Détection de l'évolution convergente à l'échelle génomique : développement de méthodes et étude des adaptations indépendantes à la vie en milieu aride chez les rongeurs

Carine Rey

Numéro National de Thèse : 2019LYSEN060

THÈSE de DOCTORAT DE L'UNIVERSITÉ DE LYON

opérée par

l'École Normale Supérieure de Lyon

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

Spécialité de doctorat : Biologie

Soutenue publiquement le 4 novembre 2019, par :

Carine REY

Détection de l'évolution convergente à l'échelle génomique : développement de méthodes et étude des adaptations indépendantes à la vie en milieu aride chez les rongeurs

---

Devant le jury composé de :

BOUSSAU,	Bastien	Chargé de recherche	CNRS Lyon 1	Co-encadrant de thèse
DELATTRE,	Marie	Directrice de recherche	CNRS ENS de Lyon	Examinatrice
DESSIMOZ,	Christophe	Professeur SNSF	Université de Lausanne	Rapporteur
GASCUEL,	Olivier	Directeur de recherche	Institut Pasteur	Examinateur
RANWEZ,	Vincent	Professeur	Montpellier SupAgro	Rapporteur
SÉMON,	Marie	Maitre de conférences	ENS de Lyon	Directrice de thèse

Résumé de la thèse

La convergence phénotypique, c'est-à-dire l'acquisition indépendante de caractères similaires par des espèces différentes, est omniprésente dans la nature et a été souvent étudiée. Mais ce processus évolutif n'est pas bien compris. Par exemple, de nombreux chercheurs cherchent à comprendre s'il existe des bases génétiques convergentes sous-jacentes à ces convergences phénotypiques.

Quelques substitutions convergentes corrélées à un phénotype convergent ont été décrites dans la littérature, mais il existe peu d'études à l'échelle génomique. Ceci peut s'expliquer par deux problèmes méthodologiques : 1/ D'une part, la difficulté de créer des jeux de données multi-espèces pour des analyses comparatives. 2/ D'autre part, le manque de méthodes dédiées à la détection de la convergence à l'échelle génomique.

Au cours de ma thèse, j'ai proposé des solutions à ces deux défis. Dans un premier temps, j'ai créé un programme (CAARS) permettant d'automatiser l'assemblage de jeux de données composés de familles d'orthologues à partir de données RNA-Seq. Puis, j'ai créé un outil (PCOC) pour étudier les substitutions convergentes au sein de séquences codantes, basé sur l'identification de changements de profils d'acides aminés. Ces outils ont été développés dans un souci de reproductibilité et de facilité d'utilisation. J'ai ensuite étudié la capacité de différentes méthodes, dont PCOC, à détecter des substitutions convergentes en présence de facteurs confondants. Enfin, j'ai appliqué ces méthodes à un cas biologique où j'ai cherché à caractériser les bases génomiques de l'adaptation aux milieux arides chez les rongeurs.

PhD thesis summary

Phenotypic convergence, the independent acquisition of similar characters by different species, is widespread in nature and has been extensively studied. But this evolutionary process is not well understood. For example, many researchers seek to understand whether there are convergent genetic bases underlying these phenotypic convergences.

Some convergent substitutions correlated with a convergent phenotype have been described in the literature, but there are few studies at the genome scale. This can be explained by two methodological problems : 1 / On the one hand, the difficulty of creating multi-species datasets for comparative analyses. 2 / On the other hand, the lack of dedicated methods to detect convergence at the genomic scale.

During my thesis, I proposed solutions to these two challenges. As a first step, I created a program (CAARS) to automate the assembly of datasets composed of orthologous families from RNA-Seq data. Then I created a tool (PCOC) to study convergent substitutions within coding sequences, based on the identification of amino acid profile changes rather than strict amino acid changes. These tools have been developed for the sake of reproducibility and ease of use. I then studied the ability of different methods, including PCOC, to detect convergent substitutions in the presence of confounding factors. Finally, I applied these methods to a biological case where I sought to characterize the genomic bases of adaptation to arid environments in rodents.

A toutes celles et ceux qui se sont intéressés à mon sujet de thèse et à qui,
pour diverses raisons, je n'ai pas eu le temps nécessaire
de le leur expliquer.

REMERCIEMENTS

Tout d'abord je voudrais remercier tous les membres du jury d'avoir accepté d'évaluer ce travail et particulièrement à Christophe Dessimoz et Vincent Ranwez, rapporteurs de cette thèse.

Merci à Marie et Bastien de m'avoir encadrée pendant ces quatre ans et demi. Merci de m'avoir fait confiance et de m'avoir accordé autant de liberté tout en veillant aux directions que je prenais. Merci de m'avoir appris votre sens du détail. Merci pour tout. Travailler avec vous est un réel plaisir. Marie, merci de m'avoir laissé l'opportunité d'enseigner et de m'avoir confié des responsabilités. Je me suis vraiment épanouie dans les enseignements que j'ai préparés et donnés.

Merci à toutes les cigognes, de passage ou installées. Merci d'avoir assuré en permanence une ambiance de travail conviviale. Le bureau était petit, mais peut-être tout aussi petit que chaleureux. Je rougis rien qu'à l'idée d'imaginer la quantité de chocolats qui a pu y séjourner de manière très très temporaire. Merci à Domitille, tu es une collègue fantastique. Il n'en manquait pas beaucoup pour que je te convertisse au badminton. Merci à Jérémy pour ta gentillesse et toutes ces anecdotes dont tu ne taris pas. Bon courage à toi pour la fin de ta thèse. Merci à Mathilde. Je ne sais pas d'où te vient toute ton énergie mais c'est un réel plaisir de partager le bureau avec toi. Merci à Sophie, pour ta sympathie et ton regard sur l'évo-dévo que tu sais si bien partager. Un jour, j'en suis convaincue, je te convertirais à R.

Merci à toi Thibault. Je t'ai toujours considéré comme un petit cigogneau. Nous avons commencé notre thèse ensemble et je pense que ma thèse aurait été très différente sans tous ces repas que l'on a partagés et toutes ces discussions. Tu es vraiment un type hors du commun. Merci pour ton aide pour les documentations de CAARS et PCOC, je pense qu'elles seraient bien moins claires si je les avais faites sans toi. Merci également de m'avoir convertie à l'escalade et puis à Corentin, Mathilde et Domitille pour toutes ces séances de blocs. On n'était pas si assidus mais on est quand même arrivés à passer quelques violettes.

Merci à toutes les membres du LBBE, ma deuxième maison. Même si je ne venais que de temps en temps, je me suis toujours sentie "comme à la maison".

Je voudrais également remercier toutes les personnes qui m'ont donné goût à la bioinformatique et qui m'ont permise de développer mes compétences. Alors merci à Laurent, Céline, Jean-Pierre, Pierre-Alexandre, Frédéric, Marie, Bastien, Coraline, Thomas, Loïs, Philippe et Vincent.

Merci également à mes vieilles potes de prépa, et oui cela va faire neuf ans. Neuf ans que vous me supportez et je ne sais toujours pas comment vous faites. Merci pour tout ces moments que l'on a passés ensemble, c'est toujours une réelle joie de vous retrouver.

Pour conclure, je voudrais remercier toutes les personnes qui m'ont soutenue, encouragée, motivée, aidée pour la rédaction de ce manuscrit, qui a été je pense, l'une des plus grandes montagnes que j'ai eu à gravir. Alors merci à ma familles, mes amis et à Marie, Bastien, Sophie, Domitille, Camille, Thibault, Philippe, Joséphine et Élise. Merci également à toutes celles et ceux qui m'ont proposé leur aide. C'est une source de motivation supplémentaire de se sentir si bien entourée et soutenue.

Merci également aux lecteurs de ce manuscrit qui vont prendre du temps pour comprendre le sujet qui m'a tenu en haleine pendant ces quatre ans.

Table des matières

Liste des figures

Liste des tableaux

1 Introduction : De la convergence phénotypique à la convergence génomique

Sommaire

1.1 L'évolution convergente, une fourberie de l'évolution bien utile

1.1.1 La caféine, un premier cas de convergence stimulant

Comme de nombreuses personnes, vous appréciez peut-être de commencer votre journée par un café ou un thé car il vous permet de vous réveiller tout en douceur. En effet, ces deux boissons contiennent une molécule, la caféine, qui a la capacité de stimuler les fonctions cognitives et donc de lutter contre la somnolence pendant une certaine durée.

1.1.1.1 Plusieurs espèces synthétisent de la caféine

Le caféier (*Coffea arabica*) et le théier (*Camellia sinensis*) ne sont pas les seules plantes contenant de la caféine. Cette molécule est aussi retrouvée, entre autres, dans les graines de cacao (*Theobroma cacao*) et de la guarana (*Paullinia cupana*), dans les feuilles de yerba maté (*Ilex paraguariensis*) - utilisées pour faire la boisson traditionnelle sud-américaine, le maté - ainsi que dans les fleurs d'orangers (*Citrus sinensis*) (ASHIHARA, 2004). On pourrait penser que toutes ces plantes capables de produire cette molécule - pas loin de 60 ! - sont issues du même genre ou de la même famille ou plus simplement sont des espèces proches les unes des autres car elles partagent une caractéristique commune.

Par exemple, tous les arbres qui produisent des agrumes appartiennent à la même famille (les Rutacées). On dit que ces espèces d'arbres sont de la même famille car elles sont toutes issues d'un même ancêtre commun présent en Asie de l'Est il y a 6 à 8 millions d'années (WU et collab., 2018). A cette époque, il n'existait pas de citronniers, ni d'orangers sur Terre mais une espèce d'arbres qui produisait des fruits qui avaient les caractéristiques des agrumes (un fruit avec des quartiers, une écorce et des pépins). Puis, des milliers d'années se sont écoulées et cette espèce a laissé place à trois nouvelles espèces, qui ont chacune gardé certaines caractéristiques de l'espèce ancestrale mais dont la taille et la couleur des fruits se sont diversifiées : les mandariniers (*Citrus reticulata*) dans le sud de la Chine, les pamplemoussiers (*Citrus maxima*) dans l'archipel Malaisien et les cédratiers (*Citrus medica*) dans le nord-est de l'Inde (WU et collab., 2018). Les citronniers et orangers que nous connaissons aujourd'hui sont issus de croisements récents réalisés par l'Homme mais ils appartiennent quand même à la famille des rutacées car ils descendent de cette espèce ancestrale.

Ce qu'il faut retenir ici n'est pas le nom scientifique de ces espèces mais plutôt que les espèces qui tendent à se ressembler et qui partagent des caractéristiques communes proviennent probablement d'un même ancêtre commun, qui avait lui même ces caractéristiques et qui les leur a "léguées". On peut alors utiliser ces caractéristiques pour retracer les relations de parenté entre espèces via un **arbre phylogénétique**, également appelé **phylogénie**.

1.1.1.2 Brève introduction à la phylogénie

La construction d'une phylogénie doit reposer sur des caractéristiques contenant de l'information sur la parenté des espèces étudiées, c'est à dire des caractères **homologues** qui ont été hérités d'un ancêtre commun (Figure 1.1). On distingue les caractères homologues des caractères **analogues** qui ont été acquis indépendamment d'un ancêtre commun. Par exemple, les fleurs d'un citronnier et d'un plan de maïs sont homologues car elles sont héritées de l'ancêtre commun des plantes à fleurs, les angiospermes. Par contre, les épines sur les cactus et les rosiers sont analogues car elles ne dérivent pas du même organe chez leur dernier ancêtre commun.

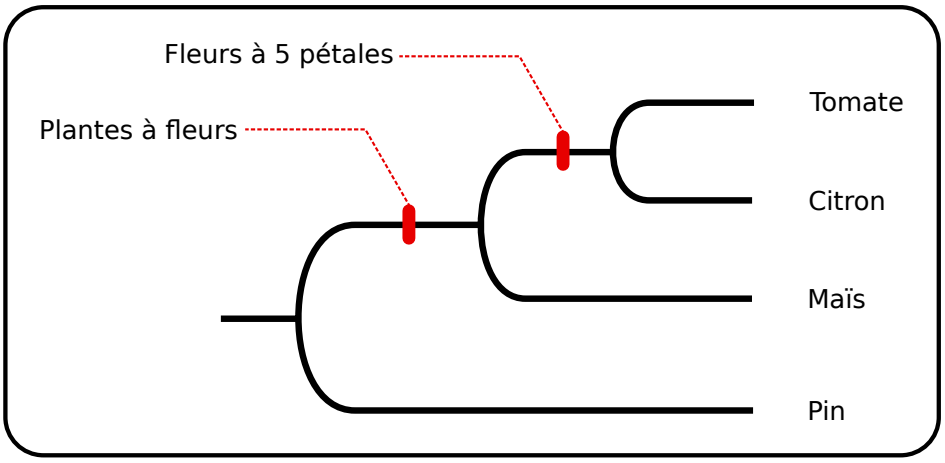

FIGURE 1.1 – Phylogénie montrant les relations entre 4 espèces de plantes ainsi que l'évolution de certains traits au cours du temps.

Pour reconstruire les relations de parenté entre des espèces, il faut donc utiliser des caractères homologues. En effet, l'utilisation de caractères analogues conduirait à la reconstruction d'une phylogénie erronée.

1.1.1.3 La synthèse de la caféine est un cas d'évolution convergente

Si l'on revient au cas de la caféine, les espèces capables de la synthétiser n'appartiennent pas du tout au même genre, ni même à la même famille. L'ancêtre commun des espèces que j'ai citées précédemment remonte à plus de 110 millions d'années et est aussi l'ancêtre commun de toutes les plantes à fleurs à 5 pétales (*Pentapetalae*). Il y a alors deux solutions possibles, soit l'ensemble des plantes à fleurs a perdu la capacité de synthétiser cette molécule, sauf ces quelques espèces, soit la capacité de produire de la caféine par ces plantes a été acquise de manière indépendante plusieurs fois dans l'évolution. La première solution ne semble pas vraisemblable. Il a en effet été démontré que la capacité de produire de la caféine par ces plantes a été acquise de manière indépendante plusieurs fois dans l'évolution (ASHIHARA, 2004) (Figure 1.2). On parle alors d'**évolution convergente**.

Cet exemple de convergence évolutive est sans doute passé inaperçu au cours de l'histoire et n'a pas eu de conséquences sur la classification de ces différentes espèces. En effet, les premiers naturalistes qui ont classifié les êtres vivants, dont les plantes, ont utilisé des caractéristiques générales pour classifier les plantes (formes des feuilles, nombre de pétales, etc.). Cependant, d'autres cas d'évolutions convergentes au niveau morphologique ont mené à des erreurs de classification des espèces qui ont perduré dans le temps et qui, pour certaines d'entre elles, n'ont pu être résolues que récemment.

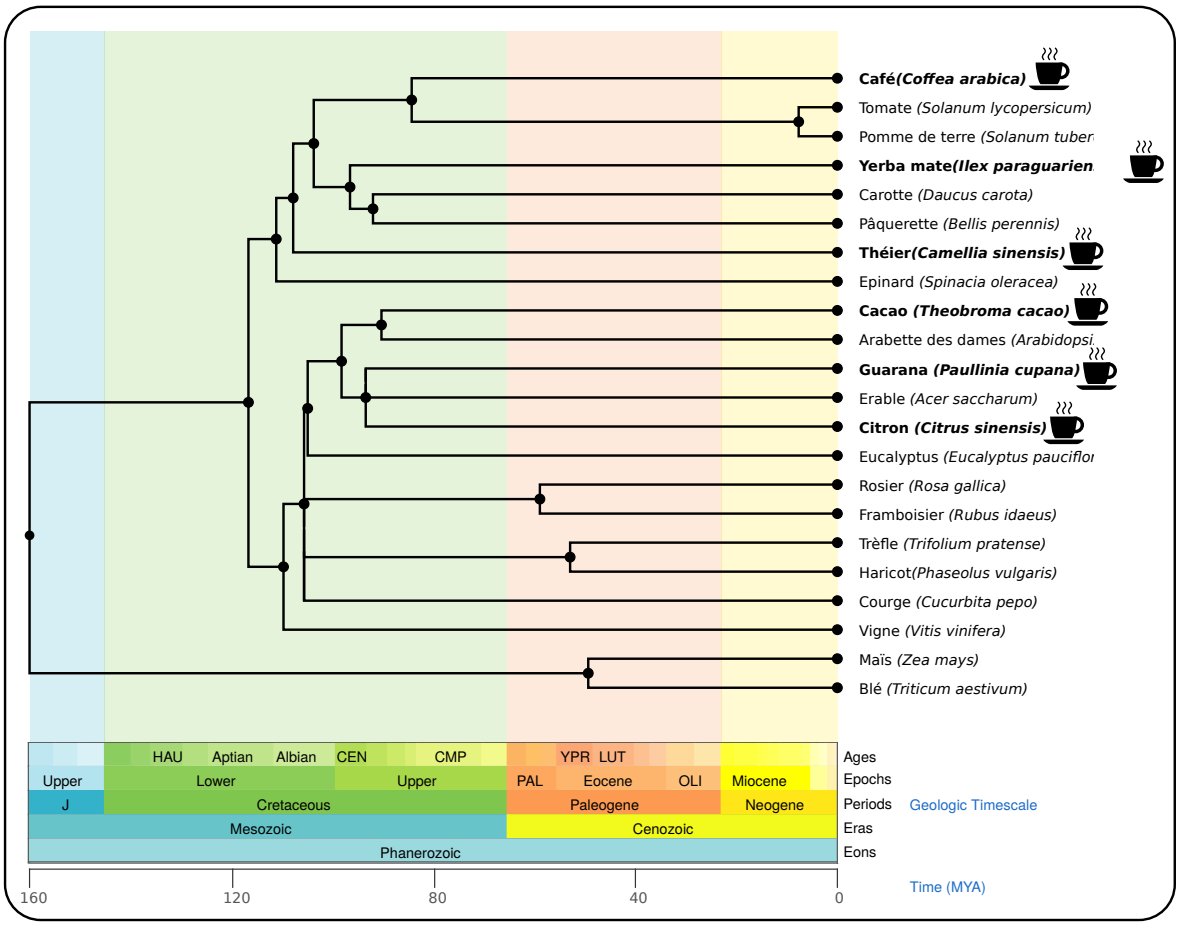

FIGURE 1.2 – Phylogénie datée d'une sélection d'angiospermes (plantes à fleurs). Les plantes capables de synthétiser la caféine sont indiquées par une tasse de café. Figure adaptée d'une figure de (LOSOS, 2017) et réalisée à l'aide de http://www.timetree.org/.

1.1.2 Des convergences trompeuses

1.1.2.1 Les cétacés, de monstres des mers à mammifères marins

Aujourd'hui, si l'on demande à quelqu'un (non passionné par les mammifères marins mais ayant tout de même quelques connaissances en biologie) de classifier les baleines, dauphins, otaries ou lamantins parmi les êtres vivants, il y a de fortes chances qu'il/elle les place parmi les mammifères car "ils ont des poils" et "allaitent leurs petits" bien qu'ils vivent dans l'eau à la différence des mammifères terrestres. Si on lui demande ensuite de les classer parmi les mammifères, il y a, à nouveau, de fortes chances qu'il réponde qu'il ne sait pas, mis à part qu'il les placerait dans un groupe à part entière car "ils n'ont pas de pattes, ils se ressemblent et ils vivent dans l'eau". En effet, ce raisonnement est logique, on apprend à l'école à classer dans une même catégorie les choses qui se ressemblent.

Mais, si cette personne est un passionné des mammifères marins ou qu'il a suivi des cours sur la classification des mammifères, il vous dira : "Tous les mammifères marins ne forment pas un même et unique groupe, il y a eu 3 transitions indépendantes vers le milieu marin chez les mammifères (Figure 1.3). La première, il y a environ 65 millions d'années, où les Siréniens qui regroupent les lamantins et les dugongs se sont séparés de leur dernier ancêtre commun avec les éléphants. La seconde, il y a environ 53 millions d'années, où les Cétacés qui regroupent entre autres les baleines, les dauphins et les orques se sont séparés de leur dernier ancêtre commun avec les hippopotames. La dernière, il y a 40 millions d'années, où les Pinnipedia qui regroupent les phoques, les otaries et les morses se sont séparés de leur dernier ancêtre commun avec les ratons laveurs et les ours. (note :

Les Canidae qui contiennent les chiens se sont séparés de ce groupe il y a 46 millions d'années) (KUMAR et collab., 2017)."

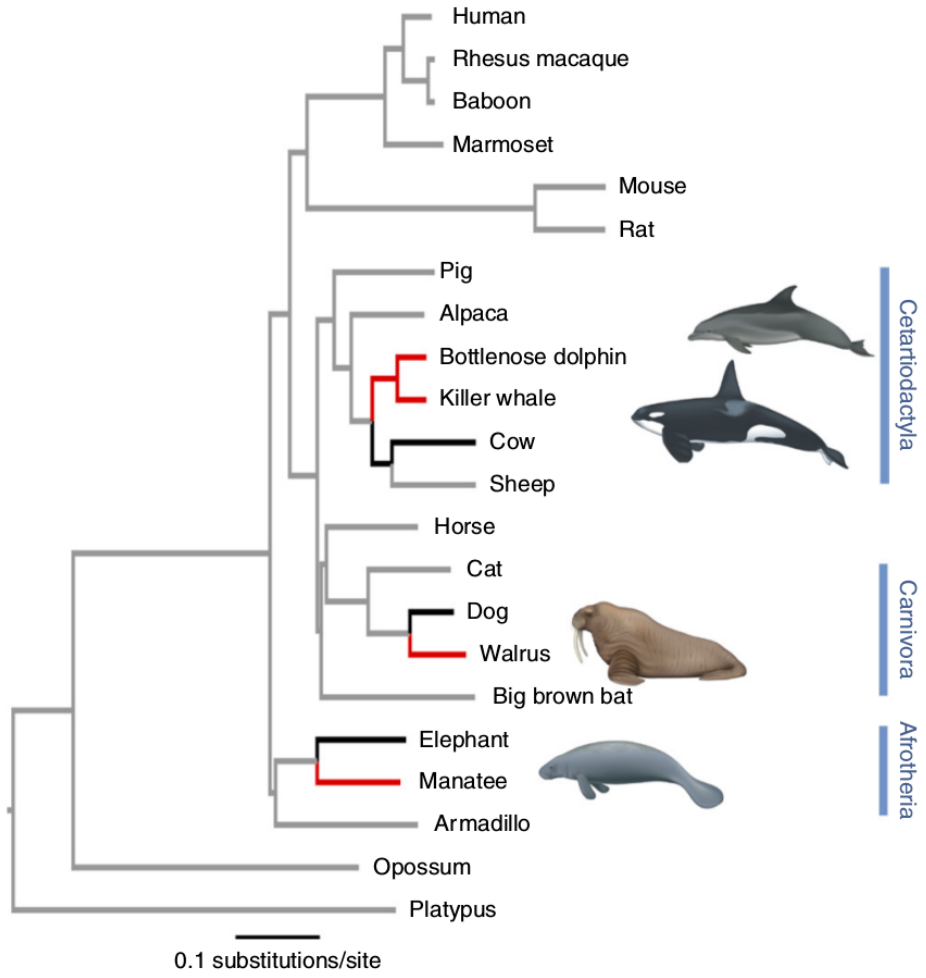

FIGURE 1.3 – Phylogénie simplifiée des mammifères avec les 3 transitions vers le milieu marin. D'après (FOOTE et collab., 2015).

Si nous sommes en mesure, aujourd'hui, de classifier correctement les groupes de mammifères (et les êtres vivants en général) et de pouvoir en parler comme si c'était une évidence, c'est grâce au travail soigneux de tous les scientifiques des années et siècles précédents, alors même qu'ils ne disposaient pas de nos moyens techniques actuels. Ce sont eux, qui par leurs observations, leurs hypothèses, leurs débats, parfois houleux, ont permis une continuelle remise en cause des connaissances contemporaines et qui ont pu accéder à la vérité actuelle (qui pourra être remise en cause demain).

Le processus historique d'affectation des mammifères marins aux mammifères et plus précisément des baleines, a été retracé par Aldemaro Romero (ROMERO, 2012). Je vous en propose un court résumé car il permet également de retracer le processus historique de la classification des êtres vivants ainsi que de la vision des espèces par l'Homme.

Aristote, qui est surtout connu pour ses travaux en philosophie, s'est aussi intéressé à l'histoire naturelle à l'époque de la Grèce Antique (IVème siècle avant JC). Il est le plus ancien auteur à avoir étudié la classification des animaux dont nous avons des écrits. Dans son livre, *Historia Animalium*, il propose une classification des êtres vivants en groupes ordonnés nommés "genus" basés sur leur stade de "perfection". Le but de cette classification n'avait pas une fin biologique, mais plutôt philosophique pour montrer l'ordre rationnel de l'Univers. Le "genus" au sommet de sa classification était "naturellement" l'Homme, puis les vivipares quadrupèdes (les animaux

à quatre pattes dont les embryons se développent dans la femelle, ce que l'on pourrait nommer aujourd'hui les mammifères terrestres), puis les ovipares quadrupèdes (les animaux qui pondent des oeufs et ayant quatre pattes, ce qui correspondrait aux reptiles et amphibiens), les oiseaux, les cétacés (venant du grecs "κῆτος" [kêtos] signifiant "gros poisson de mer" ou "monstre marin"), les poissons, les céphalopodes (calamar, pieuvres, etc.), les crustacés, les mollusques, les invertébrés, les zoophytes (animaux ressemblant à des plantes) puis les plantes supérieures (plantes à fleurs) et enfin les plantes inférieures (plantes sans fleurs).

Bien qu'Aristote avait remarqué que les "cétacés" partageaient des similarités avec les "vivipares quadrupèdes" (mammifères terrestres) tels que les poils, les poumons et les glandes mammaires, l'absence de pattes justifiait leur place en dessous des reptiles et amphibiens. Il justifia la distinction entre "cétacés" et "poissons" par, entre autres, des différences de reproduction (vivipares et ovipares) et des systèmes respiratoires (poumons contre branchies).

Les mammifères marins devront attendre jusqu'au 18ème siècle et Carl von Linné pour rejoindre les mammifères. En effet, ce naturaliste suédois, proposa une nouvelle classification dans son livre *Systema Naturae* basée sur une nomenclature hiérarchique et rationnelle où des caractéristiques morphologiques sont utilisées pour définir chacun des groupes taxonomiques. Bien que les mammifères marins aient été placés dans le groupe des poissons lors de la première édition, les connaissances amassées par les nombreux naturalistes de l'époque ont permis de les classer dans le groupe des mammifères à la 10ème édition en 1758. Il est également important de noter que les trois groupes de mammifères marins ont été attribués à trois différents sous groupes taxonomiques des mammifères.

La fine observation des naturalistes du XVIII ème siècle a permis de résoudre "le piège" tendu par l'évolution. En effet, les caractéristiques morphologiques telles que la dentition ou le nombre de doigts des espèces ont été les clés pour replacer les trois groupes de mammifères marins parmi les mammifères terrestres malgré la forte ressemblance aux autres animaux aquatiques. Cependant, des convergences évolutives au niveau de critères morphologiques plus précis compliquent la classification d'espèces relativement proches.

Jusqu'au milieu du XIX ème siècle, le but de cette classification était d'organiser les espèces tel que Dieu les avaient créées et non de retracer les relations de parentés entre les espèces. Ce n'est qu'avec les idées de Lamarck et Darwin que les espèces ne sont plus vues comme des entités fixées. Dans son célèbre livre *L'Origine des espèces* (1859), Charles Darwin fut un des premiers à formaliser la théorie de l'évolution telle qu'elle est communément acceptée aujourd'hui, malgré quelques ajouts et modifications. Il proposa notamment que les espèces étaient apparentées et que la sélection naturelle était le moteur de l'évolution.

1.1.3 L'ADN et la phylogénie moléculaire à l'aide de la classification des mammifères placentaires

1.1.3.1 Les phénotypes convergents peuvent perturber les phylogénies basées sur des critères morphologiques

Les relations de parenté entre les mammifères placentaires, qui comprennent l'ensemble des mammifères dont les embryons se développent dans le corps des femelles et qui se nourrissent grâce au placenta, n'ont été résolues que "récemment". Les quatre sous-groupes actuellement admis, les Afrotheria, les Xenarthra, les Euarchontoglires et les Laurasiatheria, ont été définis entre la fin des années 1990 et le début des années 2000 (SPRINGER, 2004), Figure 1.4. Mais pourquoi a-t-il fallu attendre si longtemps ?

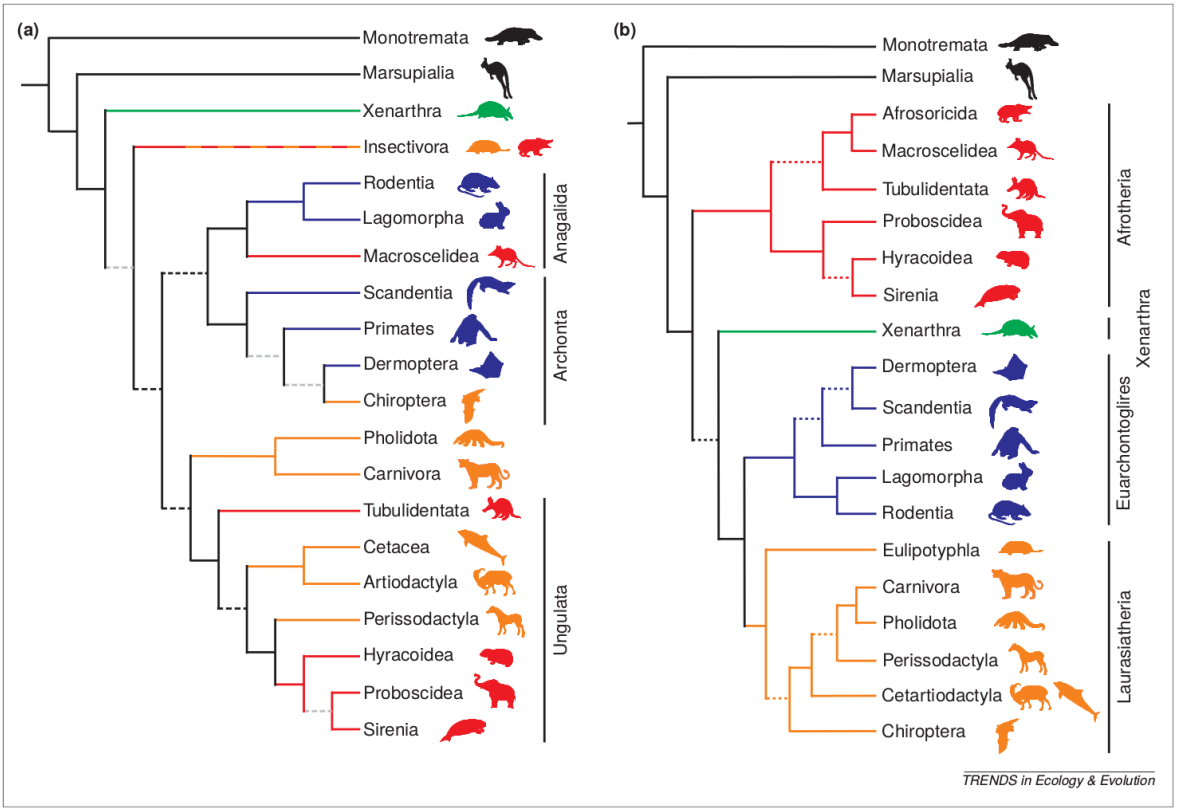

FIGURE 1.4 – Redéfinition des groupes des mammifères. Phylogénie des mammifères basée sur des critères morphologiques acceptés jusqu'à la fin des années 1990 (a) et basée sur des critères moléculaires proposés début des années 2000 (b). Les pointillés représentent les relations peu soutenues par les données. Les relations entre les Afrotheria, les Xenarthra et la paire Laurasiatheria et Eurarchontroglires ne sont pas encore résolues. D'après la Figure 1 de (SPRINGER, 2004).

