LI, Ya-Juan et al. (2011). “The origin of natural tetraploid loach *Misgurnus anguillicaudatus* (Teleostei: Cobitidae) inferred from meiotic chromosome configurations”. In : *Genetica* 139.6, p. 805.
- LI, Zhen et al. (2016). “Gene Duplicability of Core Genes Is Highly Consistent across All Angiosperms[OPEN]”. In : *The Plant Cell* 28.2, p. 326-344.

- LIANG, Feng et al. (2000). “Gene Index analysis of the human genome estimates approximately 120,000 genes”. In : *Nature Genetics* 25.2, p. 239-240.
- LIBBRECHT, Maxwell W. et al. (2019). “A unified encyclopedia of human functional DNA elements through fully automated annotation of 164 human cell types”. In : *Genome Biology* 20.1, p. 180.
- LIEN, Sigbjørn et al. (2011). “A dense SNP-based linkage map for Atlantic salmon (*Salmo salar*) reveals extended chromosome homeologies and striking differences in sex-specific recombination patterns”. In : *BMC Genomics* 12.1, p. 615.
- LIEN, Sigbjørn et al. (2016). “The Atlantic salmon genome provides insights into rediploidization”. In : *Nature* 533.7602, p. 200-205.
- LIESCHKE, Graham J. et Peter D. CURRIE (2007). “Animal models of human disease: zebrafish swim into view”. In : *Nature Reviews Genetics* 8.5, p. 353-367.
- LIN, Yu et al. (2013). “Maximum Likelihood Phylogenetic Reconstruction from High-Resolution Whole-Genome Data and a Tree of 68 Eukaryotes”. In : *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, p. 285-296.
- LINDBLAD-TOH, Kerstin et al. (2011). “A high-resolution map of human evolutionary constraint using 29 mammals”. In : *Nature* 478.7370, p. 476-482.
- LLOYD, Andrew et al. (2018). “Homoeologous exchanges cause extensive dosage-dependent gene expression changes in an allopolyploid crop”. In : *New Phytologist* 217.1, p. 367-377.
- LOH, Yong Hwee et al. (2004). “Extensive Expansion of the Claudin Gene Family in the Teleost Fish, *Fugu rubripes*”. In : *Genome Research* 14.7, p. 1248-1257.
- LOWE, Craig B. et al. (2011). “Three periods of regulatory innovation during vertebrate evolution”. In : *Science (New York, N.Y.)* 333.6045, p. 1019-1024.
- LU, Hengyun, Francesca GIORDANO et Zemin NING (2016). “Oxford Nanopore MinION Sequencing and Genome Assembly”. In : *Genomics, Proteomics & Bioinformatics. SI: Big Data and Precision Medicine* 14.5, p. 265-279.
- LUCAS, Joseph MEX, Matthieu MUFFATO et Hugues Roest CROLLIUS (2014). “PhylDiag: identifying complex synteny blocks that include tandem duplications using phylogenetic gene trees”. In : *BMC Bioinformatics* 15.1.
- LUO, Jing et al. (2020). “From asymmetrical to balanced genomic diversification during rediploidization: Subgenomic evolution in allotetraploid fish”. In : *Science Advances* 6.22, eaaz7677.
- LYNCH, M. (2000). “The Evolutionary Fate and Consequences of Duplicate Genes”. In : *Science* 290.5494, p. 1151-1155.
- EL-MABROUK, Nadia et David SANKOFF (2012). “Analysis of Gene Order Evolution Beyond Single-Copy Genes”. In : *Evolutionary Genomics*. Sous la dir. de Maria ANISIMOVA. T. 855. Totowa, NJ : Humana Press, p. 397-429.

- MACQUEEN, Daniel J. et Ian A. JOHNSTON (2014). “A well-constrained estimate for the timing of the salmonid whole genome duplication reveals major decoupling from species diversification”. In : *Proceedings of the Royal Society B: Biological Sciences* 281.1778.
- MAKINO, Takashi et Aoife MCLYSAGHT (2010). “Ohnologs in the human genome are dosage balanced and frequently associated with disease”. In : *Proceedings of the National Academy of Sciences* 107.20, p. 9270-9274.
- (2012). “Positionally biased gene loss after whole genome duplication: Evidence from human, yeast, and plant”. In : *Genome Research* 22.12, p. 2427-2435.
- MALAGUTI, Giulia, Param Priya SINGH et Hervé ISAMBERT (2014). “On the retention of gene duplicates prone to dominant deleterious mutations”. In : *Theoretical Population Biology* 93, p. 38-51.
- MARLÉTAZ, Ferdinand et al. (2018). “Amphioxus functional genomics and the origins of vertebrate gene regulation”. In : *Nature* 564.7734, p. 64-70.
- MARTIN, Kyle J. et Peter W. H. HOLLAND (2014). “Enigmatic Orthology Relationships between Hox Clusters of the African Butterfly Fish and Other Teleosts Following Ancient Whole-Genome Duplication”. In : *Molecular Biology and Evolution* 31.10, p. 2592-2611.
- MASON, Annaliese S. et Jonathan F. WENDEL (2020). “Homoeologous Exchanges, Segmental Allopolyploidy, and Polyploid Genome Evolution”. In : *Frontiers in Genetics* 11.
- MCCARTHY, Mark I. et al. (2008). “Genome-wide association studies for complex traits: consensus, uncertainty and challenges”. In : *Nature Reviews Genetics* 9.5, p. 356-369.
- MCGRATH, Casey L. et al. (2014). “Insights into Three Whole-Genome Duplications Gleaned from the *Paramecium caudatum* Genome Sequence”. In : *Genetics* 197.4, p. 1417-1428.
- MEADOWS, Jennifer R. S. et Kerstin LINDBLAD-TOH (2017). “Dissecting evolution and disease using comparative vertebrate genomics”. In : *Nature Reviews Genetics* 18.10, p. 624-636.
- MEYER, Axel et Manfred SCHARTL (1999). “Gene and genome duplications in vertebrates: the one-to-four (-to-eight in fish) rule and the evolution of novel gene functions”. In : *Current Opinion in Cell Biology* 11.6, p. 699-704.
- MINH, Bui Quang et al. (2020). “IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era”. In : *Molecular Biology and Evolution* 37.5, p. 1530-1534.
- MINOCHE, André E., Juliane C. DOHM et Heinz HIMMELBAUER (2011). “Evaluation of genomic high-throughput sequencing data generated on Illumina HiSeq and Genome Analyzer systems”. In : *Genome Biology* 12.11, R112.
- MOORE, Jill E. et al. (2020). “Expanded encyclopaedias of DNA elements in the human and mouse genomes”. In : *Nature* 583.7818, p. 699-710.
- MOREL, Benoit et al. (2020). “GeneRax: A Tool for Species-Tree-Aware Maximum Likelihood-Based Gene Family Tree Inference under Gene Duplication, Transfer, and Loss”. In : *Molecular Biology and Evolution* 37.9, p. 2763-2774.

- MORIYAMA, Yuuta et al. (2016). “Evolution of the fish heart by sub/neofunctionalization of an elastin gene”. In : *Nature Communications* 7.1, p. 1-10.
- MUFFATO, Matthieu (2010). “Reconstruction de génomes ancestraux chez les vertébrés”. PhD Thesis. Université d'Évry Val d'Essonne.
- MUFFATO, Matthieu et Hugues Roest CROLLIUS (2008). “Paleogenomics in vertebrates, or the recovery of lost genomes from the mist of time”. In : *BioEssays* 30.2, p. 122-134.
- MUFFATO, Matthieu et al. (2010). “Genomicus: a database and a browser to study gene synteny in modern and ancestral genomes”. In : *Bioinformatics* 26.8, p. 1119-1121.
- MULLER, H. J. (1925). “Why Polyploidy is Rarer in Animals Than in Plants”. In : *The American Naturalist* 59.663, p. 346-353.
- (1950). “Our load of mutations”. In : *American Journal of Human Genetics* 2.2, p. 111-176.
- MULLER, J. et al. (2010). “eggNOG v2.0: extending the evolutionary genealogy of genes with enhanced non-supervised orthologous groups, species and functional annotations”. In : *Nucleic Acids Research* 38.suppl_1, p. D190-D195.
- NAGY, László G. et al. (2017). “Genetic Bases of Fungal White Rot Wood Decay Predicted by Phylogenomic Analysis of Correlated Gene-Phenotype Evolution”. In : *Molecular Biology and Evolution* 34.1, p. 35-44.
- NAGY, László G. et al. (2020). “Novel phylogenetic methods are needed for understanding gene function in the era of mega-scale genome sequencing”. In : *Nucleic Acids Research* 48.5, p. 2209-2219.
- NAKATANI, Yoichiro et Aoife MCLYSAGHT (2017). “Genomes as documents of evolutionary history: a probabilistic macrosynteny model for the reconstruction of ancestral genomes”. In : *Bioinformatics* 33.14, p. i369-i378.
- NARUSE, Kiyoshi et al. (2004). “A Medaka Gene Map: The Trace of Ancestral Vertebrate Proto-Chromosomes Revealed by Comparative Gene Mapping”. In : *Genome Research* 14.5, p. 820-828.
- NARUSE, K et al. (2000). “A detailed linkage map of medaka, *Oryzias latipes*: comparative genomics and genome evolution.” In : *Genetics* 154.4, p. 1773-1784.
- NEAR, T. J. et al. (2012). “Resolution of ray-finned fish phylogeny and timing of diversification”. In : *Proceedings of the National Academy of Sciences* 109.34, p. 13698-13703.
- NECSULEA, Anamaria et Henrik KAESSMANN (2014). “Evolutionary dynamics of coding and non-coding transcriptomes”. In : *Nature Reviews Genetics* 15.11, p. 734-748.
- NEGRISOLO, Enrico et al. (2010). “Different Phylogenomic Approaches to Resolve the Evolutionary Relationships among Model Fish Species”. In : *Molecular Biology and Evolution* 27.12, p. 2757-2774.
- NEHRT, Nathan L. et al. (2011). “Testing the Ortholog Conjecture with Comparative Functional Genomic Data from Mammals”. In : *PLOS Computational Biology* 7.6, e1002073.
- NELSON, Joseph S., Terry C. GRANDE et Mark V. H. WILSON (2016). *Fishes of the World*. John Wiley & Sons, Inc.