L'observation de caractéristiques morphologiques, des espèces existantes et des fossiles de mammifères ont permis d'identifier trois grands groupes de mammifères : les monotrèmes (ex : ornithorynques), les marsupiaux (ex : kangourous) et les euthériens (majoritairement composés des mammifères placentaires) (O'LEARY et collab., 2013). Le dernier ancêtre commun de ces trois groupes de mammifères remonte à environ 166 millions d'années et celui des marsupiaux et des euthériens à environ 148 millions d'années (Figure 1.5). Les premiers euthériens et marsupiaux ont donc été contemporains des dinosaures et vivaient lors de la Pangée (lorsque tous les continents étaient réunis en un immense continent).

Jusqu'à 65 millions d'années, le monde était dominé par les dinosaures tandis que les marsupiaux et les euthériens étaient cantonnés à une vie nocturne pour éviter les crocs des dinosaures (MAOR et collab., 2017). [1] La chute d'une météorite, probablement combinée à une importante activité volcanique, a entraîné l'extinction de la majorité des dinosaures (il reste aujourd'hui uniquement les oiseaux) et ainsi la fin de la suprématie des dinosaures. Les mammifères ont alors profité de la place libérée par les dinosaures dans les écosystèmes et ont investi les **niches écologiques** vacantes.

1. Cette vision des mammifères de cette époque est volontairement caricaturale. En effet, de nouveaux fossiles de mammifères de cette époque montrent une diversité jusqu'à maintenant ignorée mais je ne rentrerai pas dans ces détails.

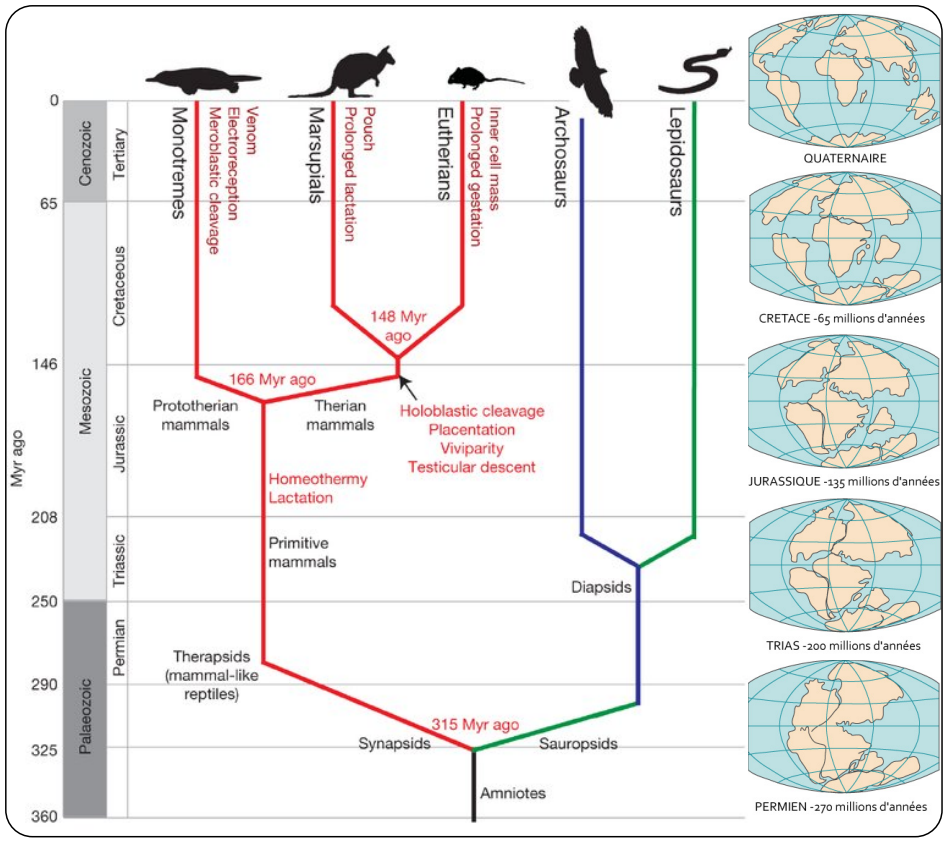

FIGURE 1.5 – Emergence des caractéristiques morphologiques distinctives des grands groupes de mammifères au cours du temps (Gauche). Position des continents au cours du temps (Droite). Figure de gauche d'après la Figure 1 (WARREN et collab., 2008). Figure de droite http://www.fondation-lamap.org/en/node/167.

Une niche écologique peut être définie par un rôle, une place dans un écosystème. Par exemple, les espèces herbivores et les insectivores occupent des niches différentes. Si deux espèces occupent la même niche écologique, elles seront alors en compétition, et d'après le principe de la **sélection naturelle**, la plus adaptée se reproduira le mieux et survivra le plus efficacement. D'autre part, des niches écologiques très similaires peuvent être occupées par des espèces différentes dans deux écosystèmes différents, par exemple en Europe et en Australie. On parle d'espèces **spécialistes**, pour les espèces qui occupent des niches écologiques très précises contrairement aux espèces dites **généralistes** qui occupent des niches écologiques plus larges.

La diversification des mammifères qui a suivi l'extinction massive des dinosaures est marquée par un taux de spéciation élevé, c'est à dire un taux élevé d'apparition de nouvelles espèces qui vont se spécialiser dans les niches écologiques vacantes. Ce type d'expansion rapide est appelée **radiation adaptative**. En effet, le relâchement de la pression de prédation causée par les dinosaures a permis une augmentation des tailles de populations et donc une variabilité plus importante dans la population des espèces de rongeurs. Cette nouvelle variabilité peut alors conférer un nouvel avantage à une partie de la population qui peut ensuite se reproduire davantage entre elle qu'avec le reste de la population et, à terme, former une nouvelle espèce. Par exemple, on peut imaginer (en caricaturant un peu) une espèce de mammifères de l'époque qui était essentiellement nocturne pour éviter de croiser son lointain cousin le *Tyrannosaurus rex*. Certains individus de cette espèce ont pu commencer à s'aventurer de plus en plus le jour pour chercher de la nourriture sans risquer de se faire dévorer. Ces individus ont pu alors avoir accès à de nouvelles ressources et de fil en aiguille, ou plutôt de génération en génération, ils ont pu adopter une vie complètement diurne où ils ne croisaient plus leurs congénères nocturnes et se reproduire sélectivement entre eux. Cette différence de comportement a entraîné une division de l'espèce en deux parties, que l'on appelle **spéciation**. Puis, l'espèce diurne a pu se diviser de nouveaux en différentes espèces, chacune

spécialisée dans un certain type de nourriture, insectes, herbes, graines ...

La disparition des dinosaures a été globale, c'est à dire qu'aux quatre coins de la Pangée, les mammifères ont pu se diversifier localement et s'adapter de manière indépendante et répétée vers les mêmes niches écologiques. Sur la figure 1.6, on peut voir que dans certains cas, l'adaptation vers des niches similaires s'est traduite par l'acquisition de traits, de caractéristiques très similaires que l'on appelle phénotype convergent. La figure 1.6 présente cinq paires d'espèces montrant des cas de convergences morphologiques très frappantes. Les espèces de chacune des paires se ressemblent de manière spectaculaire, mais sont moins proches entre elles que l'ensemble des espèces appartenant au groupe des Afrotheria ou du groupe des Laurasiatheria (taupes dorées et lamantins) ou des deux laurasiathériens (la taupe européenne et le dauphin). Par exemple, les taupes dorées (Figure 1.6. a) sont plus proches phylogénétiquement des lamantins (Figure 1.6.g) que des taupes européennes (Figure 1.6.b).

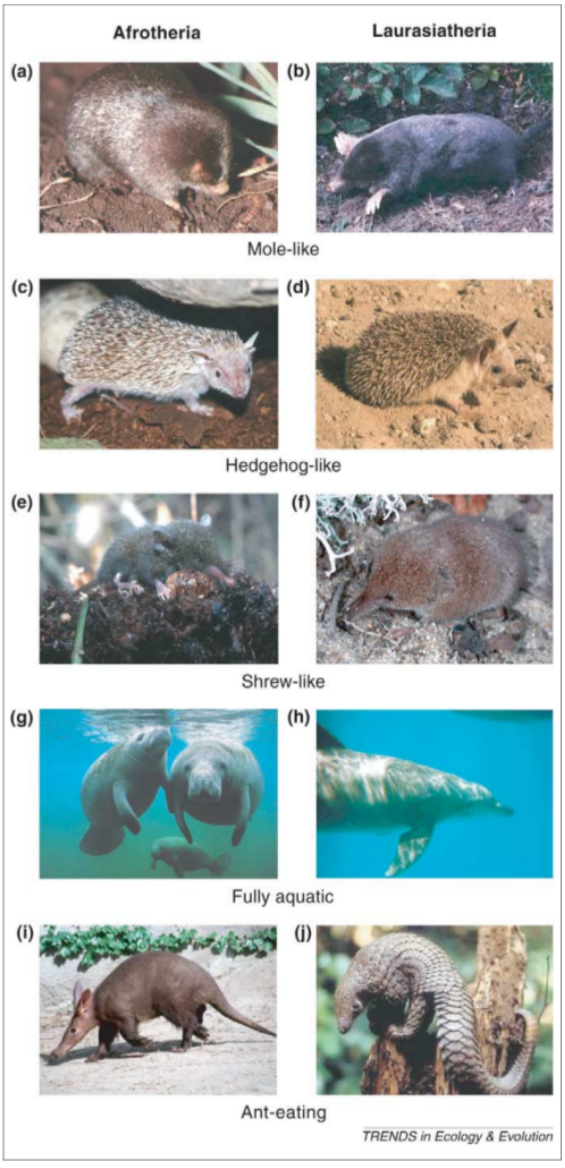

FIGURE 1.6 – Exemple de convergences morphologiques entre deux groupes de mammifères placentaires, les Afrotheria et Laurasiatheria. Les espèces (a), (c), (e), (g) et (i) sont plus proches les unes des autres que des espèces (b), (d), (f), (h) et (j) et réciproquement. (a) taupe dorée africaine (*Chrysochlorinae*), (b) taupe d'Europe (*Talpinae*), (c) hérisson malgache (*Tenrecinae*), (d) hérisson commun (*Erinaceinae*), (e) tenrec musaraigne (*Oryzorictinae*), (f) musaraigne commune (*Soricinae*), (g) lamantin (*Trichechidae*), (h) dauphin (*Delphininae*), (i) oryctérope (*Orycteropodidae*), (j) pangolin (*Maninae*). D'après la Figure 2 de (SPRINGER, 2004).

1.1.3.2 L'apport de l'ADN par rapport aux caractères morphologiques pour construire des phylogénies

Comme nous l'avons vu précédemment, les premières phylogénies étaient basées sur des caractéristiques morphologiques qui étaient considérées comme homologues. Ces caractéristiques sont rassemblées dans une matrice, appelée **matrice de similarité**, qui correspond à l'état d'une caractéristique morphologique pour chacune des espèces étudiées (Figure 1.7, haut). Puis, on reconstruit l'ensemble des scénarios possibles qui permettent d'expliquer l'acquisition de ces caractéristiques (Figure 1.7, bas). Le scénario retenu est le plus parcimonieux, c'est à dire celui qui minimise le nombre d'événements requis pour expliquer l'évolution des caractères. C'est le principe du **maximum de parcimonie**. Dans la figure 1.7, les scénarios encadrés en vert sont ceux retenus car ils impliquent le moins d'événements.

Or, si dans la matrice de similarité, on utilise des caractères non-homologues, donc analogues (caractère rouge), cela va introduire des erreurs dans la phylogénie. C'est ce que l'on peut observer entre le panel du centre et celui de gauche de la figure 1.7. Certains processus comme la convergence évolutive peuvent expliquer que le "vrai scénario" ne soit pas le plus parcimonieux. Les naturalistes de l'époque ont rassemblé suffisamment d'observations morphologiques pour ne pas complètement se laisser tromper par la convergence évolutive mais celle-ci a brouillé le signal pour reconstruire la "vraie" phylogénie (Figure 1.4.a). Une manière de se prémunir contre ce signal de convergence est d'utiliser un grand nombre de caractéristiques, de colonnes de la matrice de similarité qui supportent le "vrai" scénario (Figure 1.7, droite). Les caractères morphologiques peuvent donc se répéter ou être remplacés et ne permettent pas de garder une trace parfaite du passé.

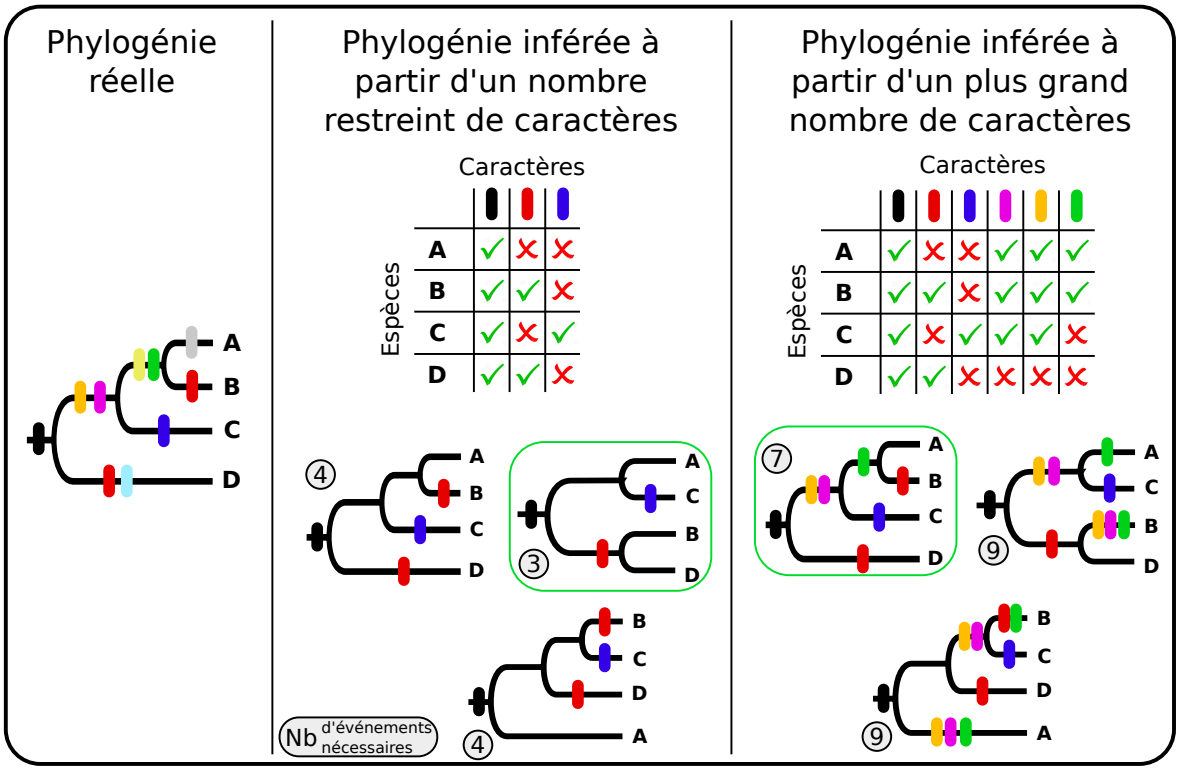

FIGURE 1.7 – Influence du nombre de caractères utilisés sur la reconstruction d'une phylogénie. La phylogénie sur laquelle est indiqué le scénario évolutif réel est représentée à gauche. Chaque trait de couleur représente un caractère différent qui a été acquis au cours de l'évolution de ces quatre espèces. Au centre et à droite sont représentées les phylogénies inférées à partir d'un nombre de caractères différents. Le nombre d'événements supportant chacune des phylogénies inférées est indiqué dans un cercle. D'après le principe de parcimonie, la phylogénie retenue (encadré vert) est celle nécessitant le moins d'événements. L'ensemble des possibilités des phylogénies inférées n'est pas représenté.

L'ADN permet également de capturer une trace évolutive, tout comme la morphologie. En effet, l'ADN est le support de l'hérédité et se transmet de génération en génération. Il contient l'histoire des individus sur de bien plus longues périodes. L'ensemble de l'ADN d'un individu est appelé **génome**. Une molécule d'ADN est une séquence de bases ou **nucléotides** composée de 4 bases différentes : A, T, G et C. Le génome humain est constitué de milliards de bases qui sont dupliquées et transmises à la descendance. Chaque position est répliquée à l'identique mais il peut se produire des erreurs, des **mutations**, qui sont très rares. L'ADN permet donc de conserver une trace de l'histoire basée sur un nombre d'événements plus important que la morphologie.

Le début de la phylogénie moléculaire, c'est à dire basée sur les molécules d'ADN, a permis d'agrandir la matrice de similarité utilisée et de minimiser l'effet de la convergence évolutive. De plus, le principe de parcimonie est abandonné et on utilise des modèles probabilistes utilisant des modèles d'évolution permettant de prendre en compte les mécanismes évolutifs.

L'utilisation de l'ADN pour faire de la phylogénie a été permise par deux avancées majeures (récompensées par des prix Nobel). La première est l'invention du séquençage par un biochimiste anglais, Frederick Sanger, en 1977. Cette méthode permet de lire et donc d'obtenir la séquence d'un fragment d'ADN d'une espèce, c'est à dire de traduire une molécule chimique en une suite de lettres constituée d'un alphabet à 4 lettres, A, T, G et C. La seconde avancée majeure est l'invention de la PCR (Polymerase Chain Reaction), en 1983 par deux scientifiques Nord-Américains (Kary Kullis et Michael Smith), qui permet d'amplifier des fragments ciblés d'ADN d'un génome.

La PCR permet de séquencer le même ou les mêmes gènes de différentes espèces et les comparer comme on peut le faire pour des caractéristiques morphologiques. La figure 1.8 présente les 96 premiers nucléotides d'un gène, le cytochrome B, qui a été l'un des premiers gènes séquencé pour différentes espèces (IRWIN et collab., 1991). Les séquences de chaque espèce sont "alignées" pour faire correspondre entre eux les sites homologues. Chacune des colonnes de cet **alignement** est appelée **site** et est considérée comme une colonne de la matrice de similarité. Il est important que les séquences soient bien alignées, sans quoi cela reviendrait à analyser des sites non-homologues. Pour faire l'analogie avec la morphologie, cela reviendrait à comparer le nombre de doigts d'un individu et la présence d'une molaire dans la mâchoire supérieure, ce qui n'a, bien sûr, aucun sens. L'utilisation de ces nouvelles informations a permis aux scientifiques de résoudre un des casses-têtes de l'évolution et de proposer la phylogénie des mammifères actuellement admise (Figure 1.4.b).

	M	T	N	I	R	K	S	H	P	L	M	K	I	V	N	N	A	F	I	D	L	P	A	P	S	N	I	S	S	W	W	N
Cow	ATG	ACT	AAC	ATT	CGA	AAG	TCC	CAC	CCA	CTA	ATA	AAA	ATT	GTA	AAC	AAT	GCA	TTC	ATC	GAC	CTT	CCA	GCC	CCA	TCA	AAC	ATT	TCA	TCA	TGA	TGA	AAT
Sheep	ATG	ATC	AAC	ATC	CGA	AAA	ACC	CAC	CCA	CTA	ATA	AAA	ATT	GTA	AAC	AAC	GCA	TTC	ATT	GAT	CTC	CCA	GCT	CCA	TCA	AAT	ATT	TCA	TCA	TGA	TGA	AAC
Goat	ATG	ACC	AAC	ATC	CGA	AAG	ACC	CAC	CCA	TTA	ATA	AAA	ATT	GTA	AAC	AAC	GCA	TTT	ATT	GAC	CTC	CCA	ACC	CCA	TCA	AAC	ATC	TCA	TCA	TGA	TGA	AAC
Pronghorn	ATG	ATC	AAC	ATC	CGA	AAA	TCC	CAC	CCA	TTA	ATA	AAA	ATT	GTA	AAC	AAC	GCA	TTC	ATT	GAC	CTC	CCA	GCC	CCA	TCA	AAC	ATC	TCA	TCT	TGA	TGA	AAC
Giraffe	ATG	ATC	AAC	ATC	CGA	AAG	TCC	CAC	CCA	CTA	ATA	AAA	ATT	GTA	AAT	AAC	GCA	CTA	ATC	GAT	CTA	CCA	GCC	CCA	TCA	AAT	ATC	TCA	TCA	TGA	TGA	AAC
Fallow	ATG	ATC	AAC	ATC	CGA	AAA	TCT	CAC	CCA	TTG	ATA	AAA	ATC	GTA	AAC	AAC	GCA	TTC	ATT	GAT	CTC	CCA	GCC	CCA	TCA	AAT	ATT	TCA	TCC	TGA	TGA	AAT
Black-tail	ATG	ACC	AAC	ATC	CGA	AAA	ACC	CAC	CCA	CTC	ATA	AAA	ATT	GTA	AAC	AAC	GCA	TTC	ATT	GAT	CTT	CCT	GCC	CCA	TCA	AAC	ATC	TCG	TCA	TGA	TGA	AAC
Chevrotain	ATG	ATC	AAT	ATC	CGA	AAA	TCA	CAC	CCA	CTA	ATA	AAA	ATT	GTC	AAC	AAT	GCA	TTT	ATT	GAC	CTC	CCA	GCC	CCA	TCA	AAC	ATC	TCT	TCA	TGG	TGA	AAC
Camel	ATG	ACA	AAC	ATC	CGA	AAA	TCA	CAC	CCA	CTT	CTA	AAA	ATT	ATA	AAC	GAC	GCA	TTC	ATT	GAC	CTT	CCA	GCC	CCC	TCC	AAT	ATT	TCA	TCA	TGA	TGA	AAC
Dolphin 1a	ATG	ACC	AAC	ATC	CGA	AAA	ACA	CAC	CCT	CTA	ATA	AAA	ATC	CTC	AAT	GAC	GCA	TTC	ATT	GAT	CTC	CCC	ACC	CCA	TCT	AAT	ATC	TCC	TCT	TGA	TGA	AAT
Dolphin 1b	ATG	ACC	AAC	ATC	CGA	AAA	ACA	CAC	CCA	CTA	ATA	AAA	ATC	CTC	AAT	GAC	GCA	TTC	ATT	GAT	CTA	CCC	ACC	CCA	TCT	AAT	ATC	TCC	TCT	TGA	TGA	AAT
Dolphin 2	ATG	ACC	AAC	ATC	CGA	AAA	ACA	CAC	CCA	CTA	ATA	AAA	ATC	CTC	AAT	GAT	GCA	TTC	ATT	GAT	CTA	CCC	ACT	CCA	TCT	AAC	ATC	TCC	TCT	TGA	TGA	AAT
Peccary	ATG	ACC	AAT	ATC	CGA	AAA	TCC	CAC	CCA	CTA	ATA	AAA	ATT	ATT	AAC	AAC	ACA	TTC	ATC	GAC	TTA	CCA	ACC	CCA	TCA	AAT	ATT	TCA	TCA	TGA	TGA	AAC
Pig	ATG	ACC	AAC	ATC	CGA	AAA	TCA	CAC	CCA	CTA	ATA	AAA	ATT	ATC	AAC	AAC	GCA	TTC	ATT	GAC	CTC	CCA	GCC	CCC	TCA	AAC	ATC	TCA	TCA	TGA	TGA	AAC
Zebra	ATG	ACA	AAC	ATC	CGA	AAA	TCC	CAC	CCG	CTA	ATT	AAA	ATC	ATC	AAT	CAT	TCT	TTC	ATC	GAC	CTA	CCA	GCC	CCC	TCA	AAC	ATC	TCA	TCA	TGA	TGA	AAC
Rhino	ATG	ACT	AAC	ATC	CGT	AAA	TCC	CAC	CCA	CTA	ATC	AAA	ATT	ATC	AAT	CAC	TCA	TTC	ATC	GAC	CTA	CCC	ACC	CCA	TCA	AAC	ATT	TCA	GCC	TGA	TGA	AAT
Elephant	ATG	ACC	GAC	ATT	CGA	AAA	TCT	CAC	CCC	TTA	CTT	AAA	ATC	ATC	AAT	AAA	TCC	TTC	ATT	GAT	CTA	CCT	ACC	CCA	TCC	AAC	ATA	TCA	ACA	TGA	TGA	AAT
Rat	ATG	ACA	AAC	ATC	CGA	AAA	TCT	CAC	CCC	CTA	TTC	AAA	ATC	ATC	AAC	CAC	TCC	TTT	ATC	GAC	CTA	CCG	GCC	CCA	TCT	AAC	ATC	TCA	TCA	TGA	TGA	AAC
Mouse	ATG	ACA	AAC	ATA	CGA	AAA	ACA	CAC	CCA	TTA	TTT	AAA	ATT	ATT	AAC	CAC	TCA	TTC	ATT	GAC	CTA	CCT	GCC	CCA	TCC	AAC	ATT	TCA	TCA	TGA	TGA	AAC
Human	ATG	ACC	CCA	ATA	CGC	AAA	ATT	AAC	CCC	CTA	ATA	AAA	TTA	ATT	AAC	CAC	TCA	TTC	ATC	GAC	CTC	CCC	ACC	CCA	TCC	AAC	ATC	TCC	GCA	TGA	TGA	AAC

FIGURE 1.8 – 32 premières positions du Cytochrome B de 20 mammifères extraites de la Figure 2 de (Irwin et al. 1991).

Résumons cette partie. On a vu, d'une part, que les caractéristiques morphologiques permettaient de retracer les relations de parentés entre espèces de manière générale mais que l'utilisation de données moléculaires (ADN) est nécessaire pour résoudre les relations de parentés plus fines. D'autre part, l'adaptation à un même environnement ou l'acquisition d'une même fonction peut se faire de manière similaire dans des lignées différentes. Mais on peut se demander pourquoi est ce important d'identifier et d'étudier les cas de convergence ?

1.1.4 Que peut nous apporter l'étude des cas de convergence ?

1.1.4.1 Comprendre la place de l'Homme dans le monde du vivant

La première raison est peut-être la curiosité humaine, le besoin de comprendre la Nature, d'approfondir ses connaissances pour en connaître les moindres recoins. En Biologie et plus particulièrement dans l'étude de la convergence évolutive, les grandes questions qui se posent concernent la répétabilité de l'Évolution et donc sa prédictibilité. Si l'on répète l'Évolution à partir d'un certain point, est-ce que le résultat serait le même ? Et, *in fine*, est-ce que l'Homme évoluerait à nouveau de la même manière ?

Il existe deux grandes théories à ce sujet. La première est que l'Évolution est liée aux événements successifs complètement aléatoires (chutes de météorites, éruptions volcaniques, etc.) qui viennent perturber l'Évolution des êtres vivants. Si on changeait l'ordre des événements passés, alors le résultat de l'Évolution des espèces ne serait pas le même que le résultat actuel. Dans ce contexte différent, la sélection naturelle aura peut-être retenu des adaptations différentes. C'est ce qui a été nommé la **contingence**. D'après cette théorie, le résultat de l'Évolution est par définition imprévisible. Les cas de convergence que l'on observe ne sont que le fruit du hasard, ce qui sous entend que la convergence n'est qu'un processus mineur. Dans la seconde théorie, la convergence est un processus majeur de l'Évolution. La multitude d'exemples de cas de convergence montre que l'évolution est prédictible et contrainte vers une direction donnée. Cela signifie que si une catastrophe naturelle venait à perturber l'équilibre des êtres vivants et engendrer une extinction massive, l'évolution des espèces qui auront survécu va retenir des adaptations qui avait déjà été retenues par des espèces éteintes mais sans lien de parenté. Les questions qui pourraient résumer ce débat sont alors : quelle est la place de la convergence dans l'Évolution ? Est-elle un processus majeur ou subsidiaire ?

Ce débat a été initié par deux scientifiques à la fin du XXème siècle, Stephen Jay Gould et Simon Conway Morris. L'ironie du sort est que ces deux scientifiques, paléontologues et collègues, ont formulé ces deux théories contradictoires en se reposant sur les mêmes observations, celles de Simon Conway Morris qui a analysé à nouveau des fossiles découverts en 1909 et qui en a revu la classification. Ces fossiles proviennent d'une carrière de schistes au Canada, appelée Schistes de Burgess. Ces schistes datent du Cambrien, c'est à dire il y a environ 540 millions d'années, donc si l'on se réfère à la Figure 1.5, bien avant l'apparition des mammifères. Le grand nombre de fossiles retrouvés, ainsi que leur très bon état de fossilisation, ont permis de faire des reconstructions complètes de la faune et de la flore de cette époque (Figure 1.9).