- NIETO FELINER, Gonzalo, Josep CASACUBERTA et Jonathan F. WENDEL (2020). “Genomics of Evolutionary Novelty in Hybrids and Polyploids”. In : *Frontiers in Genetics* 11.
- NOUTAHI, Emmanuel et al. (2016). “Efficient Gene Tree Correction Guided by Genome Evolution”. In : *PLOS ONE* 11.8, e0159559.
- NOVAK, Alicia E. et al. (2006). “Embryonic and larval expression of zebrafish voltage-gated sodium channel alpha-subunit genes”. In : *Developmental Dynamics* 235.7, p. 1962-1973.
- NÜSLEIN-VOLHARD, Christiane (2012). “The zebrafish issue of Development”. In : *Development* 139.22, p. 4099-4103.
- OHNO, Susumu (1970). “Introduction”. In : *Evolution by Gene Duplication*. Sous la dir. de Susumu OHNO. Berlin, Heidelberg : Springer, p. 1-2.
- OTTO, Sarah P. (2007). “The Evolutionary Consequences of Polyploidy”. In : *Cell* 131.3, p. 452-462.
- PALA, Irene et al. (2010). “Gene expression regulation and lineage evolution: the North and South tale of the hybrid polyploid *Squalius alburnoides* complex”. In : *Proceedings of the Royal Society B: Biological Sciences* 277.1699, p. 3519-3525.
- PANCHY, Nicholas, Melissa LEHTI-SHIU et Shin-Han SHIU (2016). “Evolution of Gene Duplication in Plants”. In : *Plant Physiology* 171.4, p. 2294-2316.
- PAREY, Elise et al. (2020). “Synteny-Guided Resolution of Gene Trees Clarifies the Functional Impact of Whole-Genome Duplications”. In : *Molecular Biology and Evolution* ().
- PASQUIER, Jeremy et al. (2016). “Gene evolution and gene expression after whole genome duplication in fish: the PhyloFish database”. In : *BMC Genomics* 17.1, p. 368.
- PELÉ, Alexandre, Mathieu ROUSSEAU-GUEUTIN et Anne-Marie CHÈVRE (2018). “Speciation Success of Polyploid Plants Closely Relates to the Regulation of Meiotic Recombination”. In : *Frontiers in Plant Science* 9.
- PLANET, Paul J. (2006). “Tree disagreement: Measuring and testing incongruence in phylogenies”. In : *Journal of Biomedical Informatics*. Phylogenetic Inferencing: Beyond Biology 39.1, p. 86-102.
- POLYCHRONOPOULOS, Dimitris et al. (2017). “Conserved non-coding elements: developmental gene regulation meets genome organization”. In : *Nucleic Acids Research* 45.22, p. 12611-12624.
- POSTLETHWAIT, J. H. et al. (2000). “Zebrafish comparative genomics and the origins of vertebrate chromosomes”. In : *Genome Research* 10.12, p. 1890-1902.
- POSTLETHWAIT, John H. et al. (1998). “Vertebrate genome evolution and the zebrafish gene map”. In : *Nature Genetics* 18.4, p. 345-349.
- POSTLETHWAIT, John et al. (2004). “Subfunction partitioning, the teleost radiation and the annotation of the human genome”. In : *Trends in Genetics* 20.10, p. 481-490.
- PRINCE, V. E. et al. (1998). “Zebrafish hox genes: expression in the hindbrain region of wild-type and mutants of the segmentation gene, *valentino*”. In : *Development* 125.3, p. 393-406.

- PROOST, Sebastian et al. (2012). “i-ADHoRe 3.0—fast and sensitive detection of genomic homology in extremely large data sets”. In : *Nucleic Acids Research* 40.2, e11-e11.
- RAMSEY, Justin et Douglas W. SCHEMSKE (1998). “Pathways, Mechanisms, and Rates of Polyploid Formation in Flowering Plants”. In : *Annual Review of Ecology and Systematics* 29, p. 467-501.
- RANDS, Chris M. et al. (2014). “8.2% of the Human Genome Is Constrained: Variation in Rates of Turnover across Functional Element Classes in the Human Lineage”. In : *PLOS Genetics* 10.7, e1004525.
- RHIE, Arang et al. (2020). “Towards complete and error-free genome assemblies of all vertebrate species”. In : *bioRxiv*, p. 2020.05.22.110833.
- RHOADS, Anthony et Kin Fai AU (2015). “PacBio Sequencing and Its Applications”. In : *Genomics, Proteomics & Bioinformatics. SI: Metagenomics of Marine Environments* 13.5, p. 278-289.
- ROBERTSON, Fiona M. et al. (2017). “Lineage-specific rediploidization is a mechanism to explain time-lags between genome duplication and evolutionary diversification”. In : *Genome Biology* 18, p. 111.
- ROBINSON-RECHAVI, Marc et al. (2001). “An ancestral whole-genome duplication may not have been responsible for the abundance of duplicated fish genes”. In : *Current Biology* 11.12, R458-R459.
- ROEST CROLLIUS, H. et al. (2000). “Estimate of human gene number provided by genome-wide analysis using Tetraodon nigroviridis DNA sequence”. In : *Nature Genetics* 25.2, p. 235-238.
- ROHLFS, Rori V. et Rasmus NIELSEN (2015). “Phylogenetic ANOVA: The Expression Variance and Evolution Model for Quantitative Trait Evolution”. In : *Systematic Biology* 64.5, p. 695-708.
- ROUX, Julien, Jialin LIU et Marc ROBINSON-RECHAVI (2017). “Selective Constraints on Coding Sequences of Nervous System Genes Are a Major Determinant of Duplicate Gene Retention in Vertebrates”. In : *Molecular Biology and Evolution* 34.11, p. 2773-2791.
- RUZICKA, Leyla et al. (2019). “The Zebrafish Information Network: new support for non-coding genes, richer Gene Ontology annotations and the Alliance of Genome Resources”. In : *Nucleic Acids Research* 47.D1, p. D867-D873.
- SACERDOT, Christine et al. (2018). “Chromosome evolution at the origin of the ancestral vertebrate genome”. In : *Genome Biology* 19.1, p. 166.
- SAITOH, Kenji, Wei-Jen CHEN et Richard L. MAYDEN (2010). “Extensive hybridization and tetraploidy in spined loach fish”. In : *Molecular Phylogenetics and Evolution* 56.3, p. 1001-1010.
- SALLAN, Lauren C. (2014). “Major issues in the origins of ray-finned fish (Actinopterygii) biodiversity”. In : *Biological Reviews* 89.4, p. 950-971.
- SALZBURGER, Walter (2018). “Understanding explosive diversification through cichlid fish genomics”. In : *Nature Reviews Genetics* 19.11, p. 705-717.