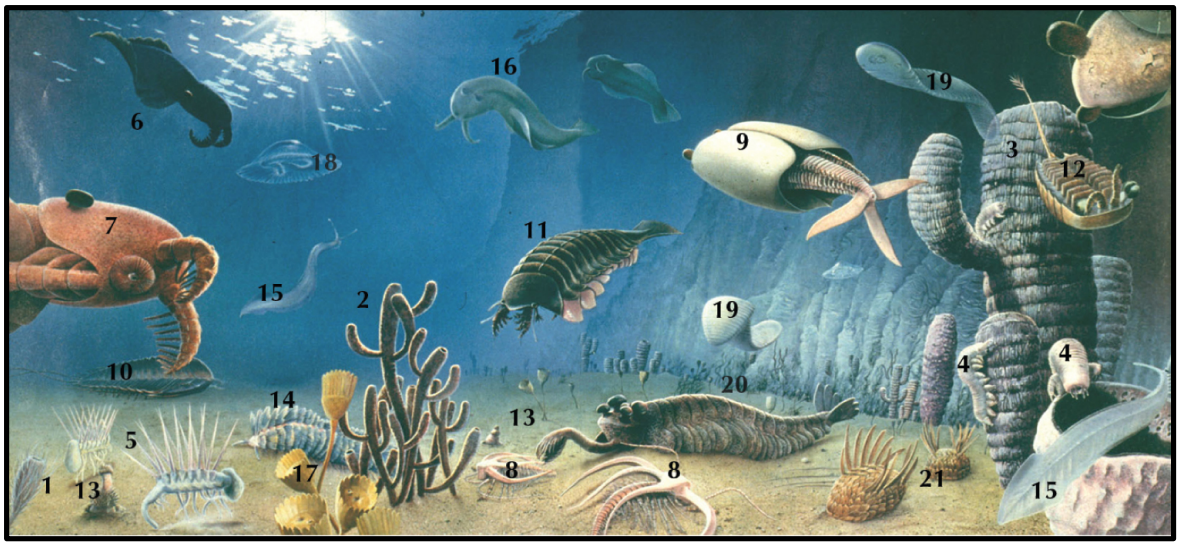

FIGURE 1.9 – Reconstruction de la faune de Burgess à partir des fossiles trouvés. Eponges : Pirania (1), Vauxia (2), Wapkia (3), Lobopodia : Aysheaia (4), Hallucigenia (5), Anomalocaris(6), Laggania (7), Marella (8), Odaria (9), Trilobites : Olenoides (10), Sanctacaris (11), Sarotrocercus (12), Priapuliens : Ottoia (13), Polychètes (annélides) : Canadia (14), Chordés : Pikaia (15), Animaux dont les relations de parentés ne sont pas résolues : Amiskwia (16), Dinomischus (17), Eldonia (18), Odontogriphus (19), opabinia (20), Wiwaxia (21). Source : http://evolution-textbook.org.

Stephen Jay Gould fut le premier à exposer sa théorie dans un livre nommé *Wonderful Life : The Burgess Shale and the Nature of History* (GOULD, 1990) [1] où il interprète les observations provenant du travail de son collègue. Pour lui, la divergence entre les formes d'êtres vivants actuels et celles de cette époque montre que l'Évolution est basée sur le hasard. En effet, si l'on revenait à cette époque et qu'on laissait à nouveau l'histoire se dérouler, la probabilité de revoir apparaître les humains est infime.

En réponse à ce livre de Gould, Simon Conway Morris publia quelques années plus tard un livre basé aussi sur les schistes de Burgess, *The Crucible of Creation : The Burgess Shale and the Rise of Animals* (MORRIS, 1999) où il s'oppose à la vision de Gould sur l'imprévisibilité de l'Évolution. Pour lui, les fossiles qu'il a observés sont apparentés à des formes existantes actuelles. C'est pour cela qu'il suggère que la convergence est la preuve que l'évolution est bornée et donc restreinte et que toutes les possibilités de formes vivantes ne sont pas possibles. Il suggère enfin qu'une force supérieure laissée à la discrétion du lecteur est à l'origine de ces contraintes.

Je pense que ces deux théories bien que contradictoires peuvent être compatibles. Pour moi, la partie aléatoire de l'Évolution est indéniable. En effet, des événements ponctuels tels que la chute d'une météorite ou une éruption volcanique sont imprévisibles et ont des conséquences importantes sur le monde vivant, sur la **biosphère**. Ces événements vont provoquer d'importantes perturbations qui vont rebattre l'ordre ponctuel de la biosphère à leurs échelles. Par exemple, une chute de météorite qui provoque une extinction majeure va permettre aux espèces qui ont survécu, par chance, de récupérer la place des espèces éteintes comme nous l'avons vu précédemment avec les mammifères. J'insiste sur le caractère aléatoire car une espèce ne peut pas s'adapter à des événements rares et imprévisibles. Une espèce peut avoir adopté une vie nocturne car cela lui permettait d'échapper à des prédateurs diurnes mais en aucun cas pour survivre à la chute d'une météorite 10 millions d'années plus tard.

Par contre, le fait que l'Évolution réutilise plusieurs fois les mêmes solutions dans différentes situations est intriguant et interroge. Parfois cela peut s'expliquer par des lois physiques. Par exemple, la convergence morphologique entre les espèces marines tels que le dauphin ou le requin

1. *La vie est belle : les surprises de l'évolution*, en francais

qui ont une forme profilée qui leur permet de nager rapidement, peut s'expliquer par les lois de la physique. Il n'y a pas des centaines de manières pour qu'un corps puisse se déplacer rapidement dans l'eau. Il doit limiter les frottements et donc avoir une forme profilée. Mais d'autres fois, il est plus compliqué de comprendre le mécanisme ou la loi qui peut se cacher derrière une convergence phénotypique. Si l'on reprend l'exemple des plantes qui synthétisent de la caféine, il a été montré que cette molécule avait différents effets bénéfiques pour ces plantes. Pour certaines telles que le caféier, la caféine est un insecticide permettant de protéger les graines. Pour d'autres telles que l'oranger, la caféine augmente les capacités de mémoire des pollinisateurs et notamment les abeilles (WRIGHT et collab., 2013). Comment expliquer que ces plantes ont eu recours à la même molécule ? Est-ce que ces molécules sont faciles à fabriquer ? Est-ce que d'autres molécules pourraient avoir les mêmes actions ? Si oui, est-ce qu'on les retrouve dans d'autres plantes ? Nous reviendrons plus en détail sur ces concepts dans une prochaine partie lorsque nous aborderons la convergence dans un contexte moléculaire.

Si l'on résume cette sous-partie, le but d'étudier les cas de convergences est de comprendre si les patterns répétés que l'on observe sont dûs au hasard ou s'il existe une contrainte qui l'explique. Ainsi, nous pourrons apporter des réponses sur la répétabilité de l'Évolution et donc sa prévisibilité.

1.1.4.2 S'inspirer de la nature pour répondre aux défis actuels et de demain

L'étude de la convergence évolutive peut aussi avoir une fin plus pragmatique. En effet, la biologie peut inspirer des innovations techniques, on parle de **biomimétisme**. Si un phénotype est utilisé de manière répétée pour s'adapter à une niche écologique, il y a de fortes chances que ce soit une solution efficace pour répondre à la contrainte de son environnement. L'étude de ce phénotype convergent et des mécanismes sous jacents peut ensuite être à l'origine d'une innovation technique. On pourrait dire cela pour toutes les adaptations que l'on peut observer dans la nature, mais si un mécanisme est utilisé de manière répétée, il paraît plausible qu'il n'existe pas de nombreuses autres façons de mieux le faire. En tous cas, il aura déjà été testé plusieurs fois. Par exemple, le radar et le sonar sont inspirés du système d'écholocation des chauves-souris et des dauphins. On peut aussi citer un autre exemple venant du monde végétal avec le velcro (pour velours et crochets) ou scratch. Ce système d'accroche composé de deux bandes, une avec de petits crochets et l'autre avec des boucles, est inspiré du fruit de bardane ou plus largement de toutes les graines possédant de petits crochets pour s'agripper aux poils des animaux et ainsi améliorer leur dispersion.

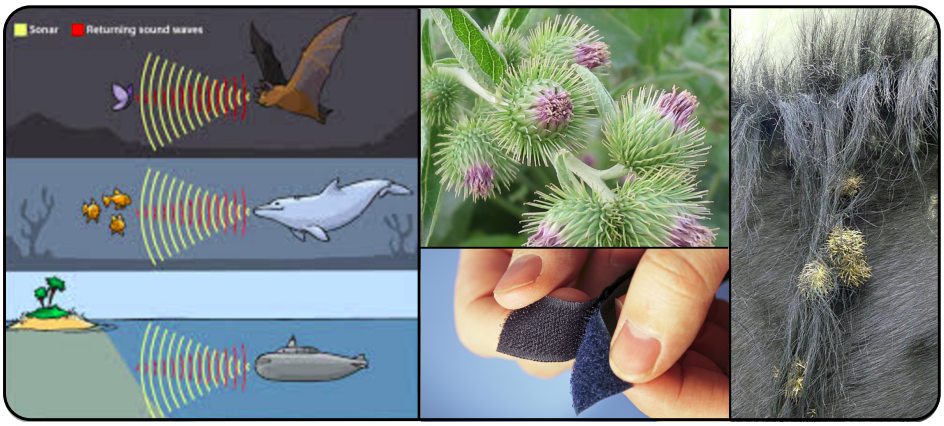

FIGURE 1.10 – Cas de biomimétisme basés sur des cas de convergence. A gauche, le sonar basé sur l'écholocation chez les dauphins et les chauves-souris et à droite le velcro, basé sur les fruits de bardane qui possèdent des petits crochets pour s'accrocher aux poils des animaux pour faciliter la dispersion des graines. Credit : Wikipedia +

Enfin, comme la Nature est remplie d'exemples de macro et micro-convergences il faut être capable de les détecter et ensuite les comprendre pour pouvoir estimer leur potentiel à être réutilisés.

Pour conclure cette première partie, nous avons vu que la convergence est un phénomène répandu dans l'Évolution. Je n'ai fait que vous citer quelques exemples mais il en existe une multitude d'autres et tout autant à découvrir. L'étude de ces cas de convergence est une source précieuse d'informations. Elle peut nous apporter des réponses sur la prévisibilité de l'Évolution mais c'est aussi un catalogue de solutions techniques déjà testées et approuvées. La prochaine étape est de savoir ce qui se cache sous ces cas de convergences au niveau biologique. Pour cela nous avons besoin de regarder à des échelles inférieures pour connaître les causes de ces convergences phénotypiques afin de mieux les comprendre.

1.2 Quelles sont les bases sous-jacentes à la convergence chez les êtres vivants et sont-elles, elles-mêmes, convergentes ?

Nous avons vu qu'il existait de nombreux cas de convergences phénotypiques. L'un des exemples que nous avons déjà abordés est l'adaptation convergente au milieu marin de trois groupes de mammifères. Dans cet exemple, la capacité de nager a été permise en partie par l'acquisition de nageoires. Mais est-ce que les bases sous jacentes de ces convergences phénotypiques sont elles-mêmes convergentes ou est-ce que les convergences phénotypiques ne sont que des similarités de surface qui proviennent de bases morphologiques complètement différentes ? Par exemple, est-ce que ces nageoires sont toutes une transformation des membres antérieurs ? Si c'est le cas, est-ce que les organisations internes (squelette, muscles) de ces nageoires se ressemblent et sont donc elles-mêmes convergentes ou sont-elles différentes ?

Nous avons abordé jusqu'à maintenant la convergence phénotypique mais nous n'avons pas encore défini précisément ce que signifie phénotype. Le terme **phénotype** peut faire référence à une caractéristique observable d'une espèce, mais si l'on parle du phénotype de l'espèce, c'est à l'ensemble des caractéristiques observables de cette espèce que l'on fait référence. Le terme phénotype est en fait un concept global qui peut aussi être appliqué à différentes échelles. La présence d'une nageoire, le squelette qui supporte la nageoire, la forme des os ou encore la disposition des muscles sont des phénotypes. Le point commun est que ce sont des caractéristiques observables. Par contre, la manière dont la nageoire se forme, l'os grandit ou les muscles s'accrochent est liée au développement, lui-même encodé dans l'ensemble du génome de l'individu. On parle du **génotype**. C'est pour cela que l'on dit que le phénotype est la réalisation du génotype en interaction avec son environnement.

1.2.1 La convergence fonctionnelle peut reposer sur ...

Pour survivre, les individus doivent être capables d'accomplir certaines fonctions comme se déplacer, se nourrir ou encore ressentir son environnement. Certaines espèces ont développé des stratégies convergentes et nous allons en étudier quelques unes pour comprendre les bases phénotypiques sous-jacentes à ces convergences fonctionnelles.

1.2.1.1 ... une convergence phénotypique macroscopique

1.2.1.1.1 Des nageoires pour se déplacer dans l'eau

Les mammifères marins ne sont pas les seuls tétrapodes [1] qui ont suivi la même transition du milieu terrestre au milieu marin (Figure 1.11, gauche). Il y a également les tortues, les pingouins ou même, parmi les tétrapodes aujourd'hui éteints, les ichthyosaures et les mosasaures. L'évolution de tous ces animaux a suivi un chemin convergent où leurs membres antérieurs ont été réutilisés comme nageoires (Figure 1.11, droite). On appelle ce type de réutilisation d'un organe pour une nouvelle fonction, une **exaptation**.

Dans le cas des tétrapodes marins, la convergence fonctionnelle, nager, est basée sur des bases morphologiques convergentes à différents niveaux, au niveau macroscopique avec les nageoires mais aussi à l'échelle inférieure avec le squelette plus ou moins marqué avec un allongement des doigts.

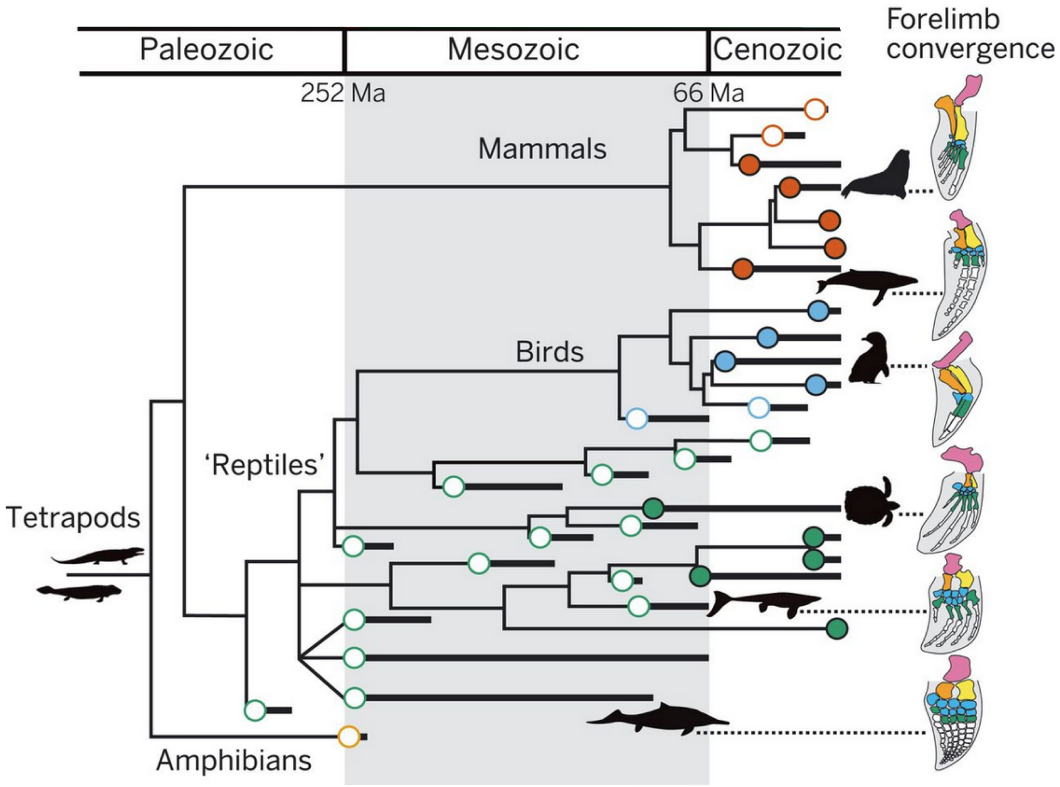

FIGURE 1.11 – Vue d'ensemble de l'évolution des tétrapodes marins (Gauche) et de l'évolution convergente du squelette des membres antérieurs (Droite). Les cercles indiquent les transitions entre le milieu terrestre et le milieu marin. Les espèces actuelles sont indiquées par des cercles pleins et les espèces éteintes par des cercles vides. Les cercles rouges correspondent aux mammifères, les bleus aux oiseaux, les verts aux reptiles excluant les oiseaux et le jaune aux amphibiens. Ma, millions d'années. D'après (KELLEY et PYENSON, 2015).

1. Animaux à 4 membres

1.2.1.1.2 Une bouche capable de filtrer de grandes quantités d'eau pour se nourrir

Les grands animaux marins, tels que les baleines ou les requins baleines, ont développé une stratégie convergente pour se nourrir. Ils sont capables de filtrer de grandes quantités d'eau à l'aide de leur bouche dans le but de retenir le plancton et ainsi se nourrir. Cependant, ils le font de manière différente. Pour la baleine, il s'agit d'utiliser ses fanons et pour le requin-baleine ses branchies (Figure 1.12, A et B). De plus, cette convergence fonctionnelle a eu lieu de manière répétée au cours du temps. Des poissons de l'ordre des Pachycormiformes, aujourd'hui éteints, étaient aussi planctivores et avaient développé cette même stratégie de filtrage d'eau (Figure 1.12, C).

La convergence fonctionnelle présentée dans ce paragraphe repose sur des similarités macroscopiques, à savoir l'utilisation de structures capables de filtrer l'eau (fanon ou branchie). Cependant, ces structures ne sont pas elles-mêmes convergentes.

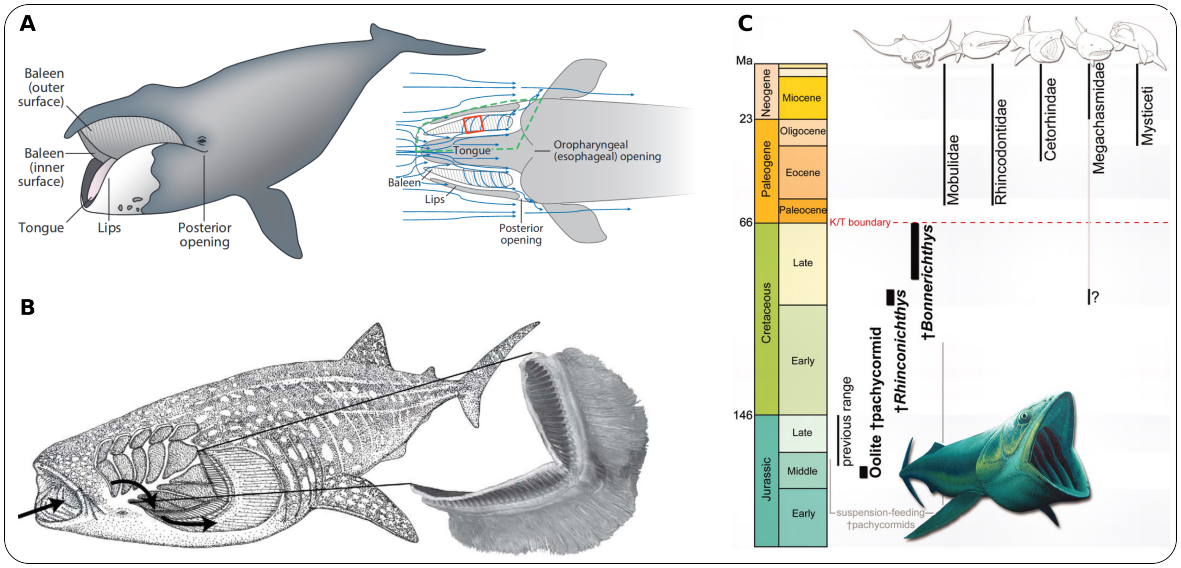

FIGURE 1.12 – Le système de filtrage de l'eau est apparu de manière répétée chez les planctivores au cours du temps. Comparaison entre le système de filtrage des baleines basé sur leurs fanons (baleen) et une différence de pression réalisée à l'aide de leur langue (tongue) (A) et des requins baleines, utilisant leurs branchies pour retenir le plancton (B) . (C) Distribution stratigraphique des espèces de grandes tailles planctivores modernes et éteintes de l'ordre des Pachycormiformes. L'échelle n'est pas respectée dans les différentes représentations. D'après (A) (GOLDBOGEN et collab., 2017), (B) (MOTTA et collab., 2010) et (C) (FRIEDMAN et collab., 2010).

1.2.1.2 … une convergence phénotypique cellulaire

1.2.1.2.1 Des yeux pour voir

Les yeux sont des organes essentiels à la survie des animaux dans leur environnement. Les premiers yeux sont apparus chez les animaux avant que les premiers animaux ne sortent de l'eau. Il existe un livre qui développe ce sujet d'un point de vue évolutif (Animal eyes, (LAND et NILSSON, 2012)). La ressemblance entre les yeux des pieuvres et des poissons pourrait laisser croire qu'il s'agit d'un organe hérité d'un ancêtre commun mais il s'agit, en fait, d'une convergence. L'œil est apparu de manière répétée dans l'évolution des animaux marins.

Si l'on observe la structure des yeux entre les céphalopodes (pieuvres, seiches, encornets, nautilus…) et les vertébrés marins, i.e. les poissons, on peut voir une structure convergente. Les deux types d'œil sont des cavités qui ont à leur entrée une lentille pour concentrer les rayons de lumière et qui sont tapissées au fond par des cellules photoréceptrices qui transmettent le signal à des cellules nerveuses (Figure 1.13, haut). Ces cellules nerveuses se ramifient pour former le nerf

optique. La structure globale est donc convergente à l'exception de l'inversion entre la position des cellules photoréceptrices et des cellules nerveuses. Dans le cas des vertébrés, les cellules photosensibles qui forment la rétine sont en dessous des cellules nerveuses, ce qui produit une zone sans récepteur à l'endroit où elles traversent la rétine, appelée tache aveugle ou tache de Mariotte. Cette tache n'est pas présente chez les céphalopodes car les cellules nerveuses se ramifient en nerf optique à l'arrière de la rétine.

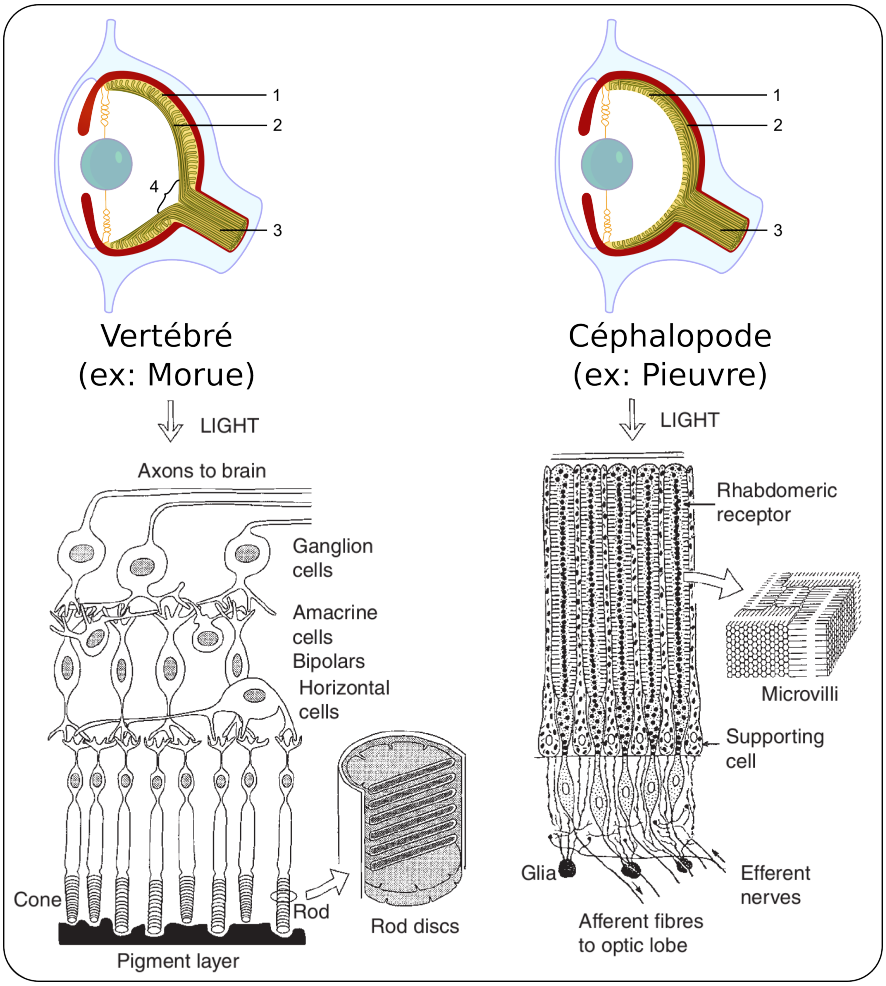

FIGURE 1.13 – Comparaison macroscopique (haut) et cellulaire (bas) entre un œil de vertébré (gauche) et un œil de céphalopode (droite). La structure macroscopique montre une convergence de structure, (1) la rétine, (2) les cellules nerveuses, (3) le nerf optique, (4) la tache aveugle. La structure cellulaire est différente entre les céphalopodes et les vertébrés. Adaptés de (haut) Wikipédia commons et (bas) de la Figure 4.8 (LAND et NILSSON, 2012).

Les structures macroscopiques ainsi que les fonctions et la structure cellulaires sont donc globalement convergentes mais si l'on descend à l'échelle inférieure (Figure 1.13, bas), on peut s'apercevoir que la fonction cellulaire est réalisée par des composants cellulaires différents. Chez les céphalopodes, les récepteurs sont rhabdomériques, c'est à dire qu'ils sont composés de microvilli contenant des pigments photorécepteurs. L'information est traitée au niveau du lobe optique à l'arrière de l'oeil (non présent sur la figure) plutôt que par les cellules nerveuses présentes à l'arrière de la rétine. Chez les vertébrés, la rétine est composée de cônes (cone) et bâtonnets (rods), eux-mêmes composés de disques contenant les pigments photorécepteurs. Au dessus de la rétine, il y a deux couches horizontales de neurones (les cellules ganglionnaires et bipolaires) connectés par les cellules amacrines qui sont en charge d'intégrer latéralement les signaux (LAND et NILSSON, 2012).

Dans cet exemple, nous avons vu que la convergence fonctionnelle peut avoir des bases macro-

scopiques communes, ici la structure de l'oeil. Cette convergence phénotypique macroscopique repose elle-même sur une fonction cellulaire convergente, des pigments photorécepteurs et des cellules nerveuses traitant le signal. Par contre, les bases moléculaires sont différentes. En effet, les pigments photorécepteurs ne sont pas les mêmes.

1.2.1.3 ... une convergence phénotypique moléculaire

1.2.1.3.1 Une molécule insecticide pour se protéger

La convergence fonctionnelle peut parfois avoir des bases moléculaires également convergentes. C'est le cas avec la caféine présentée en début d'introduction. Il a été proposé que la caféine soit à la fois un insecticide naturel (HEWAVITHARANAGE et collab., 1999; HOLLINGSWORTH et collab., 2002; MATHAVAN et collab., 1985; NATHANSON, 1984) et qu'elle ait des propriétés anti-germinatives pour inhiber la germination des autres graines du sol (SUZUKI et WALLER, 1987; WALLER, 1989). Les végétaux produisant cette molécule dans les feuilles et les graines sont ainsi protégés des prédateurs et augmentent leur compétitivité.

Cependant, la caféine peut avoir une autre fonction. Tout comme chez les êtres humains, la caféine peut avoir un effet sur le système cognitif des invertébrés (COUVILLON et collab., 2015; WRIGHT et collab., 2013). Les auteurs ont montré que la présence de caféine dans le nectar des fleurs pouvait tromper les abeilles en les faisant surestimer la qualité du nectar et ainsi en les persuadant de revenir sur la plante malgré des qualités de nectar faibles. La plante s'assure ainsi une meilleure pollinisation au détriment des abeilles ou autres pollinisateurs (COUVILLON et collab., 2015). La différence d'effet de la caféine sur les insectes (neurostimulateur et insecticide) semble dépendre de la concentration de la caféine entre les différents organes de la plante qui sont en interaction avec l'insecte.

Dans cet exemple nous avons vu qu'une convergence fonctionnelle, se protéger des insectes, avait une base moléculaire convergente avec la biosynthèse de caféine.

1.2.1.4 ... des convergences phénotypiques plus ou moins profondes

Au travers de ces différents exemples, nous avons vu que les convergences fonctionnelles peuvent être basées sur des convergences phénotypiques plus ou moins profondes. Par exemple, une fonction convergente peut être réalisée à l'aide de structures macroscopiques elles-mêmes convergentes, qui reposent sur des types cellulaires convergents, qui réalisent eux-mêmes leurs fonctions cellulaires à l'aide de molécules identiques. A contrario, une fonction convergente peut ne pas avoir une base macroscopique commune.

Comme les niveaux phénotypiques sont imbriqués les uns dans les autres, on peut donc schématiser la convergence tel un curseur qui descend plus ou moins profondément dans les différents niveaux phénotypiques (Figure 1.14). En fonction des cas étudiés, le curseur peut être placé de manière plus ou moins profonde. Ici, le phénotype est découpé en trois niveaux pour des soucis de simplification mais il pourrait être coupé en de nombreux niveaux intermédiaires si on voulait être exact.