- SANDVE, Simen R., Rori V. ROHLFS et Torgeir R. HVIDSTEN (2018). “Subfunctionalization versus neofunctionalization after whole-genome duplication”. In : *Nature Genetics* 50.7, p. 908-909.
- SANTINI, Francesco et al. (2009). “Did genome duplication drive the origin of teleosts? A comparative study of diversification in ray-finned fishes”. In : *BMC Evolutionary Biology* 9.1, p. 194.
- SCANNELL, Devin R., Geraldine BUTLER et Kenneth H. WOLFE (2007). “Yeast genome evolution—the origin of the species”. In : *Yeast* 24.11, p. 929-942.
- SCHARTL, Manfred (2014). “Beyond the zebrafish: diverse fish species for modeling human disease”. In : *Disease Models & Mechanisms* 7.2, p. 181-192.
- SCORNAVACCA, Celine et Nicolas GALTIER (2017). “Incomplete Lineage Sorting in Mammalian Phylogenomics”. In : *Systematic Biology* 66.1, p. 112-120.
- SCORNAVACCA, Celine, Edwin JACOX et Gergely J. SZÖLLÖSI (2015). “Joint amalgamation of most parsimonious reconciled gene trees”. In : *Bioinformatics* 31.6, p. 841-848.
- SÉMON, Marie et Kenneth H. WOLFE (2007a). “Rearrangement Rate following the Whole-Genome Duplication in Teleosts”. In : *Molecular Biology and Evolution* 24.3, p. 860-867.
- (2007b). “Reciprocal gene loss between Tetraodon and zebrafish after whole genome duplication in their ancestor”. In : *Trends in Genetics* 23.3, p. 108-112.
- SEOIGHE, Cathal et Kenneth H. WOLFE (1999). “Yeast genome evolution in the post-genome era”. In : *Current Opinion in Microbiology* 2.5, p. 548-554.
- SESSION, Adam M. et al. (2016). “Genome evolution in the allotetraploid frog *Xenopus laevis*”. In : *Nature* 538.7625, p. 336-343.
- SHAO, Feng, Minjin HAN et Zuogang PENG (2019). “Evolution and diversity of transposable elements in fish genomes”. In : *Scientific Reports* 9.1, p. 15399.
- SHARMA, Virag et al. (2018). “A genomics approach reveals insights into the importance of gene losses for mammalian adaptations”. In : *Nature Communications* 9.1, p. 1215.
- SHIMODAIRA, H. et M. HASEGAWA (1999). “Multiple Comparisons of Log-Likelihoods with Applications to Phylogenetic Inference”. In : *Molecular Biology and Evolution* 16.8, p. 1114-1116.
- SHIMODAIRA, Hidetoshi (2002). “An Approximately Unbiased Test of Phylogenetic Tree Selection”. In : *Systematic Biology* 51.3, p. 492-508.
- SIMAKOV, Oleg et al. (2020). “Deeply conserved synteny resolves early events in vertebrate evolution”. In : *Nature Ecology & Evolution* 4.6, p. 820-830.
- SINGH, Param Priya, Jatin ARORA et Hervé ISAMBERT (2015). “Identification of Ohnolog Genes Originating from Whole Genome Duplication in Early Vertebrates, Based on Synteny Comparison across Multiple Genomes”. In : *PLoS Computational Biology* 11.7.
- SINGH, Param Priya et al. (2012). “On the Expansion of “Dangerous” Gene Repertoires by Whole-Genome Duplications in Early Vertebrates”. In : *Cell Reports* 2.5, p. 1387-1398.

- SONG, Michael J. et al. (2020). “Gene Balance Predicts Transcriptional Responses Immediately Following Ploidy Change in *Arabidopsis thaliana*”. In : *The Plant Cell* 32.5, p. 1434-1448.
- SONNHAMMER, Erik L.L. et Gabriel ÖSTLUND (2015). “InParanoid 8: orthology analysis between 273 proteomes, mostly eukaryotic”. In : *Nucleic Acids Research* 43.Database issue, p. D234-D239.
- SPAINK, Herman P, Hans J. JANSEN et Ron P. DIRKS (2014). “Advances in genomics of bony fish”. In : *Briefings in Functional Genomics* 13.2, p. 144-156.
- SPUHLER, J. N. (1948). “On the Number of Genes in Man”. In : *Science* 108.2802, p. 279-280.
- SRIVASTAVA, Prashant K. et al. (2018). “A systems-level framework for drug discovery identifies Csf1R as an anti-epileptic drug target”. In : *Nature Communications* 9.1, p. 3561.
- STAMATAKIS, Alexandros (2014). “RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies”. In : *Bioinformatics* 30.9, p. 1312-1313.
- STEBBINS, G. Ledyard (1947). “Types of Polyploids: Their Classification and Significance”. In : *Advances in Genetics*. Sous la dir. de M. DEMEREC. T. 1. Academic Press, p. 403-429.
- STEBBAN, Scott J. et John J. SCHENK (2017). “Muroid rodent phylogenetics: 900-species tree reveals increasing diversification rates”. In : *PLOS ONE* 12.8, e0183070.
- SUN, Honghe et al. (2017). “Karyotype Stability and Unbiased Fractionation in the Paleopolyploid Cucurbita Genomes”. In : *Molecular Plant* 10.10, p. 1293-1306.
- SYMONOVÁ, Radka et Alexander SUH (2019). “Nucleotide composition of transposable elements likely contributes to AT/GC compositional homogeneity of teleost fish genomes”. In : *Mobile DNA* 10.1, p. 49.
- SZÖLLŐSI, Gergely J. et al. (2013). “Efficient Exploration of the Space of Reconciled Gene Trees”. In : *Systematic Biology* 62.6, p. 901-912.
- SZÖLLŐSI, Gergely J. et al. (2015). “The inference of gene trees with species trees”. In : *Systematic Biology* 64.1, e42-62.
- TAMPLIN, Owen J. et al. (2012). “Small molecule screening in zebrafish: swimming in potential drug therapies”. In : *WIREs Developmental Biology* 1.3, p. 459-468.
- TAN, Haihan, Daria ONICHTCHOUK et Cecilia WINATA (2016). “DANIO-CODE: Toward an Encyclopedia of DNA Elements in Zebrafish”. In : *Zebrafish* 13.1, p. 54-60.
- TAN, Justin L. et Leonard I. ZON (2011). “Chemical screening in zebrafish for novel biological and therapeutic discovery”. In : *Methods in Cell Biology* 105, p. 493-516.
- TANNIER, Eric et al. (2020). “Ancestral Genome Organization as a Diagnosis Tool for Phylogenomics”. In : *Phylogenetics in the Genomic Era*. Sous la dir. de Celine SCORNAVACCA, Frédéric DELSUC et Nicolas GALTIER. No commercial publisher \textbar Authors open access book, 2.5:1-2.5:19.
- TAYLOR, John S. et al. (2003). “Genome Duplication, a Trait Shared by 22,000 Species of Ray-Finned Fish”. In : *Genome Research* 13.3, p. 382-390.
- THE ENCODE PROJECT CONSORTIUM (2004). “The ENCODE (ENCyclopedia Of DNA Elements) Project”. In : *Science* 306.5696, p. 636-640.