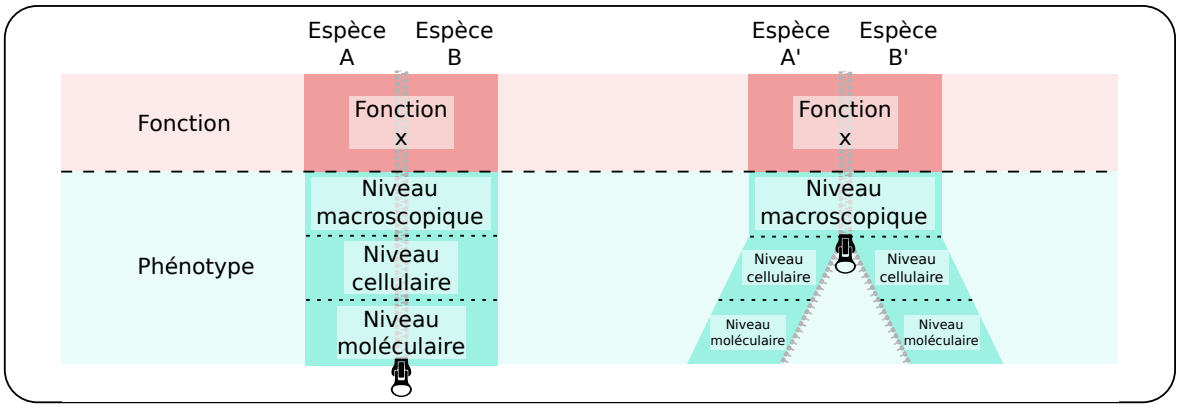

FIGURE 1.14 – La convergence peut être représentée tel un curseur qui descend le long des différents niveaux phénotypiques.

1.2.2 Le phénotype repose sur des bases génétiques : le génotype

Pour résumer ces différents niveaux d'interactions, les espèces sont soumises à différentes pressions de sélection dans leur environnement. Par sélection naturelle, les espèces qui vont survivre sont celles dont les individus auront acquis au fil des générations un ou des avantages qui leur permettent de mieux remplir une fonction que les individus des autres espèces qui sont en compétition avec elles. Cette fonction est remplie par une caractéristique acquise par l'espèce, un phénotype lui-même résultant de son génotype en interaction avec son environnement. La sélection naturelle agit donc sur la fonction, mais c'est le génotype, au travers du phénotype, qui permet sa réalisation. Ces trois niveaux d'organisation (fonction, phénotype et génotype) sont imbriqués.

Pour étudier les bases sous-jacentes de la convergence, il faut donc étudier le niveau inférieur au phénotype, le génotype (Figure 1.15). On vient de montrer que la convergence fonctionnelle pouvait reposer sur des bases phénotypiques elles-même convergentes, est-ce le cas au niveau génotypique ?

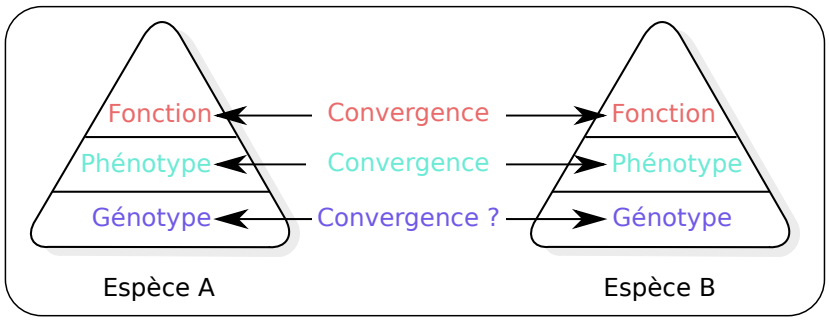

FIGURE 1.15 – La convergence évolutive repose-t-elle aussi sur des bases génétiques convergentes ?

1.2.3 La convergence phénotypique peut reposer sur . . .

Nous avons vu des exemples de convergences fonctionnelles expliquées par des convergences sous-jacentes morphologiques et où, dans certains cas, la convergence était très poussée. Or, la morphologie d'un individu est le résultat du développement de l'individu qui est lui-même codé par la génétique. Les bases génétiques de convergences phénotypiques sont souvent méconnues à cause de processus complexes que nous verrons par la suite. Cependant, depuis deux décennies, de nombreux auteurs ont documenté des exemples qui révèlent les bases génétiques de convergences phénotypiques.

1.2.3.1 … des groupes de gènes convergents

1.2.3.1.1 La modification d'un réseau de gènes à l'origine de couleurs convergentes

Les *Iochroma* sont des arbustes de la famille des Solanacées. Ils se distinguent par leurs fleurs tubulaires. La plupart des fleurs des arbustes de ce genre sont de couleur bleue ou violette (*Iochroma calycinum*). Cependant, il existe des espèces de ce genre ayant des fleurs oranges (*Iochroma edule*), blanches (*Iochroma loxense*) ou rouges (*Iochroma gesnerioides*) (Figure 1.16.A). L'apparition de ces couleurs s'est produite de manière indépendante et donc convergente au travers de l'arbre phylogénétique de ce genre (Figure 1.16.A). Des pigments, les anthocyanes, sont responsables de ces couleurs. Ils sont issus d'une même voie de biosynthèse régulée par des enzymes. (Figure 1.16.B). LARTER et collab. (2018) ont montré que l'apparition répétée de ces couleurs est permise par des modifications répétées de l'expression des gènes codants pour ces enzymes.

Dans leur étude, les auteurs ont mesuré l'expression des gènes codants pour les enzymes impliquées dans la biosynthèse de ces pigments et ils ont montré que leur profil d'expression explique la différence de synthèse des pigments de différentes couleurs et donc la couleur des fleurs. Dans le cas des espèces avec des fleurs blanches, on peut voir que les gènes *F3'H*, *F3'5'h*, *Dfr*, Ans sont sous-exprimés par rapport aux espèces à fleurs bleues. Cela diminue la quantité d'enzymes capables de réaliser les réactions menant aux anthocyanines et donc de supprimer les pigments des fleurs (Figure 1.16, C).

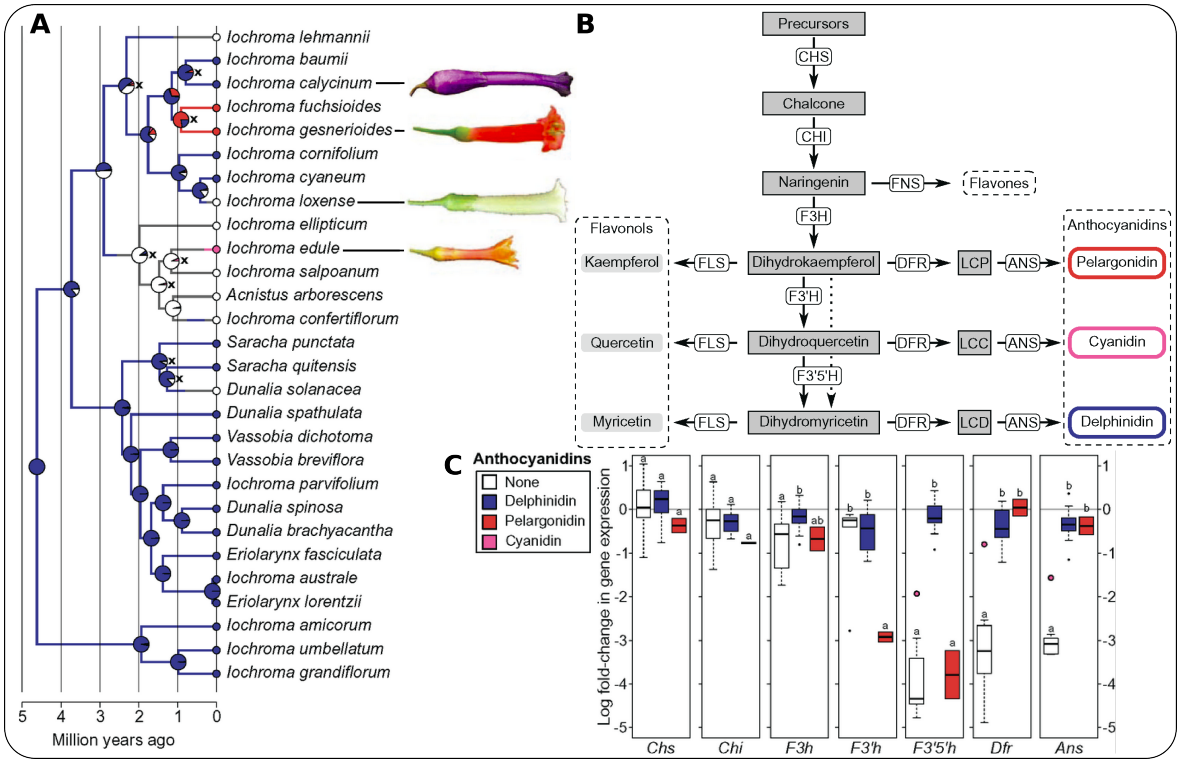

FIGURE 1.16 – Illustration de la convergence de la coloration des fleurs dans le genre *Iochroma*. (A) Reconstruction des caractères ancestraux des colorations des fleurs (LARTER et collab., 2018) et illustration provenant de (SMITH et collab., 2008). (B) Vue simplifiée de la biosynthèse des anthocyanines et des flavonoïdes (LARTER et collab., 2018), les composés chimiques sont inscrits dans des cadres colorés ou sur fond gris. Le nom des enzymes permettant de passer d'un réactif à un autre dans un cadre sur fond blanc. (C) Niveau d'expression des gènes codant pour les enzymes de la biosynthèse des anthocyanines (LARTER et collab., 2018).

On aurait pu imaginer que les différentes espèces à fleurs blanches utilisent des stratégies différentes en baissant le niveau d'expression ou supprimant l'expression de l'un des gènes de cette voie de biosynthèse mais ce n'est pas le cas. La convergence pigmentaire s'explique par la modification de manière indépendante de l'ensemble du réseau de gènes responsables de la

synthèse des pigments. Ce même mécanisme a été montré chez les ancolies (*Aquilegia*) (SMITH et RAUSHER, 2011; WHITTALL et collab., 2006). Cependant, il est également possible que ce soit un gène régulateur en amont de ce réseau de gènes qui ait été modifié et qui influe sur l'ensemble du réseau plutôt que chacun des gènes en aval aient été modifiés.

Cet exemple montre que les bases génétiques d'une convergence phénotypique, ici le changement de couleur des fleurs, peuvent être dues à la modification de l'ensemble d'un réseau de gènes.

1.2.3.1.2 La production de caféine permise par la cooptation d'enzymes

Les premières études qui se sont focalisées sur les bases génétiques de la synthèse convergente de la caféine ont été faites chez le caféier et le théier. Puis en 2016, une étude a élargi l'échantillonnage aux citronnier, cacaoyer et guarana (HUANG et collab., 2016).

Avant de rentrer dans les détails de la base génétique de cette convergence, nous avons besoin de quelques connaissances sur la synthèse de la caféine. Cette molécule est synthétisée par la suite de trois réactions enzymatiques à partir de deux précurseurs possibles : la Xanthine (X) et la Xanthosine (XR) (Figure 1.17, A). Les études antérieures à 2016 ont montré qu'un seul chemin sur les douze chemins possibles était utilisé.

HUANG et collab. (2016) ont montré que d'autres chemins étaient possibles (Figure 18, B) mais que dans chacune de ces espèces, les trois réactions enzymatiques nécessaires à la synthèse de la caféine étaient réalisées par des enzymes codées par des gènes issus de deux familles : les CS (caffeine synthase) et les XMT (xanthine méthyl-transférase). Les cinq espèces étudiées peuvent être regroupées en deux groupes : celles dont les enzymes proviennent de la famille des CS (le cacaoyer, le guarana et le camélia) et celles de la famille des XMT (le citronnier et le caféier) (HUANG et collab., 2016). Il est important de noter que la distance phylogénétique n'explique pas ce regroupement. En effet, le guarana et le citronnier sont des espèces de la famille des Sapindales et utilisent des enzymes appartenant aux deux familles d'enzymes.

La voie de synthèse de la caféine entre ces différentes espèces n'est donc pas convergente mais l'origine de la mise en place de la voie de synthèse de la caféine chez ces différentes espèces l'est. En effet, dans chacune de ces espèces, les enzymes responsables des trois réactions sont issues de duplications intra-espèces (Figure 1.18). Par exemple, bien que les enzymes TcCS1 et PcCS1 aient la même fonction, TcCS1 est plus proche de TcCS2 que de PcCS1. C'est à dire que dans chacune de ces espèces, une enzyme ancestrale ayant une fonction différente a été **cooptée** pour réaliser une nouvelle fonction.

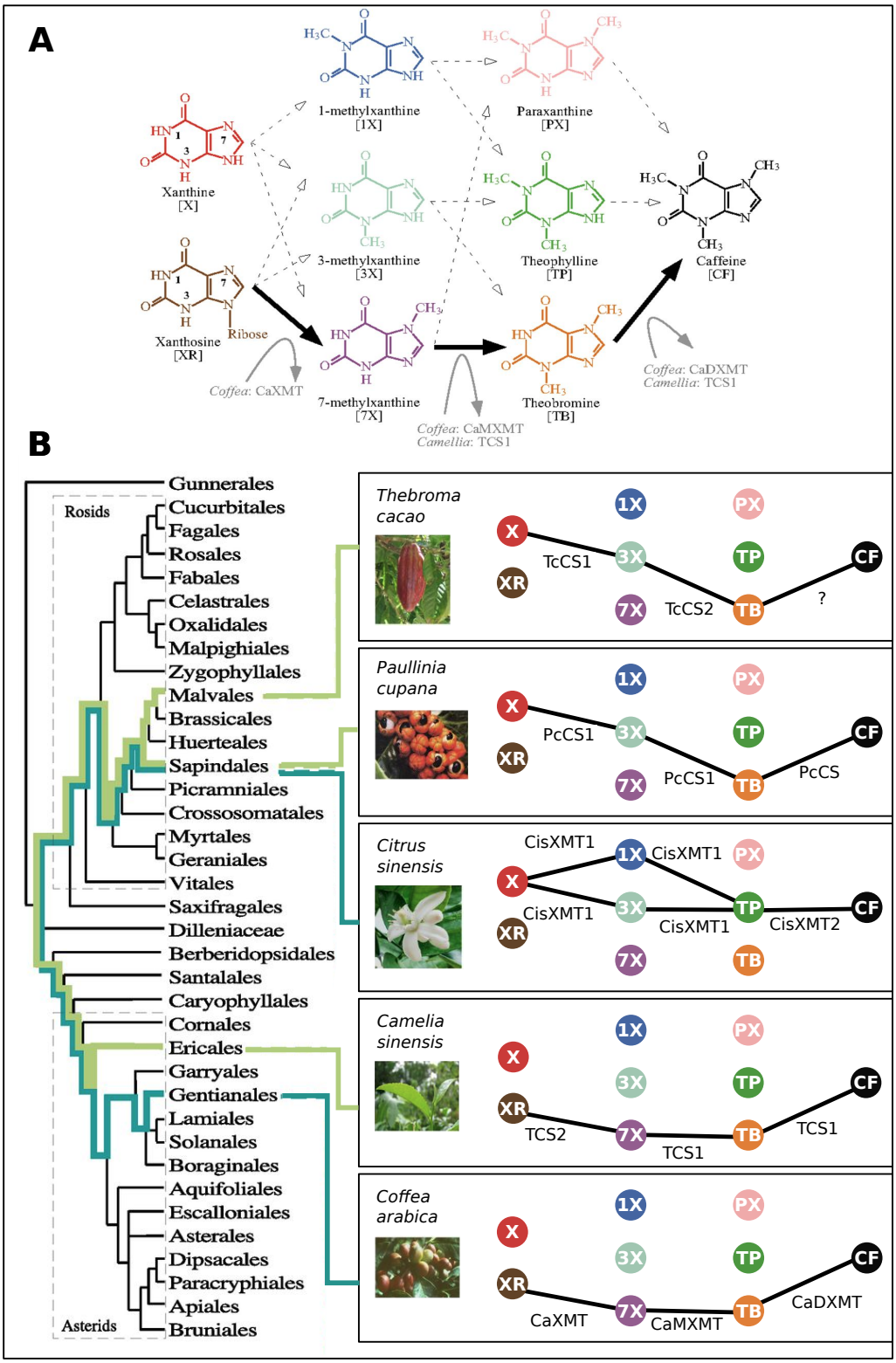

FIGURE 1.17 – Illustration de la convergence génétique de la biosynthèse de la caféine chez cinq espèces végétales. (A). La biosynthèse de la caféine a douze chemins potentiels (flèches en pointillés) dont un seul mis en évidence chez le caféier et le camélia (flèches noires) avant l'étude de (HUANG et collab., 2016). CF, caféine ; PX, paraxanthine ; TB, théobromine ; TP, theophylline ; X, xanthine ; 1X, 1-methylxanthine ; 3X, 3-methylxanthine ; 7X, 7-methylxanthine ; XR, xanthosine. (B) La biosynthèse de la caféine est apparue au cours de l'évolution dans cinq plantes utilisant des gènes et des chemins différents. Les lignes vertes foncées indiquent les plantes ayant recruté indépendamment un gène de la famille des enzymes du type XMT et en vert clair les enzymes du type CS. Les points d'interrogation représentent les enzymes non caractérisées. (A) Figure 1 et (B) adaptée de la Figure 2 de (HUANG et collab., 2016).

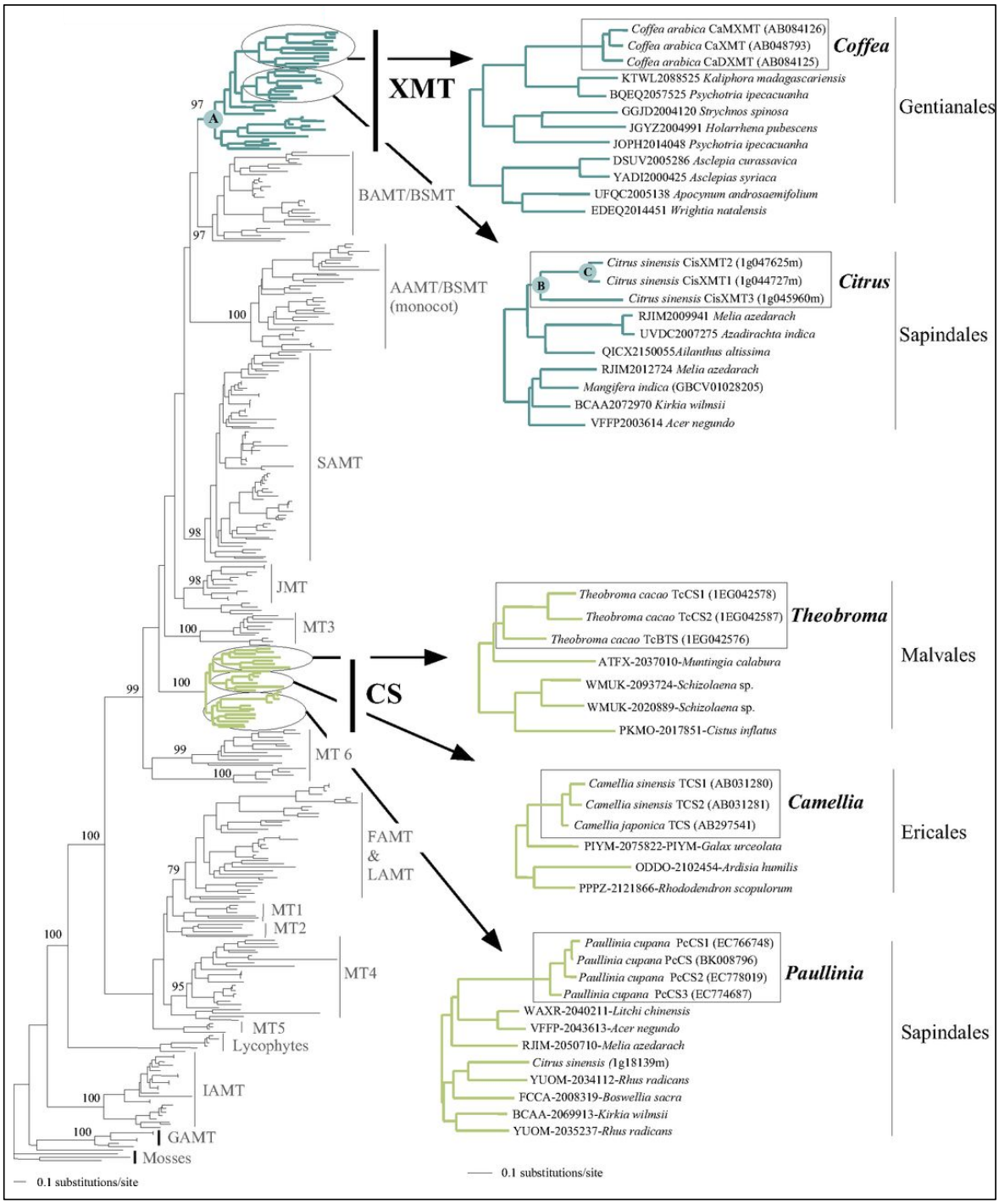

FIGURE 1.18 – La biosynthèse de la caféine est apparue de manière répétée au cours de l'évolution du fait de duplications spécifiques indépendantes. L'arbre représente les relations phylogénétiques entre 356 séquences d'enzymes de la famille des SABATH. Les sous-familles d'enzymes dont la fonction a été caractérisée sont nommées par le nom de l'enzyme et pour celles qui ne le sont pas, sont nommées de manière arbitraire entre MT1 et MT6. La famille d'enzymes CS est colorée en vert clair et la famille des enzymes XMT en vert foncé. Les flèches indiquent les duplications récentes au sein des familles d'enzymes CS et XMT pour les espèces synthétisant la caféine. Les noeuds pour lesquels les enzymes ancestrales ont été reconstruites (Figure 1.19) sont nommés A, B et C dans des cercles colorés. Figure S1 (HUANG et collab., 2016).

Les auteurs de cette étude se sont ensuite intéressés au mécanisme de cooptation chez le citronnier (Figure 1.19). Pour cela, ils ont fait des reconstructions ancestrales des séquences des enzymes à différents points de l'arbre phylogénétique. Ils ont ensuite mesuré l'activité de ces enzymes en présence des différents réactifs présents dans la synthèse de la caféine. On peut s'apercevoir que l'enzyme ancestrale au noeud A et au noeud B n'avait aucune affinité pour les

précurseurs de la synthèse de la caféine, X et Xr, mais plutôt pour l'acide benzoïque (B) et l'acide salicylique (Ba). Puis, l'évolution de l'enzyme entre le noeud B et C a permis l'acquisition d'une affinité de l'enzyme pour certains réactifs de la synthèse de la caféine. Enfin, une duplication a permis la spécialisation des deux paralogues vers des réactifs distincts.

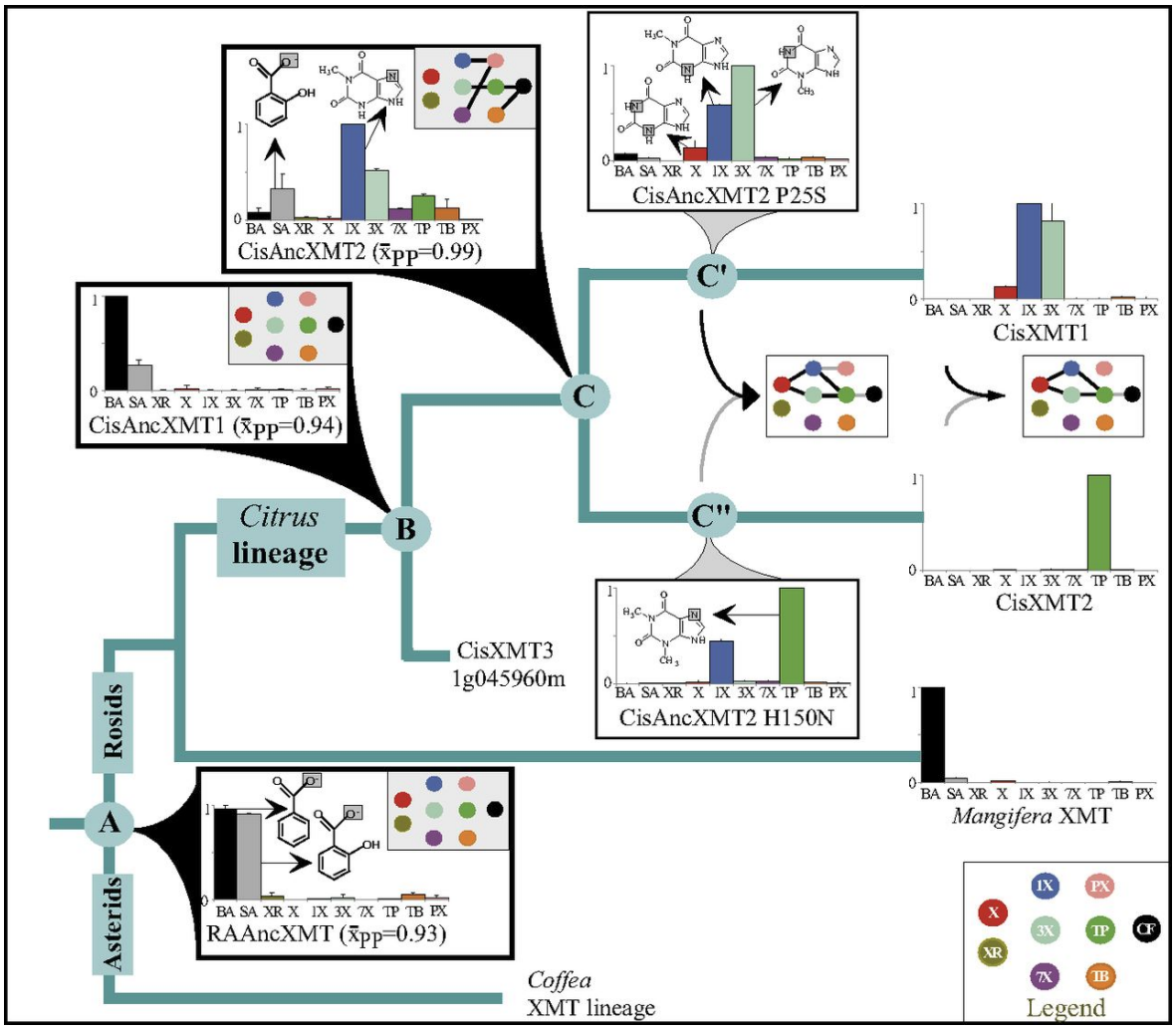

FIGURE 1.19 – La résurrection des enzymes ancestrales de la famille des XMT chez le citronnier met en évidence le chemin évolutif menant à la biosynthèse de la caféine. Les diagrammes en bâtons montrent les activités enzymatiques relatives (entre 0 et 1) de chacune des enzymes reconstruites pour 10 réactifs (BA, acide benzoïque ; SA, acide salicylique ; voire Figure 18 pour les 8 autres). Les noeuds nommés font référence à la figure 19. Le noeud A indique la séquence de l'enzyme reconstruite pour l'ancêtre des *Rosidées* et des *Astéridées* (supérieur à 100 millions d'années), le noeud B pour l'ancêtre du genre *Citrus*, le noeud C pour le dernier ancêtre commun entre CisXMT1 et CisXMT2. Enfin, les noeuds C' et C" indiquent les positions avant la dernière substitution menant à CisXMT1 et CisXMT2. L'ensemble des réactifs de la voie de biosynthèse de la caféine est schématisé par des pastilles de différentes couleurs dans les rectangles noirs. Les traits noirs représentent les réactions pouvant être catalysées par l'ensemble des enzymes présentes à un certain noeud. Le fond du rectangle est blanc lorsque l'ensemble des enzymes permet la biosynthèse de la caféine et gris si ce n'est pas le cas. Les probabilités postérieures des reconstructions ancestrales des séquences aux noeuds A, B et C sont indiquées. Figure 3 (HUANG et collab., 2016).

Dans ces deux exemples, que ce soit la pigmentation florale convergente chez les *Iochroma* ou la biosynthèse convergente de la caféine, la convergence phénotypique possède des bases génétiques communes. Dans le premier cas, il s'agit de la diminution convergente de l'expression d'un réseau de gènes et, dans le second, la formation d'un nouveau réseau de gènes en cooptant les mêmes enzymes ancestrales.

1.2.3.2 … des gènes convergents

La pigmentation des animaux est à la base de plusieurs stratégies convergentes pour se soustraire à la vue de leur prédateur. Par exemple, les animaux vivant dans des habitats clairs, tels que les étendues de sable, sont moins visibles s'ils ont une pigmentation claire. Au contraire, les animaux vivant dans des habitats foncés, tels que des milieux forestiers, sont moins visibles s'ils ont des pigmentations plutôt foncées. C'est le cas d'une espèce de *Peromyscus* : la souris des sables ou *Peromyscus polionotus* qui est une espèce de rongeurs vivant dans le sud des Etats-Unis (Figure 1.20.A). Cette espèce comprend plusieurs sous-espèces qui ont des habitats différents. Certaines vivent sur les côtes du golfe de Floride (en rouge) ou de l'Atlantique (en bleu) et d'autres dans les terres (en marron) (MANCEAU et collab., 2010; STEINER et collab., 2008).