- THOMPSON, Andrew et al. (2020). *The genome of the bowfin (Amia calva) illuminates the developmental evolution of ray-finned fishes*. preprint. In Review.
- THOMSON, Robert C. et H. Bradley SHAFFER (2010). “Rapid progress on the vertebrate tree of life”. In : *BMC Biology* 8.1, p. 19.
- TRACHANA, Kalliopi et al. (2011). “Orthology prediction methods: A quality assessment using curated protein families”. In : *BioEssays* 33.10, p. 769-780.
- UYENO, Teruya et G. R. SMITH (1972). “Tetraploid Origin of the Karyotype of Catostomid Fishes”. In : *Science* 175.4022, p. 644-646.
- VAKIRLIS, Nikolaos, Anne-Ruxandra CARVUNIS et Aoife MCLYSAGHT (2020). “Synteny-based analyses indicate that sequence divergence is not the main source of orphan genes”. In : *eLife* 9. Sous la dir. de Diethard TAUTZ, Neel PRABH et Eve SYRKIN WURTELE, e53500.
- VAN DE PEER, Yves, Steven MAERE et Axel MEYER (2009). “The evolutionary significance of ancient genome duplications”. In : *Nature Reviews Genetics* 10.10, p. 725-732.
- (2010). “2R or not 2R is not the question anymore”. In : *Nature Reviews Genetics* 11.2, p. 166-166.
- VAN DE PEER, Yves, Eshchar MIZRACHI et Kathleen MARCHAL (2017). “The evolutionary significance of polyploidy”. In : *Nature Reviews Genetics* 18.7, p. 411-424.
- VANDEPOELE, Klaas et al. (2004). “Major events in the genome evolution of vertebrates: Paraneome age and size differ considerably between ray-finned fishes and land vertebrates”. In : *Proceedings of the National Academy of Sciences* 101.6, p. 1638-1643.
- VARADHARAJAN, Srinidhi et al. (2018). “The Grayling Genome Reveals Selection on Gene Expression Regulation after Whole-Genome Duplication”. In : *Genome Biology and Evolution* 10.10, p. 2785-2800.
- VENKATESH, Byrappa (2003). “Evolution and diversity of fish genomes”. In : *Current Opinion in Genetics & Development* 13.6, p. 588-592.
- VILELLA, Albert J. et al. (2009). “EnsemblCompara GeneTrees: Complete, duplication-aware phylogenetic trees in vertebrates”. In : *Genome Research* 19.2, p. 327-335.
- VOGEL, F. (1964). “A Preliminary Estimate of the Number of Human Genes”. In : *Nature* 201.4921, p. 847-847.
- WAGNER, Gunte P, Chris AMEMIYA et Frank RUDDLE (2003). “Hox cluster duplications and the opportunity for evolutionary novelties”. In : *Proceedings of the National Academy of Sciences* 100.25, p. 14603-14606.
- WANG, Maojun et al. (2018). “Evolutionary dynamics of 3D genome architecture following polyploidization in cotton”. In : *Nature Plants* 4.2, p. 90-97.
- WANG, Wei et al. (2020). “Changes in regeneration-responsive enhancers shape regenerative capacities in vertebrates”. In : *Science* 369.6508.
- WANG, Xiyin et al. (2005). “Duplication and DNA segmental loss in the rice genome: implications for diploidization”. In : *New Phytologist* 165.3, p. 937-946.
- WANG, Yupeng et al. (2012). “MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity”. In : *Nucleic Acids Research* 40.7, e49.