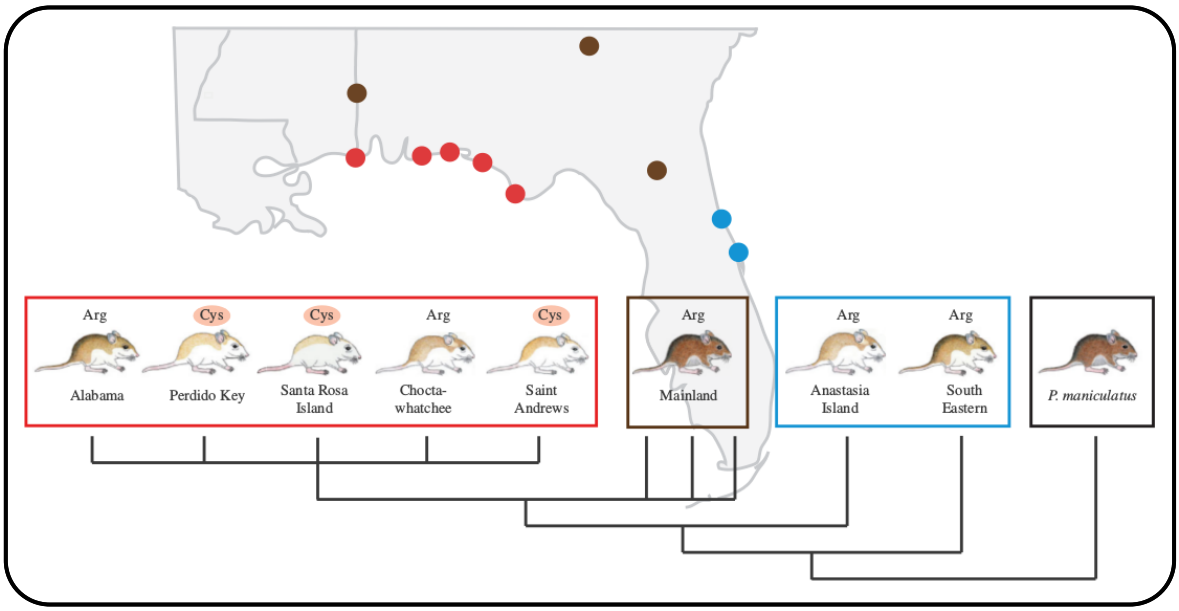

FIGURE 1.20 – La coloration claire chez les souris des sables (*Peromyscus polionotus*) a évolué de manière répétée au travers de changements dans différents gènes. Carte du sud-est des Etats-Unis sur laquelle est indiquée par des ronds de couleur la localité où ont été piégés des *P. polionotus* : trois sous-espèces présentes dans les terres (marrons), cinq sur les côtes du golfe de Floride (rouge) et deux sur la côte Atlantique (bleu). Les souris des sables provenant des côtes Atlantique ou du golfe de Floride ont de manière indépendante évolué vers une coloration plus claire par rapport à leur ancêtre qui est plus foncé. Une substitution dans le gène *Mc1r* (Arg65Cys) contribue à la couleur claire du manteau des sous-espèces des côtes du golfe de Floride mais n'est pas présente chez les souris des sables de la côte Atlantique. L'acide aminé le plus commun à la position 65 est indiqué pour chacune des sous-espèces : l'ancestral (Arg) ou le dérivé (Cys). L'arbre en bas de la figure représente les relations phylogénétiques entre les sous-espèces de souris des sables et le groupe externe, la souris sylvestre (*P. maniculatus*). Figure 2 (MANCEAU et collab., 2010) adapté de (STEINER et collab., 2008).

Cette différence de pigmentation a été caractérisée génétiquement pour les souris des sables de la côte du golfe de Floride. Elle est due en partie à une substitution d'une arginine vers une cystéine au site 65 dans le gène *Mc1r* (STEINER et collab., 2008). Dans le cas des souris des sables de la côte du golfe de Floride (en rouge), il s'agit dans trois cas sur cinq de la même substitution. Dans les autres cas, il s'agit de substitutions à des sites différents (Figure 1.21, A) (STEINER et collab., 2008).

Mc1r est un récepteur membranaire exprimé dans les mélanocytes permettant l'augmentation de la pigmentation via la transduction d'un signal au travers de la membrane. En présence d'*alpha-MSH*, son ligand, une cascade de réactions se met en place impliquant du *cAMP*, ce qui va entraîner une augmentation de la pigmentation (Figure 1.21.B). Une modification de la protéine va modifier son activité de transduction du signal et ainsi modifier la pigmentation. L'implication fonctionnelle de ces substitutions chez ces espèces de rongeurs (STEINER et collab., 2008) mais aussi chez des lézards (ROSENBLUM et collab., 2009) et même chez le mammouth (RÖMPLER et collab., 2006) a été

démontrée. Les lézards étudiés proviennent de trois espèces présentes au Nouveau-Mexique où il existe deux sous-populations : une claire vivant dans le désert de "White Sands" et l'autre foncée vivant proche de ce désert mais dans un environnement foncé. Dans chacune des ces espèces, des substitutions dans le *Mc1r* ont été décrites dans les morphes claires (ROSENBLUM et collab., 2004, 2009). Pour le mammouth, il s'agit d'étudier l'effet des deux allèles [1] de *Mc1r* caractérisés lors du séquençage et de l'annotation de son génome (RÖMPLER et collab., 2006).

Les auteurs d'une revue (MANCEAU et collab., 2010) ont compilé les résultats d'études qui ont synthétisé les protéines correspondant aux séquences de chacune de ces espèces ou sous-espèces pour l'allèle "clair" et l'allèle "foncé". L'activité de la protéine a été mesurée pour chacun des allèles en mesurant la concentration de *cAMP* pour une concentration croissante de son ligand (*alpha-MSH*). On peut voir que, malgré des substitutions différentes, la fonction de transduction du signal de la protéine diminue pour les allèles des espèces vivant dans des milieux clairs par rapport à une protéine d'une espèce vivant en milieu foncé (Figure 1.21.C). Pour deux des espèces testées, les souris des sables de la côte Atlantique et les lézards des sables, la protéine *Mc1r* ne semble pas avoir d'effet et la pigmentation peut avoir des bases génétiques autres que *Mc1r*.

Un autre gène, *Agouti*, a été caractérisé comme étant responsable de modifications de la pigmentation. Il s'agit de l'antagoniste de *Mc1r* qui séquestre le ligand *alpha-MSH* (MANCEAU et collab., 2010; MUNDY, 2009). Il faudrait regarder dans les espèces claires qui n'ont pas de réponse différente avec les différents allèles de *Mc1r*, si le ligand (*alpha-MSH*) est présent dans le compartiment extracellulaire. Si ce n'est pas le cas, les bases génétiques de cette pigmentation plus claire reposent aussi sur la voie de signalisation de *Mc1r* mais celle-ci serait coupée en amont de ce gène.

Dans cet exemple, nous avons vu qu'un même gène est responsable de la pigmentation claire chez les souris des sables de la côte du Golfe de Floride et les lézards blancs du Nouveau-Mexique. Des convergences phénotypiques peuvent donc avoir des bases génétiques communes où un même gène est impliqué dans la réalisation d'un phénotype convergent mais au niveau de sites différents.

1. Des allèles sont des versions différentes d'un gène

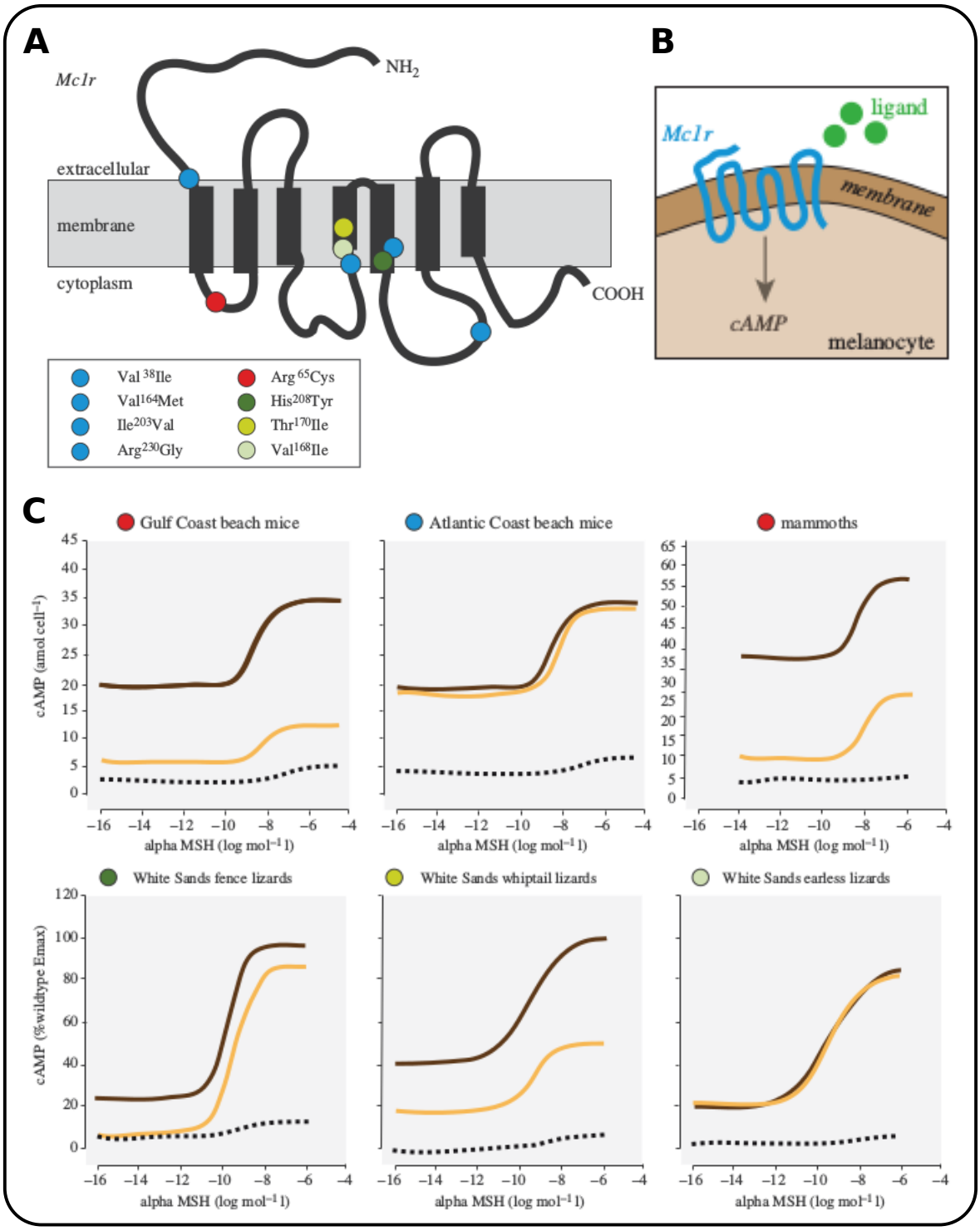

FIGURE 1.21 – Des substitutions dans *Mc1r* peuvent avoir un effet sur sa fonction signalisatrice chez les espèces avec une pigmentation claire mais pas pour toutes. (A) Représentation schématique de la protéine *Mc1r* montrant les positions des substitutions d'acides aminés chez les souris des sables des côtes du golfe de Floride et de l'Atlantique, du mammouth et des lézards blancs du désert des "White Sands" au Nouveau-Mexique. (B) Schéma simplifié du mode de fonctionnement de *Mc1r* . (C) Analyse fonctionnelle des allèles de *Mc1r* chez les différentes espèces étudiées. La concentration intracellulaire d'adénosine monophosphate cyclique (*cAMP*) est mesurée en fonction d'une concentration croissante du ligand, l'hormone mélanocortine alpha (*alpha-MSH*). Pour chaque espèce, les courbes de réponses sont présentées pour l'allèle "foncé" de *Mc1r* présent dans la population (marron), l'allèle clair (jaune) et le contrôle (noir) . Certaines substitutions, mais pas toutes, entraînent une diminution de la fonction de signalisation du récepteur qui est associée à une pigmentation plus claire. (Données provenant de (HOEKSTRA et collab., 2006; ROSENBLUM et collab., 2009; RÖMPLER et collab., 2006; STEINER et collab., 2008). Figure 3 et 4 de (MANCEAU et collab., 2010).

1.2.3.3 … des substitutions convergentes

Nous aurions pu utiliser l'exemple précédent pour montrer qu'une convergence phénotypique peut reposer sur une convergence génétique au niveau d'un site, comme entre les souris des sables de la côte du Golfe de Floride et le mammouth, mais nous allons étudier en supplément le cas de la prestine qui est pour moi un des exemples emblématiques de la convergence phénotypique basée sur une convergence génétique au niveau de la séquence codante de gènes. La prestine (*SLC26A5*) est un gène impliqué dans l'acquisition convergente de l'écholocation entre certaines espèces de chauves-souris et les cétacés (LI et collab., 2010).

Cette étude a montré que la reconstruction de l'arbre phylogénétique de ce gène regroupe le dauphin (Bottlenose dolphin) et les chauves-souris écholocatrices (Figure 1.22.A) plutôt que le dauphin et la vache, alors même que la vache est son espèce la plus apparentée dans cette phylogénie (Figure 1.22.B). Ce regroupement est un artefact de reconstruction, qui s'explique par la présence de substitutions identiques dans le gène de la prestine entre le dauphin et les chauves-souris écholocatrices. Cet ensemble de substitutions présentes de manière concomitante entre les chauves-souris écholocatrices et le dauphin contient notamment une substitution au site 7 entre une asparagine (N) et une thréonine (T) (Figure 1.22.C).

Une étude datant de 2014 a démontré le lien fonctionnel entre la convergence phénotypique et cette substitution convergente (LIU et collab., 2014). Dans un premier temps, les auteurs ont confirmé que cette substitution était bien présente en élargissant l'échantillonnage. Puis, ils ont synthétisé par mutagénèse des protéines de chauves-souris et de dauphins porteuses ou pas de cette substitution.

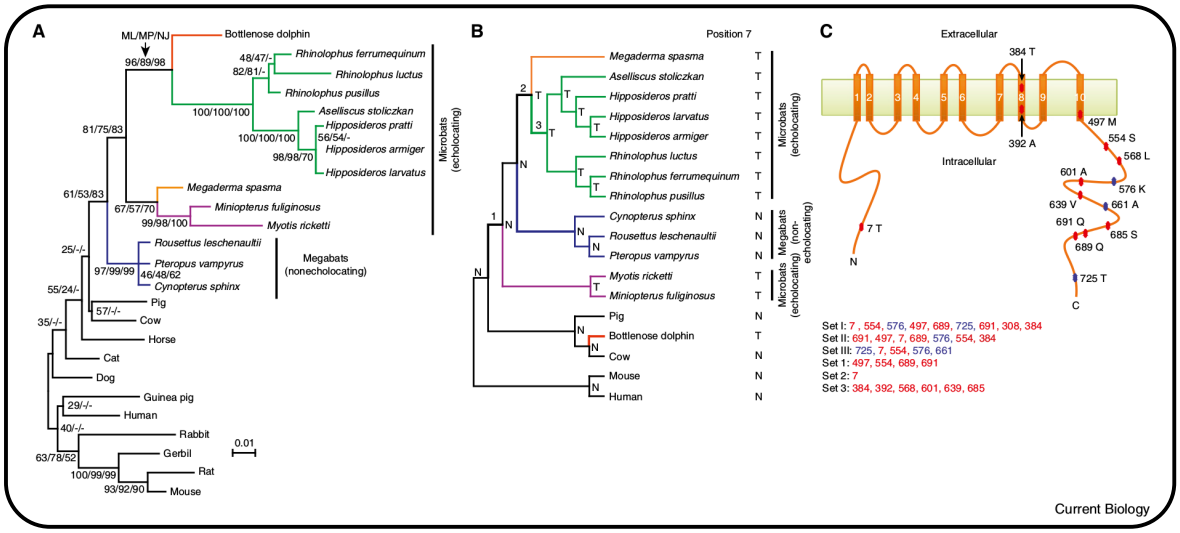

FIGURE 1.22 – (A) Arbre inféré au maximum de vraisemblance en utilisant les séquences protéiques de la prestine pour 25 mammifères. Le modèle utilisé est JTT-f avec un taux de substitutions variant entre sites selon une loi gamma. Les valeurs de bootstraps aux noeuds sont en pourcentage et sont basées sur des analyses de maximum de vraisemblance (ML) maximum de parcimonie (MP) et de neighbor-joining (NJ). (B) Arbre d'espèces pour 18 mammifères. Les longueurs de branches sont arbitraires. Les substitutions convergentes sont examinées entre la branche rouge menant au dauphin et les branches nommées 1, 2 et 3 menant à différents sous-ensembles de chauves-souris. L'acide aminé à la position 7 de la prestine est indiqué pour chacun des nœuds de l'arbre (N : Asn ; T : Thr). (C) Les points de couleur indiquent les positions des sites dans le gène de la prestine qui ont un lien avec la convergence évolutive. Le résidu affiché est celui du dauphin. Les sets I, II et III sont les groupes de sites qui entraînent le mauvais positionnement de clades dans le panel A par rapport au panel B (classé de celui qui a le plus d'impact à celui qui en a le moins). Le set I correspond au mauvais placement du dauphin et du clade de chauves-souris violettes, le set II, au mauvais placement du dauphin uniquement et le set III à celui du clade de chauves-souris violettes uniquement. Les sets 1, 2, et 3 sont les sites qui présentent des substitutions convergentes entre le dauphin et les espèces sous les noeuds, 1 ,2 et 3. Figure 1 (LI et collab., 2010).

Ensuite, ils ont testé l'effet fonctionnel de cette substitution dans des cellules humaines (HEK293). Pour celà, ils ont mesuré des paramètres liés à l'électromobilité de la membrane. En effet, la prestine est une protéine transmembranaire (transporteur d'anion) qui est exprimée dans les cellules ciliées externes de l'oreille interne. Elle a également une fonction motrice qui, en réponse à un changement de potentiel de la membrane, va entraîner un changement de la conformation de la cellule et ainsi amplifier les sons (ZHENG et collab., 2000).

Les auteurs ont montré un effet significatif de la présence de cette substitution sur ces paramètres et donc probablement sur l'audition des espèces porteuses de cette substitution. Malheureusement, les auteurs n'ont pas testé si l'ajout de cette substitution, par exemple dans la séquence de la vache, avait un effet sur ces paramètres. Cette analyse fonctionnelle a aussi été réalisée sur les autres substitutions identifiées dans (LI et collab., 2010) et confirmée dans (LIU et collab., 2014). Les résultats montrent la même tendance mais ne sont pas significatifs (LIU et collab., 2014).

Cet exemple montre qu'une convergence fonctionnelle, l'écholocation, qui est un caractère primordial pour ces espèces pour se nourrir et se déplacer dans leur environnement, est codée pour une partie par une substitution dans une séquence codante. Cette convergence a donc des bases génétiques profondes qui s'enracinent au niveau d'un site de la séquence codante d'un gène.

1.2.3.4 ... des convergences génotypiques plus ou moins profondes

La convergence est présente également à différents niveaux hiérarchiques au niveau génétique comme c'était le cas au niveau phénotypique. Dans certains cas, la convergence peut avoir une base génétique très spécifique, une substitution qui va avoir des effets à tous les niveaux hiérarchiques supérieurs (Figure 1.23.C).

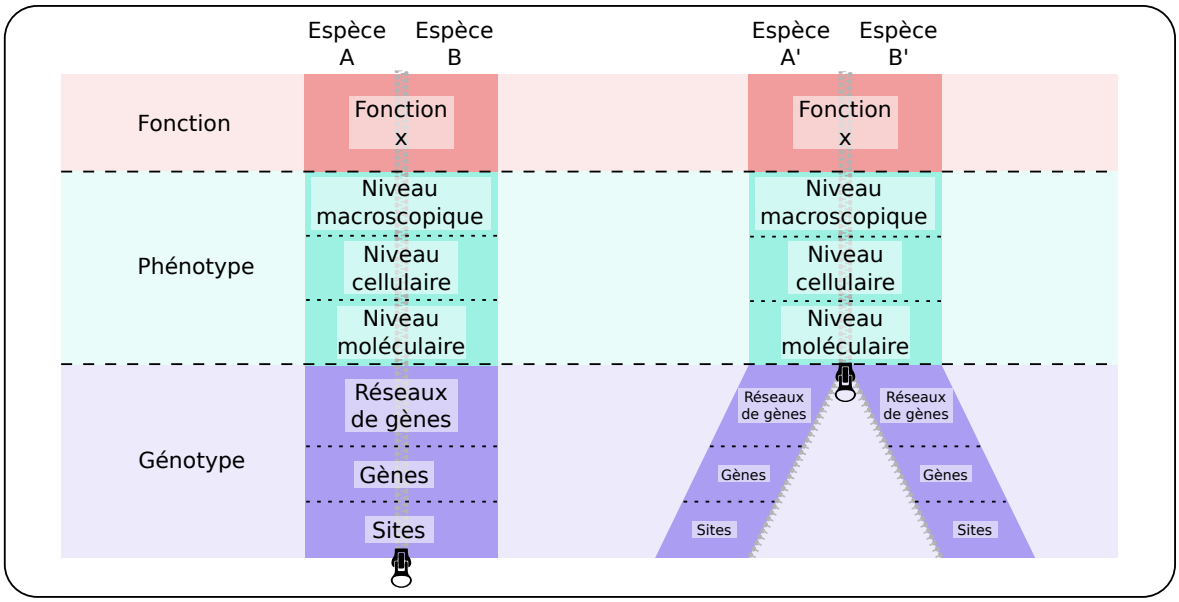

FIGURE 1.23 – Le niveau de convergence peut descendre au niveau génotypique.

1.2.4 Que peut-on conclure sur les bases de la convergence dans le monde vivant ?

Au travers des exemples des paragraphes précédents, nous avons vu que les convergences phénotypiques peuvent avoir des bases convergentes plus ou moins profondes, allant d'une simple similarité d'organes (les animaux marins planctivores) à une substitution identique (la prestine chez les animaux écholocateurs).

On peut se demander si la répartition des bases sous-jacentes à la convergence phénotypique est homogène au travers des différents étages ou bien s'il y a un biais vers un certain étage :

— Est-ce que les bases sous-jacentes à la convergence fonctionnelle sont peu profondes ? Cela signifierait qu'il existe plusieurs manières de réaliser une fonction avec des moyens phénotypiques différents (Figure 1.24.A).

— Est-ce que l'Évolution réutilise les mêmes bases phénotypiques pour réaliser une même fonction mais en utilisant des bases génétiques différentes ? (Figure 1.24.B)

— Est-ce que l'Évolution réutilise de manière préférentielle les mêmes bases génétiques pour reproduire une même fonction ? (Figure 1.24.C)

— Est-ce qu'il n'y a pas de biais vers un niveau en particulier et une convergence fonctionnelle peut avoir des bases sous jacentes au niveau phénotypique comme au niveau génotypique ou bien encore n'avoir aucune base commune ? (Figure 1.24.D)

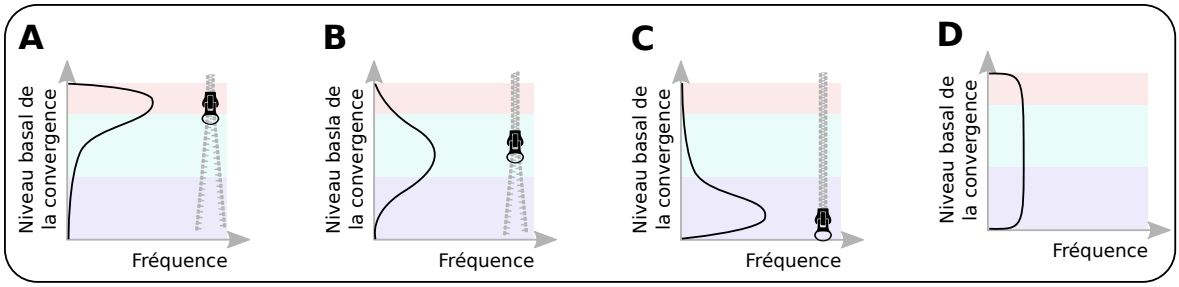

FIGURE 1.24 – Différentes hypothèses sur la répartition du niveau de convergence le plus bas au sein des êtres vivants.

Les exemples que nous avons vus ainsi que les nombreux autres présents dans la littérature ne permettent pas de généraliser à l'ensemble du vivant. Pour cela, il faudrait bien plus d'analyses où la convergence fonctionnelle serait étudiée et disséquée jusqu'aux bases génétiques. Ce processus est en cours mais il prendra énormément de temps et aussi d'argent.

De plus, ces hypothèses ne prennent pas en compte d'autres facteurs tels que la complexité du trait convergent étudié. Par exemple, synthétiser une molécule ou se nourrir de plancton sont deux choses de complexité bien différente. On peut aussi prendre en compte la distance phylogénétique entre les espèces étudiées. Des espèces proches vont avoir des différences génétiques moins fortes et il sera peut-être plus facile de réutiliser un gène en particulier car il aura moins divergé. De la même manière, des espèces très distantes auront peu de similarités génétiques et la possibilité d'avoir des bases génétiques convergentes sera peut-être impossible.

Cependant on sait que la réalisation d'une fonction a toujours des bases génétiques. Un moyen d'aborder le problème est de le regarder dans l'autre sens. Est-ce qu'il y a des ressemblances génétiques communes entre des espèces avec le même trait phénotypique ? Il est alors ensuite plus facile de tester fonctionnellement si le gène en question est lié de manière causative au phénotype convergent ou s'il s'agit seulement d'une corrélation.

Nous allons voir par la suite quels sont les processus pouvant entraîner de la convergence génétique corrélée à un phénotype convergent.

1.3 Quels sont les processus pouvant mener à la convergence génétique corrélée à un phénotype convergent ?

Nous avons jusqu'à maintenant observé que la convergence phénotypique se retrouvait à différentes échelles et que dans certains cas elle avait des bases génétiques également convergentes. Nous allons maintenant nous intéresser à la relation entre une convergence phénotypique et ses bases génétiques potentielles. Mais avant celà, nous allons étudier de manière théorique le lien entre le processus de fixation des mutations et d'un phénotype dans une espèce ainsi que les

processus pouvant mener à de la convergence génétique corrélée à une convergence phénotypique mais indépendante du phénotype.

1.3.1 Comment se fixe une mutation dans une espèce ?

Nous allons tout d'abord nous intéresser aux processus permettant à une mutation de se fixer dans une espèce, c'est à dire comment une mutation qui se produit dans un individu se répand à l'ensemble des individus d'une espèce. Cette section est écrite dans le but d'être concise et accessible à un lectorat non-spécialiste, cependant pour un développement plus complet, je réfère à tout livre de référence traitant de l'évolution tel que *Evolution : Making Sense of Life* (ZIMMER et EMLEN, 2015). Sous le terme mutation, je regroupe toutes les modifications génomiques, que ce soit à l'échelle d'un site, avec un remplacement d'un nucléotide vers un autre, une délétion [1] ou une insertion [2] ou à plus grande échelle avec une duplication de gènes ou un réarrangement chromosomique.

L'apparition de mutations est un processus aléatoire qui se produit en faible quantité dans chacun des individus. Chacun des individus est donc porteur de mutations qu'il transmettra, ou non, à sa descendance. Cela signifie que, par hasard, au terme d'un grand nombre de générations, une mutation peut se répandre dans l'ensemble des individus d'une espèce et se fixer. Elle peut également, par hasard, être éliminée de la population. Ce phénomène aléatoire est appelé la **dérive génétique**.

Certaines mutations peuvent modifier le phénotype de l'individu, ce qui va avoir un impact positif ou négatif sur la fitness de l'individu, c'est à dire la capacité de l'individu à avoir une descendance. Cet avantage ou désavantage modifie la probabilité de fixation de ces mutations. En effet, si une mutation entraîne un avantage sélectif, il y a plus de chance que les individus porteurs de cette mutation aient une descendance et donc que la mutation soit transmise de génération en génération. Il s'agit de la **sélection naturelle**.

1.3.2 La sélection naturelle favorise la propagation des phénotypes avantageux dans les populations

La sélection et la dérive sont donc deux processus permettant la fixation d'une mutation ayant un effet phénotypique. Cela signifie que pour un phénotype quantitatif donné, si on fait les deux hypothèses que 1/ le phénotype possède une valeur pour laquelle la fitness de l'individu est optimale et que 2/ les mutations permettant d'obtenir les différentes valeurs du phénotype sont équiprobables, alors la probabilité que le phénotype le plus avantageux se fixe dépend de la force de la sélection par rapport à la dérive. Si la sélection n'est pas efficace, c'est à dire que les mutations se fixent par la dérive [3] et donc indépendamment de leur fitness, alors la probabilité que le phénotype optimal se fixe dans la population est équivalente à la fixation de tout autre phénotype. Par contre, plus l'efficacité de la sélection est forte plus la probabilité d'obtenir le phénotype le plus avantageux est important (Figure 1.25).

1. perte d'une ou plusieurs bases nucléiques
2. ajout d'une ou plusieurs bases nucléiques
3. c'est à dire par hasard

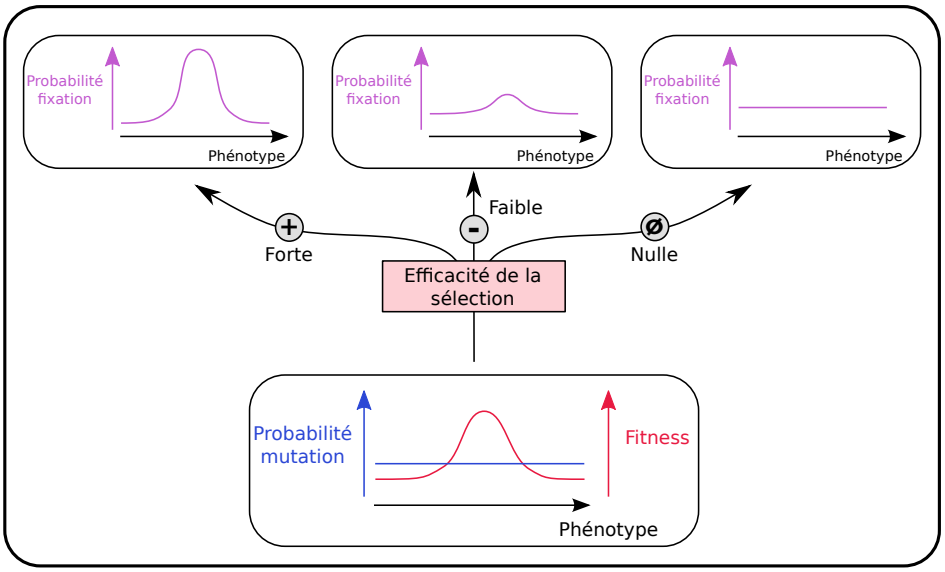

FIGURE 1.25 – L'efficacité de la sélection augmente la probabilité de fixation du phénotype le plus avantageux. Dans la partie basse de la figure est représentée la fitness (en rouge) de différentes grandeurs d'un phénotype quantitatif hypothétique et la probabilité d'obtention de la mutation produisant chacune de ces grandeurs (en bleu). Cette probabilité de mutation est considérée uniforme à des fins de simplifications.

1.3.3 Une taille efficace de population élevée est liée à une bonne efficacité de la sélection

L'efficacité de la sélection est fortement déterminée par la taille efficace (Ne) de l'espèce, c'est à dire le nombre d'individus qui composent une espèce et plus précisément le nombre d'individus participant à la diversité génétique. Plus la taille efficace d'une espèce est grande plus l'efficacité de la sélection dans cette espèce sera importante et inversement.