- WAPLES, R. K., L. W. SEEB et J. E. SEEB (2016). “Linkage mapping with paralogs exposes regions of residual tetrasomic inheritance in chum salmon (*Oncorhynchus keta*)”. In : *Molecular Ecology Resources* 16.1, p. 17-28.
- WATERHOUSE, Robert M et al. (2018). “BUSCO Applications from Quality Assessments to Gene Prediction and Phylogenomics”. In : *Molecular Biology and Evolution* 35.3, p. 543-548.
- WERTHEIM, B., L. W. BEUKEBOOM et L. van de ZANDE (2013). “Polyploidy in Animals: Effects of Gene Expression on Sex Determination, Evolution and Ecology”. In : *Cytogenetic and Genome Research* 140.2-4, p. 256-269.
- WITTBRODT, J., A. MEYER et M. SCHARTL (1998). “More genes in fish?” In : *BioEssays* 20.6, p. 511-515.
- WOLFE, Kenneth H. (2001). “Yesterday’s polyploids and the mystery of diploidization”. In : *Nature Reviews Genetics* 2.5, p. 333-341.
- WOODS, Ian G. et al. (2005). “The zebrafish gene map defines ancestral vertebrate chromosomes”. In : *Genome Research* 15.9, p. 1307-1314.
- WU, Yi-Chieh et al. (2013). “TreeFix: Statistically Informed Gene Tree Error Correction Using Species Trees”. In : *Systematic Biology* 62.1, p. 110-120.
- (2014). “Most parsimonious reconciliation in the presence of gene duplication, loss, and deep coalescence using labeled coalescent trees”. In : *Genome Research* 24.3, p. 475-486.
- XIONG, Zhiyong, Robert T. GAETA et J. Chris PIRES (2011). “Homoeologous shuffling and chromosome compensation maintain genome balance in resynthesized allopolyploid *Brassica napus*”. In : *Proceedings of the National Academy of Sciences of the United States of America* 108.19, p. 7908-7913.
- XU, Peng et al. (2019). “The allotetraploid origin and asymmetrical genome evolution of the common carp *Cyprinus carpio*”. In : *Nature Communications* 10.1, p. 4625.
- YANCOPOULOS, Sophia, Oliver ATTIE et Richard FRIEDBERG (2005). “Efficient sorting of genomic permutations by translocation, inversion and block interchange”. In : *Bioinformatics* 21.16, p. 3340-3346.
- YAO, Yao, Lorenzo CARRETERO-PAULET et Yves Van de PEER (2019). “Using digital organisms to study the evolutionary consequences of whole genome duplication and polyploidy”. In : *PLOS ONE* 14.7, e0220257.
- YATES, Andrew D. et al. (2020). “Ensembl 2020”. In : *Nucleic Acids Research* 48.D1, p. D682-D688.
- YUAN, Xuefei et al. (2018). “Heart enhancers with deeply conserved regulatory activity are established early in zebrafish development”. In : *Nature Communications* 9.1, p. 4977.
- ZAKON, Harold H. et al. (2006). “Sodium channel genes and the evolution of diversity in communication signals of electric fishes: Convergent molecular evolution”. In : *Proceedings of the National Academy of Sciences* 103.10, p. 3675-3680.

- ZHANG, Hui et al. (2019). “The effects of Arabidopsis genome duplication on the chromatin organization and transcriptional regulation”. In : *Nucleic Acids Research* 47.15, p. 7857-7869.
- ZHANG, Jianzhi (2003). “Evolution by gene duplication: an update”. In : *Trends in Ecology & Evolution* 18.6, p. 292-298.
- ZHAO, Tao et al. (2020). *Whole-genome microsynteny-based phylogeny of angiosperms*. preprint. In Review.
- ZHOU, Xiaofan et al. (2018). “Evaluating Fast Maximum Likelihood-Based Phylogenetic Programs Using Empirical Phylogenomic Data Sets”. In : *Molecular Biology and Evolution* 35.2, p. 486-503.