Plus une espèce comprend d'individus, plus les mutations avantageuses auront de chance de se fixer. En effet, les effets stochastiques [1] liés à la transmission des mutations de génération en génération seront d'autant plus faibles que le nombre d'individus sera grand. Ainsi, plus la taille efficace d'une espèce sera grande plus la sélection naturelle aura un effet par rapport à la dérive génétique sur la fixation des mutations.

1.3.4 La fixation de mutations peut être biaisée par des processus indépendants de la sélection naturelle

Les mutations peuvent se fixer sous l'effet de la dérive et de la sélection comme nous l'avons déjà vu, mais il existe aussi des processus biologiques qui biaisent la probabilité d'apparition et de fixation des mutations, indépendamment de leur fitness. Au travers de deux exemples de biais, je veux montrer comment des processus biologiques peuvent modifier la fixation des mutations indépendamment de la sélection naturelle. Cette énumération n'a pas pour but d'être exhaustive.

1.3.4.1 L'instabilité des dinucléotides CpG méthylés biaise les probabilités d'apparition des mutations

Certaines mutations sont plus probables que d'autres dans le génome. En effet, il existe des biais mutationnels pouvant modifier la production de mutations et ainsi modifier les probabilités d'apparition de mutations en favorisant certaines mutations par rapport à d'autres.

1. imprévisibles, dus au hasard

Par exemple, sur les brins d'ADN, les dinucléotides successifs cystéine (C) et guanine (G), notés CpG, sont particulièrement instables lorsque la cystéine (C) est méthylée. En effet, la cystéine peut spontanément se désaminer pour produire de l'uracile (U) qui est ensuite corrigée par la machinerie de réparation de l'ADN en thymine (T). A terme, le dinucléotide CpG est donc muté en dinucléotide TpG. Chez les vertébrés, ce processus est particulièrement actif et explique la sous représentation des dinucléotides CpG par rapport à la quantité attendue sous une hypothèse homogène (LI et ZHANG, 2014).

1.3.4.2 La gBGC biaise la fixation des mutations indépendamment de leur valeur sélective

Il existe également des biais de fixation des mutations qui vont faciliter la fixation des mutations indépendamment de leur valeur sélective et de leur probabilité d'apparition. Par exemple, une mutation qui est apparue dans un individu peut avoir une plus grande chance de se fixer par sa nature que d'autres mutations de natures différentes.

C'est le cas de la conversion génique biaisée vers le GC, ou gBGC (GC-biased gene conversion), qui modifie les compositions en base des séquences. Lors de la recombinaison méiotique et plus particulièrement lors de la résolution des hétéroduplex formés par les brins de chromosomes homologues, le polymorphisme potentiel entre les chromosomes est corrigé de manière préférentielle vers la guanine (G) et la cystéine (C) par la machinerie de réparation de l'ADN. La gBGC entraîne donc un biais de la composition en base des séquences vers G et C via la fixation préférentielle de mutation vers les bases G et C (DURET et GALTIER, 2009; PESSIA et collab., 2012).

La fixation d'une mutation dans une population va donc dépendre de différents facteurs : sa probabilité d'apparition qui peut être biaisée, sa valeur sélective qui aura d'autant plus d'importance que la sélection sera efficace et des biais pouvant modifier sa probabilité de fixation. De plus, ces biais peuvent interagir en s'annulant ou s'additionnant.

1.3.5 Pourquoi observe-t-on de la convergence au niveau génétique corrélée à un phénotype convergent ?

La convergence génétique corrélée à un phénotype convergent se produit lorsque des mutations identiques se fixent dans deux espèces différentes menant à un même phénotype. La probabilité de convergence entre deux espèces est donc corrélée à la similarité entre les distributions des probabilités de fixation des mutations entre les deux espèces.

Comme nous avons vu précédemment la probabilité de fixation d'une mutation dans une espèce est corrélée à sa valeur sélective mais également à des biais pouvant modifier les probabilités de fixations et interférer dans cette relation. La convergence génétique entre différentes espèces peut provenir soit de la fixation de mutations ayant des valeurs sélectives élevées dans leur environnement respectif, soit du partage de biais menant à la fixation préférentielle de mutations indépendamment de leur valeur sélective.

1.3.5.1 Une mutation est liée à un phénotype très avantageux dans différentes espèces

Les espèces vivant dans un environnement similaire sont celles qui ont développé des phénotypes permettant de s'adapter aux pressions de sélection communes de cet environnement. Or, s'il existe seulement un phénotype permettant de répondre de manière optimale aux contraintes de cet environnement, alors les mutations permettant de l'obtenir seront très avantageuses. Si le nombre de ces mutations est faible et que ces mutations peuvent exister dans les deux espèces alors il y a une forte probabilité de convergence génétique entre ces espèces.

Par exemple, il a été montré chez l'épinoche qu'une délétion observée plusieurs fois dans la région régulatrice d'un gène, *Pitx1*, permettait une adaptation au milieu lacustre (CHAN et col-

lab., 2009). En effet, une délétion dans cette zone régulatrice entraîne des changement majeurs dans l'expression de ce gène ce qui provoque des répercussions phénotypiques importantes avec notamment la réduction de plaques osseuses lui permettant de se protéger en milieu marin.

Cet exemple montre également que si l'acquisition d'un phénotype est accessible par une ou un petit nombre de mutations, c'est à dire que la carte génotype-phénotype est simple, alors la probabilité de convergence devrait être plus élevée.

Une mutation à l'origine d'un phénotype très avantageux et qui peut se produire dans différentes espèces peut donc être à l'origine d'une convergence génétique. Dans ce cas là, la convergence génétique est liée à l'acquisition du phénotype convergent.

1.3.5.2 De faibles distances phylogénétiques devraient augmenter la probabilité de convergence génétique

Une espèce ne peut pas acquérir subitement un nouvel organe tel qu'un nouveau membre. Les innovations développementales, comme par exemple l'acquisition des nageoires chez les mammifères marins, proviennent généralement de la réutilisation d'un organe à de nouvelles fins. En effet, chacune des espèces hérite des contraintes développementales de leurs ancêtres. Cette empreinte du passé se retrouve également dans la similarité d'organisation du génome entre les espèces proches. Des mutations dans des espèces relativement proches auront probablement les mêmes conséquences phénotypiques et inversement une même mutation dans des espèces très éloignées aura des conséquences différentes sur ce même phénotype.

Dans ce cadre on peut donc imaginer que la distribution des probabilités de fixation des mutations est plus similaire entre des espèces relativement apparentées que dans des espèces plus éloignées et donc que la probabilité de convergence génétique est plus forte entre les espèces relativement proches phylogénétiquement (ROSENBLUM et collab., 2014).

Cependant, cette hypothèse n'est pas toujours vérifiée. Dans le cas de l'acquisition convergente de coloration plus claire du corps pour se cacher des prédateurs, nous avons vu qu'un même gène (cf exemple *Mc1r*) était impliqué entre des reptiles et des mammifères (ARENDT et REZNICK, 2008; GOMPEL et PRUD'HOMME, 2009) mais que des gènes différents étaient impliqués entre deux espèces de souris (STEINER et collab., 2008).

1.3.5.3 Des biais convergents pourraient induire de la convergence génétique

Les biais d'apparition de mutations ou de fixation des mutations peuvent modifier la probabilité de fixation des mutations indépendamment de leur valeur sélective. Il est donc possible que de la convergence génétique soit produite par ces biais entre des espèces et donc indépendamment de la sélection naturelle. Elle serait alors non-adaptative.

Par exemple, la gBGC influe sur le taux de GC chez les mammifères et a également une intensité variable entre les clades (ROMIGUIER et collab., 2010). Ce biais pourrait donc induire de la convergence génétique dans les régions fortement recombinantes, les enrichir en GC et provoquer de la convergence au niveau des séquences indépendamment du phénotype. Pour un autre exemple, dans le cas des dinucléotides CpG, il a déjà été montré dans un gène candidat que de la convergence génétique pouvait être liée à une substitution dans un dinucléotide CpG (ZHU et collab., 2018).

En conclusion, la convergence génétique peut être liée à de la convergence adaptative mais il se peut que d'autres phénomènes viennent ajouter une quantité de convergence génétique indépendante de la sélection naturelle et qu'elle soit par hasard corrélée à un phénotype convergent. Si on s'intéresse aux bases génétiques de la convergence phénotypique, c'est à dire à la part liée à la sélection naturelle, il faudra prendre en compte ces facteurs qui vont pouvoir ajouter une proportion inconnue de convergence génétique indépendante du phénotype convergent.

1.4 Objectifs de ma thèse

Le projet de ma thèse était d'étudier la convergence génétique sous jacente à une convergence phénotypique à l'échelle génomique. Pour cela, j'ai mis en place des outils pour étudier la convergence génomique, étudier de façon théorique les processus confondants et enfin, appliquer ces outils et connaissances à un jeu de données réel.

L'étude de la convergence génétique sous jacente à un phénotype convergent implique l'étude d'un cas concret afin de se confronter à la réalité. Les études à l'échelle de gènes candidats permettent d'étudier le lien entre le phénotype et ses bases génétiques sur des gènes ciblés mais ne permettent pas d'étudier la quantité globale de convergence. Je me suis donc intéressée à l'ensemble des gènes, c'est à dire à l'échelle génomique.

L'étude d'un jeu de données réel induit la construction d'un jeu de données constitué des génomes des espèces convergentes étudiées mais également d'espèces non-convergentes proches. L'étude comparative de ces génomes permettra de détecter la convergence génétique présente dans les données, c'est à dire les mutations corrélées au phénotype convergent. Elle permettra également de quantifier la proportion de convergence génétique non adaptative qui est indépendante des phénotypes convergents.

Cependant, le passage pratique à l'étude empirique est délicat. La construction du jeu, c'est à dire la première étape et qui sera celle sur laquelle repose la suite des analyses, n'est pas triviale. Les séquences génomiques des espèces sont difficilement accessibles. En effet, il n'est pas possible de "lire" le génome d'une espèce, il faut le séquencer. Cette étape nécessite de fragmenter l'ADN, puis de décoder chacun des fragments et ensuite de ré-assembler les morceaux tel un puzzle, ce qui n'est pas évident. Dans ma thèse, je me suis intéressée dans un premier temps à une possibilité d'améliorer la construction des jeux de données destinés à la génomique comparative en créant un outil, CAARS que je présente dans le premier chapitre (Partie 2).

Ensuite, je me suis intéressée à la détection de la convergence génétique et ai créé un nouvel outil (PCOC) afin de proposer une alternative aux méthodes existantes. J'ai ensuite fait une analyse comparative de la capacité de PCOC et des méthodes existantes à détecter des substitutions convergentes en présence de facteurs confondants. L'ensemble de ces travaux sur la détection de la convergence est présenté dans le second chapitre (Partie 3).

Ces travaux préalables ont ensuite été utilisés pour réaliser une étude de la convergence dans un cas réel que je présente dans le troisième et dernier chapitre de ma thèse. Dans cette étude, j'ai cherché à caractériser les bases génomiques de l'adaptation aux milieux arides chez les rongeurs (Partie 4).

Ces trois chapitres sont donc plus ou moins indépendants mais correspondent aux prérequis qui ont été nécessaires pour étudier un cas réel de convergence.

2 Construction automatisée de jeux de données pour des études de génomique comparative avec CAARS

Sommaire

2.1 Pourquoi un nouvel outil ?

La première partie de ma thèse a été consacrée à développer un outil bioinformatique nommé CAARS. Cet outil permet la création automatique de jeux de données destinés à des analyses de génomique comparative regroupant des organismes modèles et non-modèles.

La génomique comparative vise à étudier les différences et les similarités entre les génomes de différentes espèces en termes de structures, fonctions, évolution, composition en gènes... Ces analyses utilisent donc les génomes de l'ensemble des espèces étudiées, ou uniquement une partie de ces génomes, par exemple l'ensemble des gènes. La première étape de ces analyses est de créer le jeu de données permettant de répondre à la problématique posée. Cela implique que la qualité de cette analyse repose entièrement sur la qualité du jeu de données.

Cependant il n'existe pas de méthode automatisée et standardisée pour réaliser de tels jeux de données. De plus, nous avons identifié dans les analyses de génomique comparative des moyens d'améliorer de manière significative la construction, et donc la qualité, des jeux de données.

C'est pour cela que nous avons voulu mettre à disposition de la communauté un outil complet, permettant la création de jeux de données de bonne qualité destinés à la génomique comparée, s'assurant de sa reproductibilité et permettant de gagner du temps sur sa construction. CAARS a également pour objectif d'améliorer la qualité de ces jeux de données en incluant un plus grand nombre de paralogues annotés et en améliorant l'assemblage *de novo* des séquences des gènes.

2.1.1 Fournir un outil complet permettant de...

2.1.1.1 créer des jeux de donnés destinés à la génomique comparative, ...

La plupart des analyses de génomique comparée partent d'un même type de jeu de données, des alignements multi-géniques, c'est à dire les séquences codantes de l'ensemble des espèces étudiées regroupées en familles de gènes où les sites homologues sont alignés. En effet, nous avons besoin de comparer ce qui est comparable, il faut donc s'assurer de l'homologie de ce que l'on compare.

Ce type d'analyse tend à comparer de plus en plus d'espèces non modèles entre elles, c'est à dire des espèces dont le génome annoté n'est pas disponible. Actuellement, il est plus simple d'obtenir les séquences codantes des gènes de ces espèces en séquençant leurs transcriptomes puis en les assemblant à partir de données RNA-Seq plutôt que de séquencer leur génome puis d'inférer les séquences codantes.

Il existe des bases de données publiques contenant des alignements de séquences de gènes de différentes espèces modèles pour l'ensemble des familles de gènes connus (Ensembl Compara (Herrero et collab., 2016), Orthomam (Douzery et collab., 2014), EGNOGG (Huerta-Cepas et collab., 2015)). Le but de CAARS est de réutiliser cette connaissance et d'incorporer les séquences des gènes d'une ou de plusieurs nouveaux transcriptomes dans ces alignements existants plutôt que de refaire l'annotation des familles à partir de zéro en utilisant l'ensemble des séquences de toutes les espèces étudiées (Figure 2.1).

En 2015, au début de ma thèse, il n'existait pas d'outil intégré permettant de partir de données RNA-Seq de nos espèces et d'obtenir les alignements. Cette transformation nécessite de nombreuses étapes, chacune mobilisant des outils particuliers. Cependant, aucun outil ne les regroupait en un seul et même pipeline. Nous avons donc voulu fournir un outil qui pourrait combler ce manque : CAARS.

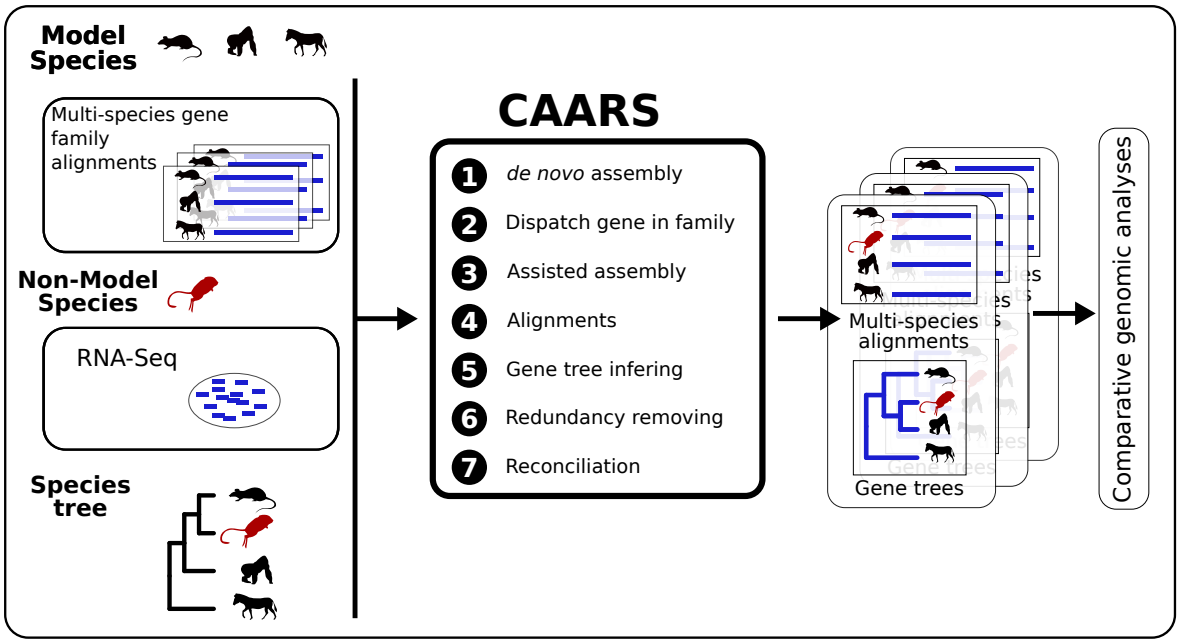

FIGURE 2.1 – Aperçu du fonctionnement de CAARS. A gauche sont représentées les principales entrées du programme et à droite les sorties. Le rectangle au centre récapitule l'ensemble des étapes réalisées par CAARS permettant d'intégrer les séquences de l'espèce non-modèle (en rouge) dans les familles de gènes des espèces modèles (en noir). Les traits bleus représentent les séquences d'un gène.

CAARS réalise différentes tâches entièrement automatisées :

1. Dans un premier temps, les données de RNA-Seq de chaque nouvelle espèce sont assemblées *de novo*.
2. Ces séquences, appelées contigs, sont ensuite réparties dans les familles existantes en utilisant comme critère la similarité de séquence.
3. Dans un second temps, CAARS travaille au niveau de la famille de gènes, c'est à dire que les tâches suivantes sont réalisées famille par famille et sont donc facilement parallélisables. Pour chaque famille, les contigs attribués à cette famille sont raffinés par une méthode d'assemblage qui est spécifique à CAARS et qui sera détaillée dans un prochain paragraphe.
4. L'ensemble des séquences des espèces modèles provenant de l'alignement existant pour cette famille et les contigs raffinés sont alignés.
5. Cet alignement est utilisé pour inférer un arbre phylogénétique de ce gène.
6. Les séquences très similaires seront plus proches dans l'arbre que des séquences partageant moins de similarités. Cette caractéristique est utilisée pour supprimer la redondance produite par les assembleurs de données RNA-seq. En effet, les outils réalisant les assemblages peuvent créer de la redondance en construisant des séquences quasi identiques différant par des erreurs de séquençage ou des polymorphismes, ou par des régions plus ou moins courtes, comme dans le cas des isoformes. Ces séquences seront très similaires et seront donc placées côte-à-côte dans l'arbre. Elles seront ainsi plus facilement identifiables.
7. Finalement l'arbre est réconcilié, c'est à dire que l'information contenue dans l'arbre d'espèces est utilisée pour raffiner les arbres de gènes. Cet arbre réconcilié est ensuite utilisé pour annoter les relations d'homologie entre les séquences des gènes de la famille et définir des sous-groupes de gènes orthologues.

2.1.1.2 d'assurer la reproductibilité, ...

En génomique comparée, le manque de reproductibilité vient souvent du fait que chacune des analyses utilise son propre pipeline pour créer son jeu de données. Comme dans les articles scientifiques, la partie "matériels et méthodes" est souvent réduite à son minimum, il manque parfois la version des programmes utilisés ou certaines options ou paramètres. Il y a aussi certains choix ou filtres qui parfois ne sont pas complètement explicités. Dans ces conditions, il est impossible de recréer le jeu de données ayant servi comme base d'une analyse et donc il devient très difficile de reproduire les résultats. Cela est d'autant plus vrai qu'auparavant les scripts utilisés pour réaliser les analyses étaient rarement disponibles, mais cela commence à changer. Les revues scientifiques demandent de plus en plus la publication des codes ayant servi à réaliser l'analyse en même temps que l'article (STODDEN et collab., 2013).

Pour éviter ces problèmes de reproductibilité, nous avons utilisé lors du développement de CAARS un gestionnaire automatique de pipeline et une technologie de containerisation (voir ci-dessous).

Le gestionnaire de pipeline permet d'automatiser le lancement de toutes les tâches. Cela permet de ne lancer aucune tâche à la main si ce n'est la commande générale. Il est donc plus facile d'assurer la reproductibilité de la création de son jeu de données. Il suffit de rendre disponible les données de départ et de documenter l'utilisation de CAARS. De plus, l'utilisation d'un gestionnaire de pipeline permet la parallélisation automatique de CAARS et également la reprise sur erreurs. Si CAARS s'arrête ou est arrêté par une coupure de courant par exemple, il reprendra là où il s'est arrêté et l'utilisateur n'a pas à se soucier de ce qui a déjà tourné ou pas.

D'autre part nous avons choisi de containeriser CAARS. Les containers permettent de fournir un environnement indépendant de la machine hôte contenant tous les outils nécessaires aux différentes tâches de CAARS. Cela permet d'éviter l'installation des dépendances de CAARS sur la machine hôte pour gagner du temps, éviter les conflits entre les versions de logiciels installés sur la machine hôte et le container et enfin pour la reproductibilité, les versions installées sont fixées pour une version donnée dans le container.

2.1.1.3 ... et gagner du temps

Enfin, la création de ce nouvel outil a pour but de faire gagner du temps aux bioinformaticiens débutants mais aussi confirmés en charge de la création de jeux de données.

Dans le cas des bioinformaticiens débutants, la création d'un jeu de données peut prendre beaucoup de temps car il faut apprendre à gérer les conversions entre les différents formats, être capable de choisir parmi les nombreux outils différents et leurs options pour réaliser chacune des tâches. CAARS permet de leur faire gagner du temps et les rassurer sur les choix des outils responsables des tâches. Les solutions retenues dans CAARS sont celles utilisées et acceptées par la communauté, bien que ce ne soit pas une garantie de qualité. Lorsqu'il y avait des alternatives possibles nous avons essayé de faire les choix qui nous semblaient les plus judicieux. Par exemple, nous avons choisi comme assembleur *de novo* Trinity (GRABHERR et collab., 2011). Il n'est pas l'assembleur le plus rapide mais il montre des statistiques de qualité d'assemblages parmi les meilleures (CLARKE et collab., 2013; HÖLZER et MARZ, 2019; VOSHALL et MORIYAMA, 2018).

Dans le cas des bioinformaticiens confirmés, CAARS leur évite de coder des choses maintenant existantes et apporte en plus la garantie de la reproductibilité de la création de leur jeu de données sans effort supplémentaire. De plus, éviter la duplication des codes réalisant les mêmes choses simplifie la reproduction des analyses.

Finalement, le gain de temps obtenu grâce à CAARS sur la création du jeu de données permet de passer plus de temps sur l'analyse biologique en elle-même.

2.1.2 Pouvoir inclure les paralogues dans les analyses de génomique comparative

Une autre motivation pour la création de CAARS est d'améliorer la qualité des jeux de données de génomique comparative en permettant une annotation des paralogues.

En lisant la littérature au début de ma thèse, j'ai constaté que de nombreuses études excluaient les paralogues de leur jeu de données (ISHIKAWA et collab., 2016; MARRA et collab., 2014; PEREIRA et collab., 2016; THOMPSON et ORTÍ, 2016). En effet, ces études utilisaient une annotation basée sur les similarités de séquences par Reciprocal Best Hit (RBH) qui, par construction, ne permet d'annoter que les orthologues 1:1.

L'annotation des relations d'orthologie entre les séquences de deux espèces, une espèce référence et une espèce questionnée par RBH est basée sur l'analyse des scores de similarités entre les séquences de 2 espèces. Dans un premier temps, on calcule les scores de similarité entre les séquences d'une espèce et les séquences de l'autre espèce. Puis inversement. L'orthologie entre deux séquences de chacune des espèces est établie si le meilleur score de similarité pour chacune des séquences correspond à l'autre séquence. Cette méthode fonctionne parfaitement dans le cas de gène orthologue 1:1 (dans l'exemple set n°1) mais lorsqu'il s'agit de gènes dupliqués le RBH peut mener sur une mauvaise association (Figure 2.2). Dans CAARS, nous utilisons une méthode permettant d'annoter les relations d'orthologie et de paralogie en inférant l'histoire du gène grâce à l'étape de réconciliation.

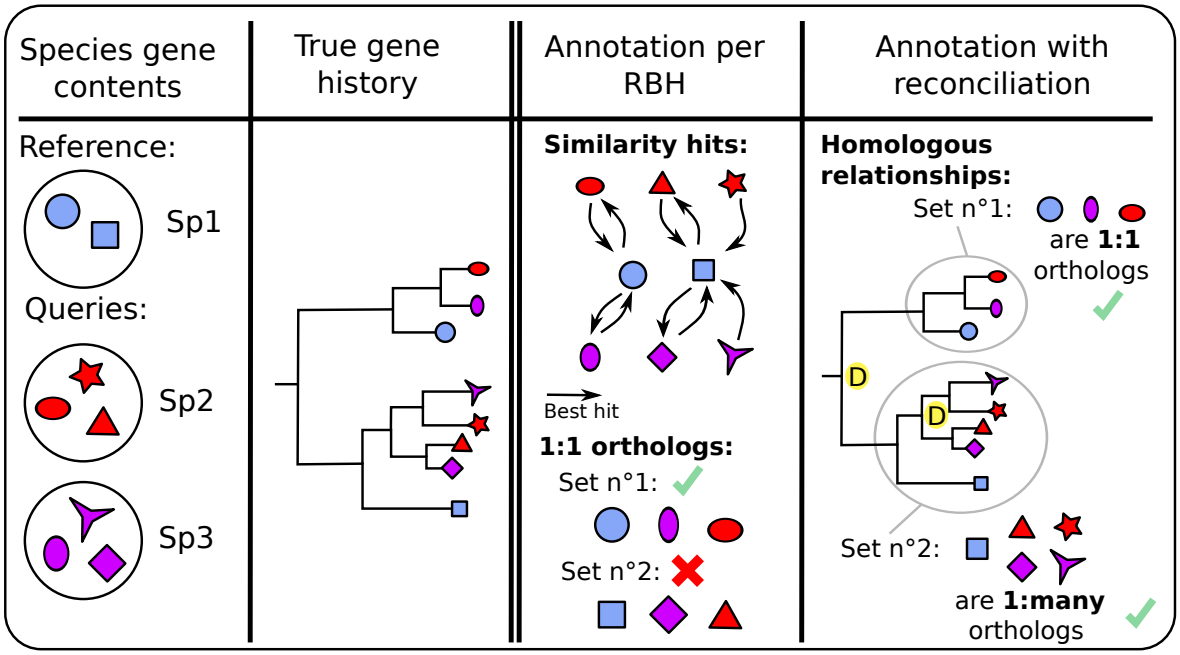

FIGURE 2.2 – Comparaison d'une annotation par Reciprocal Best Hit (RBH) et par réconciliation comme utilisée dans CAARS. De gauche à droite sont représentés, le contenu en gènes de 3 espèces représentés par des formes géométriques, l'histoire réelle de cette famille de gènes qui comprend une duplication à la racine et une duplication dans la partie inférieure de l'arbre chez l'ancêtre commun des espèces 2 (rouge) et 3 (violette), les résultats des meilleurs hits basés sur la similarité entre les gènes qui sont ensuite utilisés pour l'annotation par RBH en utilisant l'espèce 1 (bleu) comme référence. Et enfin, à droite, les résultats de l'annotation par RBH ou par réconciliation. Dans les cas de l'annotation par RBH, le set n°2 est faux car il ne regroupe pas des orthologies 1 :1 dans le sens où il contient des paralogues. Dans le cas d'une annotation par réconciliation, les séquences des gènes sont utilisées pour faire un arbre de gènes qui est ensuite réconcilié avec l'arbre d'espèces pour annoter les duplications (lettre D sur fond jaune) et donc annoter les sets de gènes ainsi que leur relation d'homologie.

L'utilisation du RBH implique que tous les gènes qui ont été perdus ou dupliqués de manière différentielle entre l'espèce étudiée et l'espèce de référence sont exclus de l'analyse ou bien mal annotés comme ici dans l'exemple dans le cas du set n°2 (Figure 2.2).

Pourtant les paralogues sont une source d'innovation importante (néo fonctionnalisation, sub fonctionnalisation (THOMPSON et collab., 2013)) et je trouvais regrettable de les exclure. Dans CAARS, l'annotation par RBH est remplacée par une annotation basée sur des arbres de gènes, ce qui permet de conserver un plus grand nombre de paralogues et surtout de les annoter comme tels.

2.1.3 Pouvoir améliorer de manière significative les assemblages

La dernière motivation pour développer CAARS est d'inclure un programme que j'ai implémenté, apytram, et qui permet d'améliorer les assemblages (Figure 2.3). Le fonctionnement d'apytram est décrit dans l'article de CAARS présenté ci-après mais je vais en reprendre ici les grandes lignes. Apytram a pour but de parfaire l'assemblage d'un gène basé sur le principe de TRAM ((JOHNSON et collab., 2013), Target Restricted Assembly Method). Pour cela, un processus itératif en trois étapes est mis en place (Figure 2.3).

La première étape est la récupération des lectures RNA-Seq qui ont des fortes similarités avec une séquence appât. Ces lectures sont ensuite assemblées. Les contigs issus de cet assemblage sont ensuite filtrés pour conserver uniquement les contigs homologues avec la séquence appât et ainsi éviter de reconstruire des séquences partageant une similarité réduite (comme par exemple, un motif structural tel que les doigts de zinc). Les contigs qui ont ensuite passé ce filtre sont utilisés comme nouvelles séquences appâts. Ce processus s'arrête lorsqu'il n'y a plus d'amélioration (Figure 2.3). Dans CAARS, le processus est initialisé avec les séquences des espèces de référence de la famille étudiée ainsi que les contigs provenant de l'assemblage global. Cette méthode permet donc d'étendre les séquences présentes dans l'assemblage global en récupérant des lectures qui n'avaient pas été intégrées.

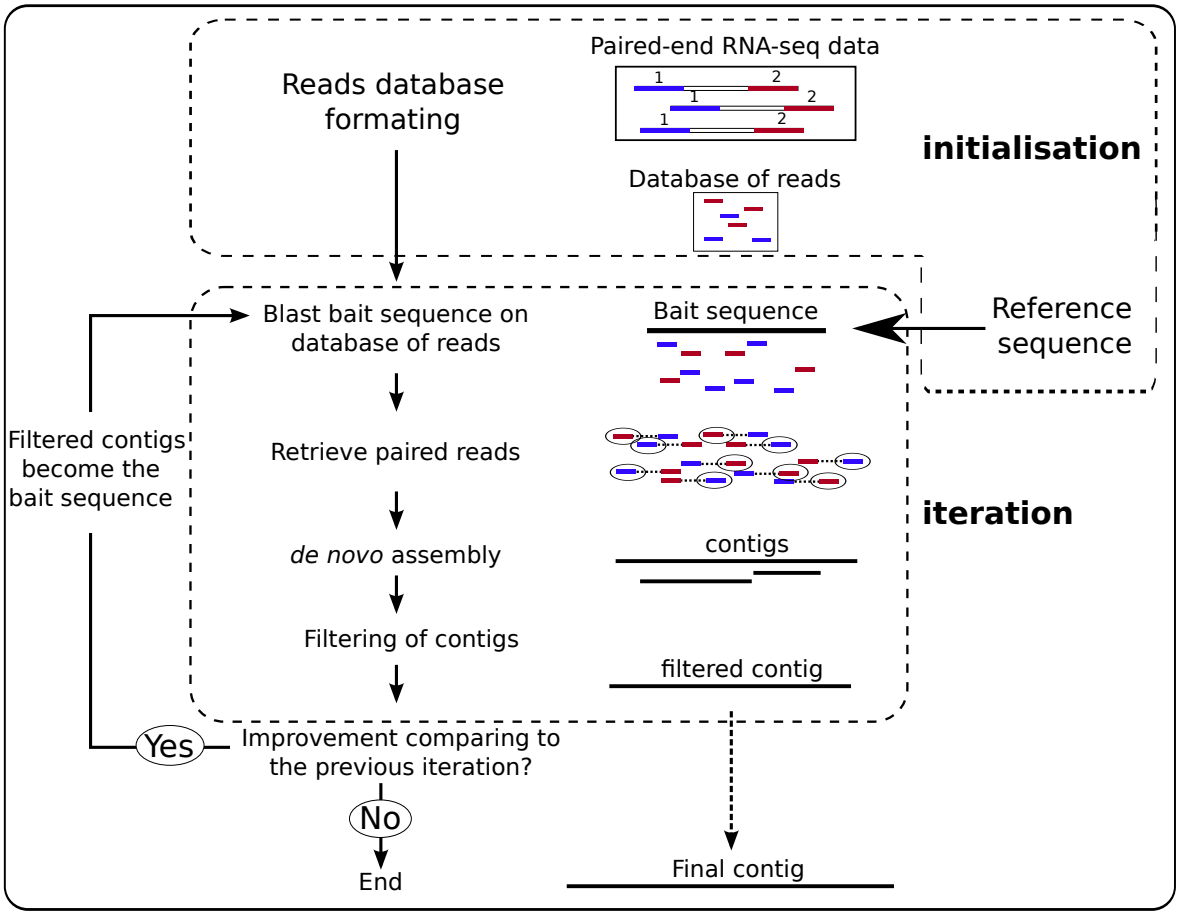

FIGURE 2.3 – Fonctionnement simplifié d'Apytram.

L'étape d'amélioration des assemblages par apytram ajoute de la complexité au cours de la

construction du jeu de données. Le gestionnaire de pipeline compris dans CAARS permet de masquer totalement cette complexité à l'utilisateur. Sans ce gestionnaire, il deviendrait compliqué d'utiliser apytram à grande échelle.

En conclusion, CAARS a été conçu dans l'objectif de faciliter la création de jeux des données pour la génomique comparative et d'en améliorer la qualité.

2.2 CAARS : un pipeline automatisé pour la création de jeux de données pour des analyses de génomique comparative

2.2.1 Avant-propos

L'article de publication de CAARS, qui est incorporé ci-après, présente le fonctionnement de CAARS ainsi qu'une mise en situation de CAARS où l'on construit un jeu de données utilisable pour des analyses de génomique comparative.

Cette mise en situation a consisté à se mettre dans un cas réel où l'on cherchait à construire un jeu de données pour des espèces non-modèles. Or, il n'y a pas de moyen de vérifier la qualité du jeu de données pour des espèces non-modèles. C'est pour cela que nous avons choisi d'utiliser des espèces modèles, la souris et l'épinoche, mais en les traitant comme des espèces non-modèles afin de pouvoir comparer les résultats de CAARS par rapport à leur annotation réelle qui est de très bonne qualité. De plus, nous avons comparé les performances de CAARS par rapport au pipeline plus ou moins standard trouvé dans la littérature et que nous avons essayé de reproduire.

Le but de cette mise en situation fictive était de montrer que CAARS était efficace par rapport à un pipeline standard et d'autre part qu'il fonctionnait même lorsque l'ensemble des espèces étudiées représentait une diversité phylogénétique conséquente et donc qu'il était potentiellement utilisable sur de nombreux cas biologiques. .

2.2.2 Article : CAARS : comparative assembly and annotation of RNA-Seq data

L'implémentation de CAARS et la réalisation de la publication associée à CAARS est un travail que j'ai réalisé avec la collaboration de Philippe Veber, Bastien Boussau et Marie Sémon. CAARS a été publié dans Bioinformatics, le 19 Novembre 2018 (https://doi.org/10.1093/bioinformatics/bty903). Il s'agit d'un logiciel libre hébergé et disponible sur github (https://github.com/CarineRey/caars). Un tutoriel ainsi que des précisions sur son implémentation sont disponibles dans une page dédiée (https://github.com/CarineRey/caars/wiki).

Bioinformatics, 2019, 1–9
doi: 10.1093/bioinformatics/bty903
Advance Access Publication Date: 19 November 2018
Original Paper

OXFORD

Sequence analysis

CAARS: comparative assembly and annotation of RNA-Seq data

Carine Rey[1,*], Philippe Veber[2], Bastien Boussau[2,†] and Marie Sémon[1,†]

[1]UnivLyon, Université Claude Bernard Lyon 1, ENS de Lyon, CNRS UMR 5239, INSERM U1210, LBMC, F-69007, Lyon, France and [2]UnivLyon, Université Claude Bernard Lyon 1, CNRS, UMR 5588, LBBE, F-69100, Villeurbanne, France

*To whom correspondence should be addressed.
†The authors wish it to be known that, in their opinion, the last two authors should be regarded as Joint Last Authors.
Associate Editor: Janet Kelso

Received on June 28, 2017; revised on September 13, 2018; editorial decision on October 10, 2018; accepted on November 16, 2018

Abstract

Motivation: RNA sequencing (RNA-Seq) is a widely used approach to obtain transcript sequences in non-model organisms, notably for performing comparative analyses. However, current bioinformatic pipelines do not take full advantage of pre-existing reference data in related species for improving RNA-Seq assembly, annotation and gene family reconstruction.
Results: We built an automated pipeline named CAARS to combine novel data from RNA-Seq experiments with existing multi-species gene family alignments. RNA-Seq reads are assembled into transcripts by both *de novo* and assisted assemblies. Then, CAARS incorporates transcripts into gene families, builds gene alignments and trees and uses phylogenetic information to classify the genes as orthologs and paralogs of existing genes. We used CAARS to assemble and annotate RNA-Seq data in rodents and fishes using distantly related genomes as reference, a difficult case for this kind of analysis. We showed CAARS assemblies are more complete and accurate than those assembled by a standard pipeline consisting of *de novo* assembly coupled with annotation by sequence similarity on a guide species. In addition to annotated transcripts, CAARS provides gene family alignments and trees, annotated with orthology relationships, directly usable for downstream comparative analyses.
Availability and implementation: CAARS is implemented in Python and Ocaml and is freely available at https://github.com/carinerey/caars.
Contact: carine.rey@ens-lyon.org
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Large scale RNA sequencing (RNA-Seq) is often used in non-model species as a pragmatic alternative to genome sequencing, in particular for comparative analyses (Ozsolak and Milos, 2011; Todd *et al.*, 2016; Wang *et al.*, 2009). However, the assembly of short reads from transcriptome assays into full length transcript sequences poses difficult issues related to repeated regions, variable expression levels, alternative splicing, sequencing errors and composition biases (Garber *et al.*, 2011). Further, the clustering of those sequences into gene families, their alignment and the step of gene tree reconstruction all represent challenges that studies of comparative genomics face without agreed-upon standards.

Different strategies can be used for transcript assembly, depending on the existence of genomic data for closely related species (Conesa *et al.*, 2016; Ockendon *et al.*, 2016). If no sister species with a sequenced genome is available, reads are assembled *de novo* based on overlapping sequences [e.g. Trinity, Grabherr *et al.* (2011)]. Otherwise, genome-guided assembly may be used [e.g. Tophat, Trapnell *et al.* (2009) and Cufflinks, Trapnell *et al.* (2010)]. In that case, reads are aligned to this guide genome, creating clusters

of reads that are used for local transcript assembly. This strategy is obviously restricted to very closely related species, for which trans-species read mapping is feasible. On more distantly related species, no approach has been proposed for RNA-Seq assembly, but developments have been made for genome assembly. In particular, the Target Restricted Assembly Method (TRAM) by Johnson *et al.* (2013), automated in aTRAM (Allen *et al.*, 2015), reconstructs a gene sequence by an iterative process where reads are collected by sequence similarity to a reference genome using BLAST (Camacho *et al.*, 2009) and then assembled. A different implementation was proposed in Kollector (Kucuk *et al.*, 2017) based on a *k*-mers approach. These methods show encouraging results, but have not been designed to be used on RNA-Seq data and for thousands of genes at a time.

After assembly, transcripts should ideally be annotated with a gene name. Commonly, transcriptome annotation is based on sequence similarity between the transcripts and the transcriptome of already annotated species. This step is most often treated by Reciprocal Best Hits (RBHs) (Rivera *et al.*, 1998), typically using BLAST (Camacho *et al.*, 2009), which cannot handle species–specific duplications (Altenhoff and Dessimoz, 2009; Tekaia, 2016). This is an issue because many genes are duplicated. For instance, in the Ensembl database (Herrero *et al.*, 2016; Yates *et al.*, 2016) 10% of all Human genes have no one-to-one orthology relationships with mouse genes.

In principle, relying on gene phylogenies instead of RBH for annotation allows handling complex homology relationships (Chen *et al.*, 2007; Kristensen *et al.*, 2011; Kuzniar *et al.*, 2008; Tekaia, 2016). We suggest to take such an approach: genes from annotated transcriptomes are clustered into homologous gene families either *de novo* (Kristensen *et al.*, 2011) or using existing families [EnsemblCompara (Herrero *et al.*, 2016), TreeFam (Finn *et al.*, 2014), Hogenom (Penel *et al.*, 2009), PhylomeDB (Huerta-Cepas *et al.*, 2014)]. Then, reconstructed transcripts are integrated into these gene families based on sequence similarity. Alignments and trees are reconstructed for these enlarged gene families. Quality of the trees can be improved by using reconstruction methods that use the information provided by the species tree (Boussau *et al.*, 2013; Ullah *et al.*, 2015). Finally, gene trees are reconciled with a species tree to annotate speciations, duplications and losses (Kristensen *et al.*, 2011). Based on this scenario of gene family evolution, orthology and paralogy relationships are derived, and gene names are propagated from annotated sequences to novel transcripts (Kristensen *et al.*, 2011). In this approach, accurate annotations are an outcome of accurate gene trees.

Here, we present an automated tool, named CAARS, to assemble and annotate the whole transcriptome of non-model organisms from RNA-Seq data, using sequences from one or several species that can be closely or distantly related to guide transcript assembly and annotation. CAARS relies on reference gene alignments and outputs homologous gene sets with high quality phylogenetic trees and orthology relationships that can directly be used for downstream comparative analyses. CAARS improves upon a well-established pipeline in terms of transcriptome completeness, transcript accuracy and annotation accuracy. Thanks to its high quality output gene trees, CAARS also improves upon Ensembl Compara in terms of the number of orthologs it can recover.

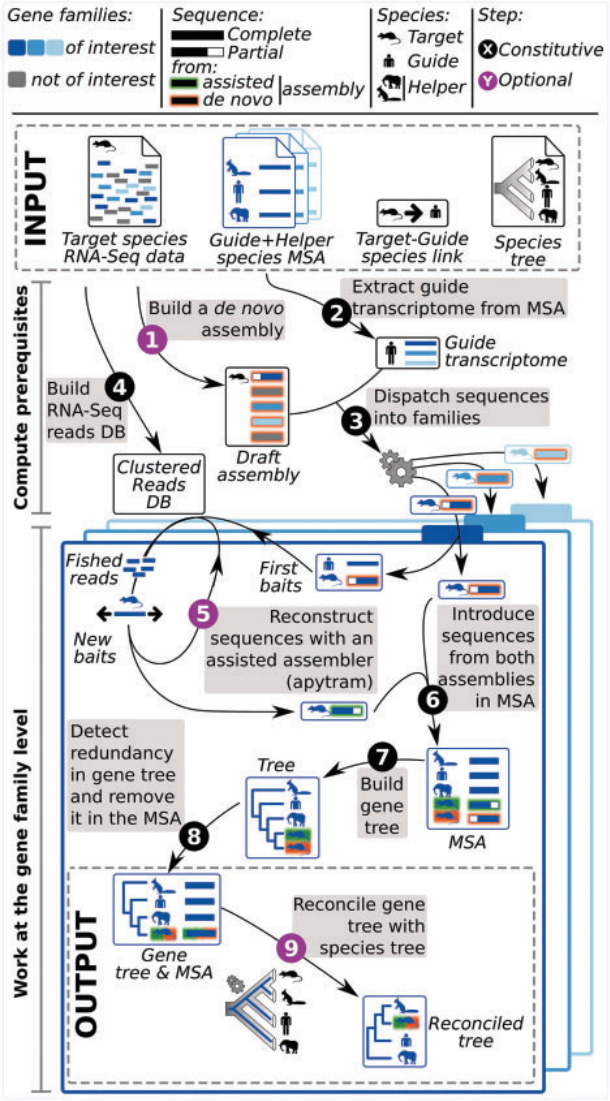

Fig. 1. CAARS overview. Representation of the major steps of CAARS. Steps 1–4 group pre-requisite computations. (**1**) If no draft transcriptome is given in input, RNA-Seq data are *de novo* assembled into a draft transcriptome and coding sequences are parsed to remove 5′ and 3′ UTR. (**2**) Transcriptomes from guide species are extracted from input MSAs to form guide transcriptomes. (**3**) Transcripts from the draft transcriptome are associated to the corresponding gene families by BH against guide transcriptomes. Steps 5–10 group computations made for each family. (**4**) RNA-Seq reads are clustered and formated into a database. (**5**) Transcripts are assembled again with an assisted and iterative method (Apytram). At the first iteration, genes from the guide species and target transcripts from the draft transcriptome corresponding to this family are used as bait sequences to fish reads in the reads database. Mate reads are used to enlarge this batch of reads. Reads are then *de novo* assembled, and a new iteration can begin with the reconstructed sequences as baits. (**6**) Coding sequences from both assemblies are added to the existing gene family alignments. (**7**) A primary gene tree is obtained for the family. (**8**) Redundancy is removed by selecting the longest sequence or by merging sequences from the same species when appropriate. Then, sequences (from target species) with a low-scoring alignment to their sister sequences (from guide or helper species) can be discarded (not shown). A gene tree is re-computed to take into account potential changes. (**9**) The species tree and the gene family tree are used jointly to infer a reconciled gene tree placing gene losses and duplications along the gene family tree

2 Materials and methods

2.1 Outline of CAARS and implementation

The general structure of CAARS is illustrated in Figure 1. As input CAARS requires data from three types of species: the species whose transcriptomes need to be assembled, which we call target species, the species with transcriptomes serving as guide for assembly and annotation, which we call guide species and the species with already assembled transcriptomes because they improve the resolution of

the gene trees in the pipeline, which we call helper species. More particularly, CAARS requires Multiple Sequence Alignments (MSAs) corresponding to gene families containing sequences from guide and helper species, a rooted species tree with all the species and RNA-Seq data for the target species. Finally, we require to specify a set of guide species for each target species (Fig. 1 top): since target species may belong to various taxonomic groups, it may be useful to adapt guide species to each target species.

CAARS is organized in two major parts. The first one sets up several pre-requisites for the second, which is the execution in parallel of a series of steps for each gene family.

First, CAARS performs a *de novo* assembly using the commonly used program Trinity (Grabherr *et al.*, 2011). The *de novo* assembled transcripts are dispatched to gene families using BLAST. In addition, RNA-Seq reads are formatted as BLAST databases, one per target species.

Second, independently for each gene family, CAARS performs another assembly assisted by sequences from the guide species and by *de novo* assembled transcripts. This latter assembly is performed by our in-house software Apytram (Supplementary Fig. S1) (Rey *et al.*, 2017), a multi-species implementation of the TRAM algorithm (Johnson *et al.*, 2013). Importantly Apytram is able to deal with several RNA-Seq samples simultaneously, which improves on the initial implementation (Allen *et al.*, 2015).

Coding regions of transcripts from both *de novo* and assisted assemblies are extracted using Transdecoder (v3.0.1) (http://transdecoder.github.io) and then integrated into the MSAs. At this step, gene families typically include redundant transcripts, which can be alternative transcripts of the same gene at the same locus, or identical transcripts that have been assembled independently by the two methods.

To remove this redundancy, the default option is to select the longest sequence (raw or aligned length), following Yang and Smith (2014). Alternatively, it is possible to merge transcripts from the same species that branch at the same position in the tree, by maximizing the information content in the alignments (Supplementary Fig. S3). Partial sequences may be filtered out based on their alignment to their sister sequence in the tree (see Section 3 or the detailed implementation on the CAARS website).

Then, accurate gene trees are inferred using a phylogenetic pipeline that uses the information coming from both the alignments and the species tree (Boussau *et al.*, 2013) and identifies events of gene duplication and loss in a reconciliation step. Orthology and paralogy relationships are naturally deduced from the reconciled gene trees. Because CAARS grounds assembly and annotation on several guide species at the same time, and not on a single one, it is robust to species-specific gene duplications or losses in the guide transcriptomes.

The method and implementation are detailed in the Supplementary Material provided on the CAARS website. CAARS is written in the Python programing language for all intermediate steps and in the OCaml language for the main program orchestrating all computational steps. This program relies on bistro (Veber, 2017), an OCaml library that manages dependencies between steps, distributed computation and recovery upon error [also known as resume-on-failure ability (Leipzig, 2016)]. For ease of deployment, users do not need to install all external dependencies and may instead use the dedicated docker image available on DockerHub called *carinerey/caars*.

2.2 Assessing CAARS performance

We selected the Human as a guide species for assembling mouse and stickleback transcriptomes. We used their annotated transcriptomes as reference against which to compare the performance of CAARS and of a standard pipeline commonly used for *de novo* transcriptome assembly and annotation. In the following, to avoid ambiguities, we will use 'guide' to name the transcriptomes used to help the assemblies and 'reference' to name the transcriptomes used to benchmark the assemblies. In the standard pipeline, used e.g. in Marra *et al.* (2014); Konczal *et al.* (2014); Pereira *et al.* (2016); Thompson and Ortí (2016) and Ishikawa *et al.* (2016), the assembly is performed *de novo* by Trinity (Grabherr *et al.*, 2011) and the annotation by RBH using BLAST (Camacho *et al.*, 2009).

2.2.1 Dataset, common inputs

We used paired-end RNA-Seq libraries from adult mouse kidneys (2×51 bp, about 12.5–15.3 million reads per library, SRR636916, SRR636917, SRR636918) and from adult stickleback kidneys (2×100 bp, about 16.5 million reads per library, SRR528539, SRR528540).

2.2.2 CAARS additional inputs and assembly

In addition to the reads for the target species and an annotated transcriptome for the guide species, CAARS also requires MSAs corresponding to gene families containing sequences from guide and helper species. We downloaded the Ensembl Compara dataset from the Ensembl database [release 91, Herrero *et al.* (2016); Yates *et al.* (2016)]. This dataset contains MSAs for 22 340 gene families with sequences from 97 chordates, including 71 mammals. From this set of species, we extracted a subset of 17 species of which 13 representative mammals, 1 bird, 1 reptile and 2 fishes. We obtained their phylogeny from the Ensembl Github repository (Herrero *et al.*, 2016; Yates *et al.*, 2016) (Supplementary Fig. S2). We did not want to favor CAARS during the tests, and we voluntarily removed the rodents belonging to the mouse sub-order and the fishes belonging to the stickleback order.

To be usable in CAARS, MSAs must contain at least one sequence from the guide species (Human), and at least 3 species in total (for the reconciliation step). A total of 8622 MSAs satisfy both criteria.

We launched CAARS on a Linux server (16 threads, 64 G RAM) with a running installation of Docker and an imported CAARS image from DockerHub (*carinerey/caars*). CAARS tutorial contains the material (dataset and scripts) to replicate the analysis presented here as a demo and can be found on the wiki page of the Github repository. A smaller data set is also provided for a quick test.

2.2.3 Additional inputs for the standard pipeline and assembly

On the same hardware, for each target species, we first assembled RNA-Seq reads into transcripts with Trinity (Grabherr *et al.*, 2011) using default parameters. Then, we used Cap3 (Huang and Madan, 1999) (default parameters) to assemble overlapping Trinity contigs. We removed redundancy using CD-HIT-EST (Fu *et al.*, 2012), which clusters nucleotide sequences that meet a sequence identity threshold (99%) and finds one representative sequence per cluster (-c 0.99 -n 11 -d 0). We then used Transdecoder (v3.0.1, –retain_long_orfs 150) to extract the coding region of each transcript. Finally, we retained only transcripts associated by RBH with a guide species transcript using BLAST (Camacho *et al.*, 2009) with an *e*-value of $1e-6$ and using `blastn` as task option. In order to assess the impact of the evolutionary distance between the guide and the target species, we run this final step using two different guide species for each target species.

2.2.4 Reference transcriptome

To assess the accuracy of the assemblies and annotations made by CAARS or the standard pipeline, we compared them for each target species with their corresponding known transcriptome. We extracted mouse and stickleback sequences from the Ensembl Compara dataset (v91), as reference transcriptomes (Herrero *et al.*, 2016; Yates *et al.*, 2016). They are composed of 22 388 coding DNA sequences (CDSs) for the mouse and 20 072 for the stickleback distributed in 10 350 gene families. We removed strictly identical sequences using CD-HIT-EST (Fu *et al.*, 2012), keeping respectively 22 060 and 20 020 sequences.

Of note, the mouse and the stickleback sequences are distributed in more families (10 350) than were used as input for CAARS (8622) because CAARS needs Human homologous sequences as bait sequences. So, because they have no homolog in Human, in this intentionally difficult test, CAARS cannot find 1822 mouse and 1766 stickleback sequences. In a real-life situation, users can use more closely related genomes when available, or can use multiple genomes as bait sequences. This would drastically reduce the number of genes without homologs.

2.2.5 Sensitivity measure

We compared the completeness of each CAARS and standard assemblies with respect to the corresponding reference transcriptome. For each gene of the reference transcriptome and each assembly, we retrieved the RBH sequence when available, using BLAST with a stringent *e*-value threshold ($1e-10$) and otherwise default parameters (Camacho *et al.*, 2009). In the case where no RBH was found, we considered that this gene was missing from the assembly. Completeness statistics are provided in the Table 1.

2.2.6 Identification of partial and alternative transcripts

Assembled transcripts may be incomplete compared to the reference transcriptome, because they represent shorter alternative transcripts, or because the coverage for this transcript in the kidney expression data is low. To identify these partial and alternative transcripts, we aligned each transcript of each assembly to the reference transcriptome [BLAST (Camacho *et al.*, 2009) with `evalue=1e-10`, and otherwise default parameters], and retrieved the Best Hit sequence (BH). By using BH instead of RBH, we allow that several sequences in a given assembly match a single transcript of the reference transcriptome. This ensures all sequences of an assembly with a possible hit have an associated reference sequence.

2.3 Overlap between reconstructed and reference transcript sequences

We analyzed the sequences of transcripts present in both the reference transcriptome and a reconstructed assembly to estimate whether reconstructed transcripts are longer, shorter or generally different from the reference transcripts. To this end we computed two indices, for a given reference sequence R of length len_R and a given query sequence Q of length len_Q. We used Mafft (Katoh *et al.*, 2002) with default parameters to align the R and Q sequences and we computed the number of aligned positions, len_{ali}, between R and Q (without gaps) in the alignment.

The two indices are then calculated using these formula: $P_{\text{reference}} = \frac{len_{\text{ali}}}{len_R} \times 100$ and $P_{\text{query}} = \frac{len_{\text{ali}}}{len_Q} \times 100$.

2.4 Quantification of expression levels

For each target species, we quantified the levels of gene expression for the three transcriptomes (the reference transcriptome and the CAARS and standard assemblies) using Kallisto (Bray *et al.*, 2016) and the RNA-Seq libraries mentioned earlier.

2.5 Evaluation of sets of orthologs

Orthologs predicted by CAARS were extracted for different sets of species, including or not including target species. We compared the set of orthologs obtained without target species to the 'high confidence' orthologs available on the Ensembl Compara database.

3 Results

We used CAARS to assemble and annotate transcriptomes of target species using a guide species too divergent for a genome-guided assembly, and we compared it with a standard pipeline combining *de novo* assembly and annotation by RBH. We selected two target species, the mouse and the stickleback, for which gene sequences and annotations are well-established. Kidney RNA-Seq libraries of these two species were used for assemblies, and their genomes were used later to evaluate the accuracy of CAARS. We used the Human as

Table 1. Statistics of the CAARS assembly compared to a standard assembly

Target species	Assembly method	Options	Guide or reference species	Divergence (in Mya)	# of seqs in the assembly	Precision: # of seqs associated with a seq in the target species	Sensitivity: % of seqs of the target species associated with a seq in the assembly[a] (%)
Mouse	CAARS	by/By default	Human	90[b]	12 421	11 500 (92.6%)	88.1
		Filter at 25%	Human	90[b]	11 093	10 779 (97.2%)	82.6
	Standard pipeline	Filter based on RBH	Human	90[b]	10 808	10 749 (99.5%)	82.4
			Guinea pig	70[c]	10 572	10 496 (99.3%)	80.5
			Squirrel	70[c]	10 594	10 511 (99.2%)	80.6
Stickleback	CAARS	by/By default	Human	400[b]	10 878	9570 (88.0%)	66.7
		Filter at 25%	Human	400[b]	9789	8931 (91.2%)	62.2
	Standard pipeline	Filter based on RBH	Human	400[b]	8758	8189 (93.5%)	57.1
			Zebrafish	225[d]	11 273	10 483 (93.0%)	73.1

[a]This is calculated as the ratio of the number of target reference sequences with an associated sequence in the assembly over the number of target reference sequences expressed more than 1 count per base in the library (for the Mouse, 13 046 seqs, and for the Stickleback, 14 349 seqs).

[b]Hedges *et al.* (2015).

[c]Fabre *et al.* (2012).

[d]Betancur-R *et al.* (2015).

guide species because it is well-annotated and also because it is quite divergent from the mouse and the stickleback [divergence around 90 million years ago—Mya—and 400Mya (Hedges *et al.*, 2015)], far too distant for trans-species mapping (Conesa *et al.*, 2016; Ockendon *et al.*, 2016; Torres-Oliva *et al.*, 2016). Overall we conservatively chose unfavorable test settings to assess the performance of CAARS. To this end, first we chose a single distant guide species to assemble target transcriptomes. Second, we also chose a Teleost fish as one of our target species because it contains many duplicate genes due to whole genome duplications, which makes the reconstruction difficult.

3.1 CAARS has a better sensitivity than a standard assembly pipeline

CAARS took 4 days to reconstruct 12 421 mouse CDSs and 10 878 stickleback CDSs included in 7049 families (Table 1). These figures are in accordance with the fact that about 10 000 ∼ 13 000 genes are expressed in Human kidney [10 000 using a threshold > 5 FPKM (Fagerberg *et al.*, 2014; Uhlen *et al.*, 2015).)]. This demonstrates that CAARS may be used with a distant guide species in a reasonable amount of time.

Independently, we ran a standard *de novo* assembly on the same data and the same hardware, which took about 4 h. After filtering transcripts using RBH, there is a large influence of the divergence between the guide and the target species on the accuracy of the standard pipeline (Table 1). For example, for the stickleback assembly, we obtained 8758 sequences using the Human as guide species, against 11 273 if we used zebrafish, a less divergent guide species. The genes only recovered with the zebrafish as a guide species probably mostly correspond to fish-specific gene families. To interpret the differences between the pipelines in terms of their power to detect transcripts rather than in terms of whether they can detect fish-specific gene families or not, we discuss in the following the assembly with the Human as guide for both the mouse and the stickleback and for both methods.

To examine the completeness of both assemblies, we associated by RBH the target species reference transcripts to the reconstructed sequences. Out of all 22 060 mouse reference sequences, 10 672 sequences were found by both assemblies, 828 sequences only by CAARS, 77 only by the standard assembly (Fig. 2A). Therefore, in this case, CAARS retrieves 7% more transcripts than a pipeline classically used for *de novo* assemblies.

For the stickleback, on a total of 20 020 sequences, 7687 sequences were found by both assemblies, 1883 sequences only by CAARS, 502 only by the standard assembly (Fig. 2B), meaning CAARS allows a gain of 17% of transcripts.

The sequence of genes not expressed in the kidney cannot be reconstructed from our RNA-Seq data. Conversely, the sequence of highly expressed genes should be easier to obtain. We wished to establish more systematically the link between the level of gene expression and the accuracy of sequence assembly. For the mouse, we measured gene expression levels in kidney using the reference transcriptome. As expected, very weakly expressed genes (≤1 count per base) are rarely retrieved by CAARS, or by the standard assembly.

If we focus only on genes expressed more than two counts per base pair (meaning with an average sequencing coverage ≥2×), CAARS detects more genes than the standard assembly. In the mouse, 93% of the reference sequences have an associated sequence in the CAARS assembly and 89% in the standard assembly (Fig. 2A). In the stickleback, the improvement of CAARS over the standard assembly is more pronounced: 66% compared to 59% (Fig. 2B).