RÉSUMÉ

Les duplications complètes de génome sont des événements majeurs dans l'histoire évolutive des espèces. Elles produisent des copies surnuméraires de gènes qui peuvent acquérir de nouvelles fonctions et ainsi contribuer aux processus d'adaptation et de diversification. Deux duplications complètes de génome ont eu lieu dans la lignée précédant l'ancêtre des Vertébrés, suivies d'une troisième à la base des poissons téléostéens (datée à 320 millions d'années). L'impressionnante diversité du clade téléostéen, représentant plus de la moitié des espèces de Vertébrés actuelles, permet d'explorer un large éventail de questions fonctionnelles et évolutives. De fait, le séquençage récent et en cours de nombreuses espèces de poissons promet de compléter le modèle bien établi du poisson-zèbre. Néanmoins, leur événement partagé de duplication complète représente un défi pour l'analyse et la comparaison des génomes de poissons. En effet, suite à la duplication, de nombreux gènes demeurent en deux copies dans les génomes, ce qui complexifie la caractérisation des relations d'homologies entre gènes de différentes espèces. Afin de résoudre ce problème, j'ai développé une nouvelle méthodologie spécifique à la reconstruction d'arbres de gènes dans le contexte de duplications complètes de génomes, nommée SCORPiOs (Synteny-guided CORrection of Paralogies and Orthologies). L'innovation notable derrière SCORPiOs est l'intégration d'information provenant de l'organisation des gènes dans les génomes (synténie) afin de compléter les méthodes basées sur l'évolution moléculaire des séquences. Je présente comment l'application de cette nouvelle méthode à différents jeux de génomes de poissons améliore notre compréhension de l'évolution et de la structure des génomes de téléostéens. Dans un premier temps, je montre que SCORPiOs met en évidence la contribution des gènes dupliqués aux innovations évolutives des téléostéens. L'identification précise de gènes orthologues et paralogues m'a également permis d'établir la première cartographie à grande échelle des régions dupliquées entre génomes de poissons. Ce second résultat représente une nouvelle ressource qui devrait faciliter l'extrapolation d'annotations fonctionnelles entre espèces modèles et non-modèles. Enfin, je démontre comment l'analyse fine des désaccords de prédictions basées sur la synténie et la séquence permet de préciser les patrons spatio-temporels du retour à l'état diploïde après la duplication complète. Mon travail propose un cadre pour faciliter les analyses comparatives chez les poissons téléostéens et améliore nos connaissances concernant l'évolution des génomes après duplication complète.

MOTS CLÉS

Génomique comparative ; duplication complète de génome ; évolution moléculaire ; arbres de gènes ; synténie.

ABSTRACT

Whole-genome duplications are major events in the evolutionary history of species. They produce additional gene copies that can acquire new functions and thus contribute to adaptation and diversification processes. Two rounds of whole genome duplications occurred in the lineage leading to the Vertebrate ancestor, followed by a subsequent one at the stem of the teleost fish clade (dated 320 million years ago). The impressive diversity of the teleost clade, accounting for over half of extant vertebrate species, allows us to address a vast panel of functional and evolutionary questions. As such, the recent and on-going sequencing of many fish species promises to neatly complement the well-established zebrafish model. However, their shared whole genome duplication represents an additional layer of complexity that has to be accounted for when comparing fish genomes. Indeed, many genes still remain in two copies after the duplication, which renders the identification of homologous genes across species extremely complex. To tackle this challenge, I have developed a novel method, named SCORPiOs (Synteny-guided CORrection of Paralogies and Orthologies), which reconstructs more accurate phylogenetic gene trees in the context of whole genome duplications. The major innovation behind SCORPiOs is that it integrates information from the genomic organisation of genes (synteny) to complement classical sequence-based methods. I present how the application of SCORPiOs to various fish genomes datasets enhances our understanding of fish genome structure and evolution. First, I show that SCORPiOs links duplicate gene retention to evolutionary novelties in the teleost clade. Further, the precise identification of orthologous and paralogous genes allowed me to establish the first large-scale cartography of WGD-duplicated regions across fish genomes. This second result represents a novel resource that should facilitate the transfer of functional annotations between model and non-model fish species. Last, I demonstrate how the analysis of discordances between sequence and synteny predictions sheds light on the spatio-temporal pattern of rediploidization following the duplication event. My work provides a framework that facilitates comparative analyses across teleost fish genomes and reveals insights into the evolution of genomes following whole genome duplication.

KEYWORDS

Comparative genomics ; whole-genome duplication ; molecular evolution ; gene trees ; synteny.