As CAARS annotates transcripts using a phylogeny, it is expected to be more effective than the standard pipeline, which used a RBH annotation, to retrieve genes in the target species that have been duplicated since the divergence with the guide species (one-to-many or many-to-many orthologs in Ensembl Compara). Such orthologs correspond to 8% of the expressed genes (>1 count/pb) in the mouse and 30% in the stickleback. Expectedly, they are more often retrieved by CAARS than the standard pipeline, whether for the mouse, where 63.0% of them are retrieved compared to 57.5% in the standard pipeline, or for the stickleback (58.9 versus 46.6%).

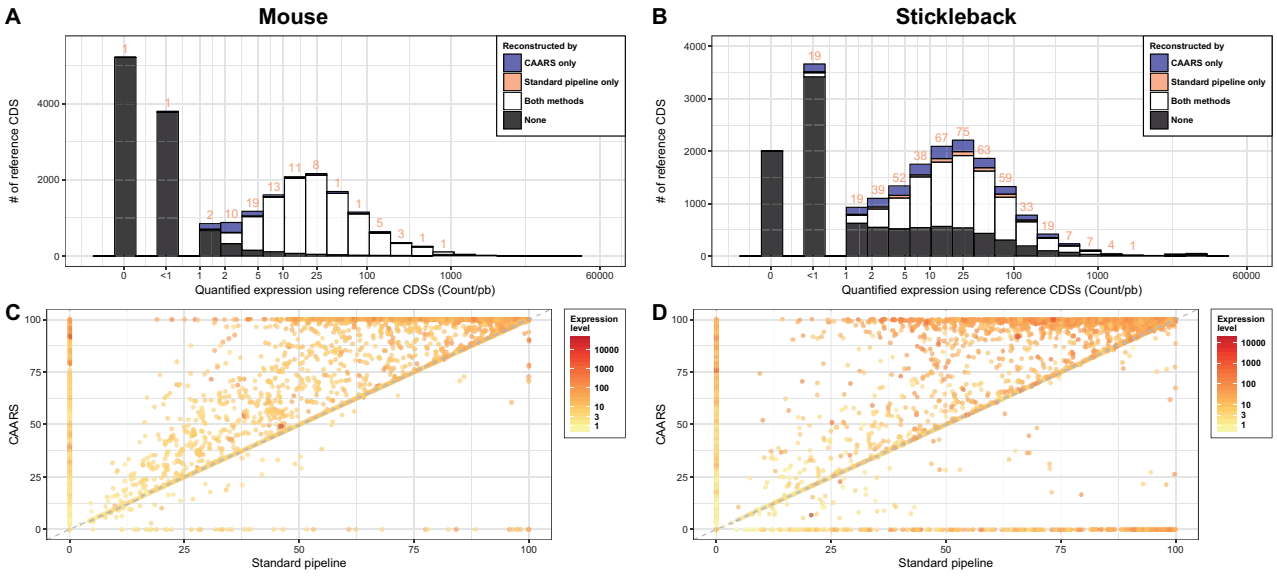

Fig. 2. Comparison between the CAARS and standard assemblies for the mouse and the sickleback. (**A, B**) Number of reference CDSs associated by a RBH with a transcript from CAARS assembly only, the standard assembly only, both assemblies or none. Expression was quantified on the reference transcriptome in Count per base; 0 means no detected expression. (**C, D**) Proportion of the reference transcript ($P_{reference}$) aligned with its CAARS assembly counterpart (*y* axis) or its standard assembly counterpart (*x* axis). Marginal distributions can be seen in Supplementary Figures S5A and S6A. Dots with a null value on one of the axes represent genes not recovered by the corresponding method. The results for CAARS are similar whether we use the filtering step or not

3.2 CAARS assembly provides more complete transcripts

The completeness of an assembly cannot be assessed solely on the basis of the number of recovered transcripts. We measured the coverage of each reference transcript, and compared the results between the two assemblies for each target species. Gene coverage was estimated on pairs of sequences with a transcript from the reference transcriptome and the corresponding reconstructed transcript. We computed the percentage of the reference transcript that aligns with the reconstructed transcript ($P_{\text{reference}}$) for 11 577 Ensembl CDS with a matched sequence in at least one of the assemblies (Fig. 2C). Both for the mouse and the stickleback, CAARS better recovers the reference transcripts. In the experiment on mouse libraries, the percentage of reconstruction of reference transcripts is identical for 8937 sequences, better in the CAARS assembly for 2509 sequences (with an average increase of 24.8%) and better in the standard assembly for 131 sequences (smaller improvement, 10.4%) (Fig. 2C). For the stickleback, 3813 sequences (out of 10 072) are better in the CAARS assembly (26.5% average increase) against 566 for the standard pipeline (12.7% average increase) (Fig. 2D).

CAARS transcripts are longer than those from the standard assembly but they are also more often complete. Total of 8134 CAARS mouse transcripts are complete or sub-complete ($P_{\text{reference}} > 95$), which is better than the 7246 sub-complete transcripts obtained with the standard pipeline and respectively 6465 compared to 4853 for the stickleback transcripts (Table 2).

A potential issue for assembled transcripts is the merging of two transcripts into a chimeric sequence. Such chimeric sequences will be longer than the reference CDSs and characterized by a low P_{query} value. For the mouse and the stickleback, distributions are similar for CAARS and the standard pipeline, with no excess of low values (Supplementary Figs S5 and S6), meaning that the potential numbers of chimeric sequences found in the CAARS assembly and in the standard pipeline are small.

We estimated the quality of the assembly using another criterion, gene expression levels. The levels of expression obtained from both RNA-Seq assemblies are well correlated to reference expression levels ($R^2 = 0.98$) (Supplementary Fig. S4A and B).

3.3 An optional filter can discard redundant and low quality sequences

CAARS assemblies have a better sensitivity than the standard pipeline but, with default parameters, the precision, the number of sequences associated by RBH with a reference transcript, is lower. Among sequences without RBH, most (914, ie. 92%) have a unidirectional blast hit in the reference transcriptome. Besides, they have a low $P_{\text{reference}}$ (Supplementary Fig. S5A left) and a high P_{query} (Supplementary Fig. S5B left), meaning they are partial assemblies or assemblies of small redundant transcripts.

Table 2. Comparison of the alignment statistics on reference genes of the CAARS assembly and a standard assembly using the Human as guide species

Assembly method	# of sub-complete CDSs[a]	
	Mouse	Stickleback
CAARS (by default)	8134 (65.5%)	6465 (59.4%)
CAARS (with filter at 25%)	8131 (73.3%)	6457 (66.0%)
Standard pipeline	7246 (67.0%)	4853 (55.4%)

[a]A CDS is counted as sub-complete if its $P_{reference}$ is superior to 95, ie. it covers at least 95% of its reference CDS. The proportion of sub-complete transcripts is obtained by dividing the number of sub-complete transcripts by the total number of transcripts predicted by the method.

Partial CDSs can introduce noise in subsequent analyses and users may wish to flag them. We reasoned that transcript length should not vary too much among related species. Hence, transcripts assembled by CAARS with a length similar to the length of their sister species in the tree are expected to be complete. We verified that, indeed, the percentage of reconstruction of the neighborhood sequence in the gene tree is a good proxy for the percentage of reconstruction of the reference sequence (Supplementary Figs S5B and S6B). An alternative would have been to filter them out based on their expression level but we have found that it is not correlated to the assembly quality (Supplementary Fig. S7A). A threshold can be applied on this criterion to discard partial CDSs or select high quality CDSs (Supplementary Fig. S7C).

For instance, with a threshold at 25%, CAARS puts aside 1328 (10.7%) sequences for the mouse and 1089 (10.0%) for the stickleback. This filter allows increasing the precision from 92.6 to 97.2% for the mouse and from 88.0 to 91.2% for the stickleback (Table 1). The filter increases the proportion of complete sequences in the assembly from 65.5 to 73.3% for the mouse and from 59.4 to 66.0% for the stickleback, which is better than the standard pipeline (Table 2). The sensitivity decreases a little but stays above the standard pipeline, so some sequences with a RBH have been discarded but these sequences are partial (low $P_{\text{reference}}$ and high P_{query}) (Supplementary Fig. S5C). The stringency of this filter can be set by the user when using CAARS.

3.4 CAARS produces sets of orthologs defined by phylogeny

CAARS returns MSA and reconciled gene trees ready to use for comparative analyses. From these reconciled gene trees, the user may use CAARS to infer orthology relationships between all sequences. This information is stored in a table which can be easily mined to retrieve all one-to-one orthologs for a given subset of species (Fig. 3). These subsets can include, or not, the target species. For example, we find 4850 sets of one-to-one orthologs with one gene per mammalian species of our dataset (Fig. 3). This is substantially more than the number found by the equivalent request in Ensembl Compara (4505 sets of genes with high confidence one-to-one orthology relationship), a gain attributable to our reconciliation step, which improves gene trees (Boussau *et al.*, 2013). We also extracted sets of orthologs for subset of species that include a target species, and obtained reasonable numbers (8705 for Human/mouse comparisons, 6435 for comparisons across rodents, 5 666 for comparisons across fishes Fig. 3). We cannot compare these numbers to numbers from the Compara database, because the mouse or stickleback gene complements are partial, being reconstructed from libraries of specific organs.

4 Discussion

CAARS is a pipeline that can be used to assemble transcriptomic data sets for comparative analyses. To assess its first steps, we compared CAARS with a standard pipeline for comparative assembly and annotation of RNA-Seq data. We found that CAARS is more sensitive since it finds more transcripts that are more complete. This better sensitivity is accompanied by a high precision (in particular with the optional filter), as a large majority of the sequences can be associated by RBH with a sequence of its reference transcriptome.

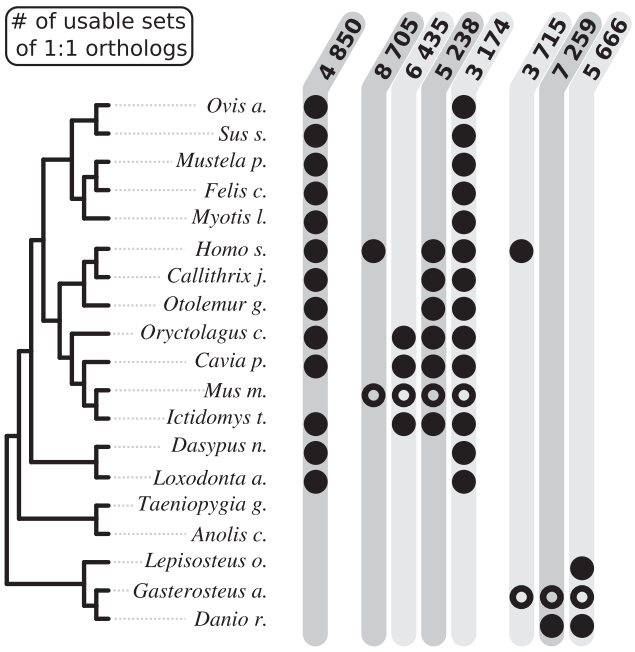

Fig. 3. CAARS outputs ready to use sets of one-to-one orthologs, inferred from reconciled phylogenies. Each column corresponds to the number of sets of one-to-one orthologs containing at least the species indicated by circles. Empty circles correspond to target species (reconstructed with RNA-Seq data), filled circles to guide or helper species (genome available)

Expression levels are at least as well estimated by CAARS as with the standard pipeline.

The improvement over the standard pipeline can be attributed to two novel features: (i) CAARS implements a trans-species assembly, based on one or several guide sequences, which can be distantly related. This is demonstrated here with mouse and stickleback RNA-Seq data assembled using a very distant Human guide species; (ii) CAARS annotates the transcripts by integrating them in phylogenies built with a set of helper sequences. This is demonstrated here with families from the Ensembl Compara database.

In addition to the steps of transcript detection and assembly, CAARS generates gene alignments, trees and sets of orthologs that can be directly used in subsequent comparative analyses. Notably, we found that it could recover more sets of orthologs than the Ensembl Compara pipeline and was better at recovering one-to-many or many-to-many orthologs than the standard pipeline, probably because it relies on gene trees reconstructed with a reconciliation approach. Besides, CAARS is easy to use, robust and modular.

4.1 CAARS uses one or several possibly divergent species to generate assemblies

In our test we used Human transcripts as a guide to assemble mouse and stickleback RNA-Seq data. Guide transcriptomes with good annotations are important as they reduce the likelihood that a gene has been mis-annotated or missed altogether.

In the case of the stickleback, the sensitivity of the CAARS assembly is better than that of the standard pipeline using the Human as guide species. It is not as good as the sensitivity of the standard pipeline using the zebrafish, but that is expected since the zebrafish is much more closely related to the stickleback than the Human is.

In the case of the mouse, very few transcripts are recovered by the standard pipeline only. For the stickleback, the number of transcripts found only by the standard pipeline is larger (see figures in red, Fig. 2B). This difference is due to a more stringent threshold inside CAARS (not shown). CAARS nonetheless clearly outperforms the standard pipeline in sensitivity in both cases (Fig. 2B).

It is not always easy to find well-annotated and closely related species that can serve as guide species. In many cases, well annotated genomes will be distant and closely related genomes of weaker quality. To improve the performance of guide-based assembly in those situations, CAARS can use several guide species at the same time. This reduces the likelihood that a gene is missed because of a missing bait, since it would have to be absent from all guide species. In a real study, where we want to optimize the result and not challenge CAARS we would add the zebrafish and the spotted gar as guide species to assemble the stickleback transcriptome and the squirrel and the guinea pig to assemble the mouse transcriptome.

In addition, several target species can be assembled at the same time. This can benefit the assembly of target species because during the step of assisted assembly (by Apytram), all the target species sharing the same guide species will participate and help each other in fishing the reads. This also allows breaking the distance between guide and target species.

4.2 CAARS integrates assembled transcripts into families and builds gene phylogenies

CAARS belongs to a small group of pipelines [e.g. Agalma, Dunn *et al.* (2013)] that explicitly aim at assembling data sets for phylogenomic analyses providing homologous and orthologous sequences, MSAs and gene trees from RNA-Seq data. However making use of one or several distant guide species at the same time, using phylogeny for annotation and providing sets of one-to-one orthologs are to our knowledge new features.

Other automatized methods that can assemble RNA-Seq data using closely related helper species [Agalma, Dunn *et al.* (2013), BRANCH, Bao *et al.* (2013) or FRAMA, Bens *et al.* (2016)] are based on direct sequence similarities (mapping or alignment on guide genome). However studies showed the negative correlation between annotation quality and divergence with guide species used for trans-species assembly/annotation (Ockendon *et al.*, 2016; Torres-Oliva *et al.*, 2016; Ungaro *et al.*, 2017; Vijay *et al.*, 2013). For instance, Vijay *et al.* (2013) recommends not to map directly on the guide genome when there is more than 15% of sequence divergence between the target and the guide species, which corresponds to the median nucleotide divergence in one-to-one orthologs between mouse and Human (Church *et al.*, 2009).

The phylogenetic framework used by CAARS allows identifying redundant transcripts that have been assembled more than once, and collapsing them or selecting the longest (the default). Gene trees are also used by CAARS to identify incomplete transcripts by comparison with neighboring sequences and filter them out (Supplementary Fig. S5). It remains to be seen how CAARS would behave on datasets containing lots of recent duplicates; in particular, the option to merge monophyletic transcripts from the same species may create chimeric transcripts and should be used knowingly.

A limitation of CAARS remains the usage in input of MSAs. However, nowadays, there are several public database containing such MSAs [Orthomam (Ranwez *et al.*, 2007), Hogenom (Penel *et al.*, 2009), PhylomeDB (Huerta-Cepas *et al.*, 2014), TreeFam (Finn *et al.*, 2014), EnsemblCompara (Herrero *et al.*, 2016)].

4.3 CAARS is robust and easy to install and use

CAARS is a complex pipeline combining existing software and newly developed programs, and is built to be able to analyze

thousands of gene families at once. In particular, CAARS includes Apytram (Rey *et al.*, 2017), a multi-species and more accurate re-implementation of TRAM (Johnson *et al.*, 2013), which initially introduced the idea of a trans-species assembler at the level of a single gene. For robustness and traceability, and to enable recovery upon error or iterative use, it uses the *bistro* library (Veber, 2017).

For an easy installation, we packaged CAARS into a Docker image, so that there is no need to install any dependency once Docker has been installed on the system, which can be a Mac, Windows or Linux system. Further, by using Docker we ensure that the intended versions of all tools of the pipeline will be used, irrespective of what is installed on the host machine system. The results produced by CAARS are thus fully reproducible. Finally, the use of Docker ensures only minimal penalties on computational efficiency. However, users can also opt to install the full pipeline without using Docker.

Here we demonstrated the use of CAARS at the whole transcriptome level but the program can be used at a much smaller scale, for a single gene family, or even for a single gene. CAARS may also be used for integrating transcripts obtained from a pre-existing assembly into gene families. This is easily feasible by switching off the step of assisted reconstruction in the input option file (explained in the tutorial on CAARS's website).

Although CAARS is slower than the standard pipeline, it provides not only the assembly and annotation, but also gene family alignments, reconciled genes trees and sets of orthologous genes. In many cases, such data may be used directly for subsequent analyses.

5 Conclusion

We have introduced CAARS, a new pipeline for the comparative assembly and annotation of transcripts in non-model species. Because it operates within a phylogenetic framework, it can use both closely related and distantly related species. In addition to annotated transcripts, it also provides gene family alignments and trees built using state-of-the-art methods, which can be directly used for downstream analyses. On data coming from the Ensembl database, it compared favorably to a pipeline combining Trinity and BLAST, and provided more complete sets of orthologs than Ensembl. CAARS could therefore be used in a variety of situations where transcript assembly needs to be of high quality, for instance for comparing gene expression or gene sequences across species.

Acknowledgements

We would like to thank T. Lorin, L. Taulelle, R. Allio for their contributions to testing and releasing of CAARS. We also thank the French Institute of Bioinformatics (IFB, ANR-11-INBS-0013) and the PSMN computing center of ENS de Lyon for providing storage and computing resources.

Funding

The research presented here was supported by the Convergenomix project [ANR-15-CE32-0005]. C.R. was supported by a PhD fellowship (CDSN) from the Ecole Normale Supérieure of Lyon.

Conflict of Interest: none declared.

References

Allen,J.M. *et al.* (2015) aTRAM - automated target restricted assembly method: a fast method for assembling loci across divergent taxa from next-generation sequencing data. *BMC Bioinformatics*, **16**, 98.

Altenhoff,A.M. and Dessimoz,C. (2009) Phylogenetic and functional assessment of orthologs inference projects and methods. *PLoS Comput. Biol.*, **5**, e1000262.

Bao,E. *et al.* (2013) BRANCH: boosting RNA-Seq assemblies with partial or related genomic sequences. *Bioinformatics*, **29**, 1250–1259.

Bens,M. *et al.* (2016) FRAMA: from RNA-seq data to annotated mRNA assemblies. *BMC Genomics*, **17**, 54.

Betancur-R,R. *et al.* (2015) Fossil-based comparative analyses reveal ancient marine ancestry erased by extinction in ray-finned fishes. *Ecol. Lett.*, **18**, 441–450.

Boussau,B. *et al.* (2013) Genome-scale coestimation of species and gene trees. *Genome Res.*, **23**, 323–330.

Bray,N.L. *et al.* (2016) Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.*, **34**, 525.

Camacho,C. *et al.* (2009) BLAST+: architecture and applications. *BMC Bioinformatics*, **10**, 421.

Chen,F. *et al.* (2007) Assessing performance of orthology detection strategies applied to eukaryotic genomes. *PLoS One*, **2**, e383.

Church,D.M. *et al.* (2009) Lineage-specific biology revealed by a finished genome assembly of the mouse. *PLoS Biol.*, **7**, e1000112.

Conesa,A. *et al.* (2016) A survey of best practices for RNA-seq data analysis. *Genome Biol.*, **17**, 13.

Dunn,C.W. *et al.* (2013) Agalma: an automated phylogenomics workflow. *BMC Bioinformatics*, **14**, 1–17.

Fabre,P.-H. *et al.* (2012) A glimpse on the pattern of rodent diversification: a phylogenetic approach. *BMC Evol. Biol.*, **12**, 88.

Fagerberg,L. *et al.* (2014) Analysis of the human tissue-specific expression by genome-wide integration of transcriptomics and antibody-based proteomics. *Mol. Cell. Proteomics*, **13**, 397–406.

Finn,R.D. *et al.* (2014) Pfam: the protein families database. *Nucleic Acids Res.*, **42**, D222–D230.

Fu,L. *et al.* (2012) CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics*, **28**, 3150–3152.

Garber,M. *et al.* (2011) Computational methods for transcriptome annotation and quantification using RNA-seq. *Nat. Methods*, **8**, 469–477.

Grabherr,M.G. *et al.* (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.*, **29**, 644–652.

Hedges,S.B. *et al.* (2015) Tree of life reveals clock-like speciation and diversification. *Mol. Biol. Evol.*, **32**, 835–845.

Herrero,J. *et al.* (2016) Ensembl comparative genomics resources. *Database (Oxford)*, **2016**, bav096.

Huang,X. and Madan,A. (1999) CAP3: a DNA sequence assembly program. *Genome Res.*, **9**, 868–877.

Huerta-Cepas,J. *et al.* (2014) PhylomeDB v4: zooming into the plurality of evolutionary histories of a genome. *Nucleic Acids Res.*, **42**, D897–D902.

Ishikawa,M. *et al.* (2016) Different endosymbiotic interactions in two hydra species reflect the evolutionary history of endosymbiosis. *Genome Biol. Evol.*, **8**, evw142.

Johnson,K.P. *et al.* (2013) Next-generation phylogenomics using a target restricted assembly method. *Mol. Phylogenetics Evol.*, **66**, 417–422.

Katoh,K. *et al.* (2002) Mafft: a novel method for rapid multiple sequence alignment based on fast fourier transform. *Nucleic Acids Res.*, **30**, 3059–3066.

Konczal,M. *et al.* (2014) Accuracy of allele frequency estimation using pooled RNA-Seq. *Mol. Ecol. Resour.*, **14**, 381–392.

Kristensen,D.M. *et al.* (2011) Computational methods for gene orthology inference. *Brief. Bioinform.*, **12**, 379–391.

Kucuk,E. *et al.* (2017) Kollector: transcript-informed, targeted de novo assembly of gene loci. *Bioinformatics*, **18**, 821–829.

Kuzniar,A. *et al.* (2008) The quest for orthologs: finding the corresponding gene across genomes. *Trends Genet.*, **24**, 539–551.

Leipzig,J. (2016) A review of bioinformatic pipeline frameworks. *Brief. Bioinform.*, **18**, 530–536.

Marra,N.J. *et al.* (2014) Natural selection and the genetic basis of osmoregulation in heteromyid rodents as revealed by RNA-seq. *Mol. Ecol.*, **23**, 2699–2711.

Ockendon,N.F. *et al.* (2016) Optimization of next-generation sequencing transcriptome annotation for species lacking sequenced genomes. *Mol. Ecol. Resour.*, **16**, 446–458.

Ozsolak,F. and Milos,P. (2011) RNA sequencing: advances, challenges and opportunities. *Nat. Rev. Genet.*, **12**, 87–98. Epub 2010 Dec 30. ST – RNA sequencing: adv.

Penel,S. *et al*. (2009) Databases of homologous gene families for comparative genomics. *BMC Bioinformatics*, **10** (Suppl. 6), S3.

Pereira,R.J. *et al*. (2016) Transcriptome-wide patterns of divergence during allopatric evolution. *Mol. Ecol.*, **25**, 1478–1493.

Ranwez,V. *et al*. (2007) Orthomam: a database of orthologous genomic markers for placental mammal phylogenetics. *BMC Evol. Biol.*, **7**, 241.

Rey,C. *et al*. (2017) apytram v1.1. *Zenodo*. (doi: 10.5281/zenodo.804416).

Rivera,M.C. *et al*. (1998) Genomic evidence for two functionally distinct gene classes. *Proc. Natl. Acad. Sci. USA*, **95**, 6239–6244.

Tekaia,F. (2016) Inferring orthologs: open questions and perspectives. *Genomics Insights*, **9**, 17–28.

Thompson,A.W. and Ortí,G. (2016) Annual Killifish transcriptomics and candidate genes for metazoan diapause. *Mol. Biol. Evol.*, **33**, 2391–2395.

Todd,E.V. *et al*. (2016) The power and promise of RNA-seq in ecology and evolution. *Mol. Ecol.*, **25**, 1224–1241.

Torres-Oliva,M. *et al*. (2016) A robust (re-)annotation approach to generate unbiased mapping references for RNA-seq-based analyses of differential expression across closely related species. *BMC Genomics*, **17**, 392.

Trapnell,C. *et al*. (2009) TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics*, **25**, 1105–1111.

Trapnell,C. *et al*. (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.*, **28**, 511–515.

Uhlen,M. *et al*. (2015) Tissue-based map of the human proteome. *Science*, **347**, 1260419–1260419.

Ullah,I. *et al*. (2015) Integrating sequence evolution into probabilistic orthology analysis. *Syst. Biol.*, **64**, 969–982.

Ungaro,A. *et al*. (2017) Challenges and advances for transcriptome assembly in non-model species. *PLoS One*, **12**, e0185020.

Veber,P. (2017) bistro v0.3.0. *Zenodo*. (doi: 10.5281/zenodo.815611).

Vijay,N. *et al*. (2013) Challenges and strategies in transcriptome assembly and differential gene expression quantification. A comprehensive in silico assessment of RNA-seq experiments. *Mol. Ecol.*, **22**, 620–634.

Wang,Z. *et al*. (2009) RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.*, **10**, 57–63.

Yang,Y. and Smith,S.A. (2014) Orthology inference in nonmodel organisms using transcriptomes and low-coverage genomes: improving accuracy and matrix occupancy for phylogenomics. *Mol. Biol. Evol.*, **31**, 3081–3092.

Yates,A. *et al*. (2016) Ensembl 2016. *Nucleic Acids Res.*, **44**, D710–D716.

2.3 Application de CAARS : intérêts et limites de CAARS

A ma connaissance, CAARS a été utilisé ou est en cours d'utilisation pour la création de trois jeux de données :

— Le premier, dans le cadre de l'étude de la convergence morphologique dans la première molaire des rongeurs dans mon équipe au LBMC. Je me suis occupée de la création du jeu de données.

— Le second, dans le cadre de l'étude de la convergence chez les mammifères myrmécophages dans l'équipe de Frédéric Delsuc (ISEM, Montpellier) dont la création du jeu de données est réalisée par un doctorant, Rémi Allio.

— Le dernier, dans le cadre de l'étude des chromosomes sexuels chez une espèce de plante (Silene acaulis) dans l'équipe de Gabriel Marais (LBBE, Lyon) dont la création du jeu de données est réalisée également par un doctorant, Djivan Prentout.

Ces premières applications ont montré que CAARS fonctionnait et permettait la création de jeux de données exploitables. La qualité des jeux de données en est même parfois nettement améliorée. Cette nette amélioration a été observée dans l'équipe, pour une étude comparative de profils d'expression développementaux de dents de rongeurs. J'ai ainsi observé que la différence d'expression entre espèces est moins importante dans un jeu de données construit avec CAARS, que dans un jeu de données construit par un pipeline traditionnel (obtenu par assemblage *de novo*, puis association des contigs par BLAST dans des familles de gènes préexistantes). Ceci est nettement visible sur les matrices de corrélation des niveaux d'expression entre espèces présentées dans la Figure 2.4. On voit qu'avec CAARS, les niveaux d'expression sont beaucoup plus corrélés entre espèces (R^2 de 0.88 à 0.93 selon la paire d'espèces comparée) qu'avec le pipeline traditionnel (R^2 de 0.39 à 0.81 selon la paire d'espèces comparée). Ces résultats sont beaucoup plus conformes à l'attendu, car la phylogénie des quatre espèces considérées est en étoile, avec des espèces relativement proches (40 MA divergence). Le nombre de gènes reconstruits n'est pas très différent entre les méthodes (12 004 pour CAARS et 12 211 pour le pipeline traditionnel).

FIGURE 2.4 – Matrices représentant les corrélations des niveaux d'expression entre quatre espèces de rongeurs. Les données sont des moyennes obtenues sur plusieurs stades de développement de bourgeons dentaires. La couleur des carrés figure la valeur des corrélations par paires d'espèces. A gauche, quantifications obtenues avec des assemblages de CAARS, et à droite, par le pipeline traditionnel.

Ces premières applications pratiques ont également montré des points sur lesquels CAARS pouvait s'améliorer notamment : la réconciliation lors de la dernière étape et la vitesse d'exécution. En effet, bien que la réconciliation soit performante, il reste des familles de gènes où la réconciliation pourrait être plus efficace. Dans certains cas, l'annotation des relations d'homologie est erronée

dans le sens où des orthologues sont annotés comme paralogues au lieu d'orthologues. Ce problème diminue le nombre de sous familles de gènes orthologues 1:1. Le second problème est la vitesse d'exécution de CAARS qui augmente rapidement avec le nombre de gènes et le nombre d'espèces inclus dans le jeu de données et notamment lors de l'utilisation d'apytram dans les options.

Enfin, ces cas d'utilisation ont permis et permettront de répondre à des questions que je me posais lors sa mise à disposition.

D'un point de vue pratique :

— Est-ce que CAARS est aussi facile d'utilisation que j'ai essayé de le faire ?

CAARS a été utilisé par des bioinformaticiens non débutants sans trop de problèmes. Les problèmes qu'ils ont rencontrés ont plutôt été sur l'obtention des données d'entrées (familles de gènes déjà existantes). Il faudrait peut être ajouter au tutoriel une partie dédiée à ce travail.

— Quelles peuvent être les cas d'utilisation et notamment les besoins en sorties de CAARS ?

Actuellement, les sorties de CAARS sont nombreuses. Il y a, par exemple, les alignements finaux, les arbres de gènes avant et après réconciliations, des tableaux résumant les relations d'homologie entre gènes et les sous-groupes d'orthologues 1:1, un fichier séparé par nouvelle espèce regroupant l'ensemble des séquences construites dans CAARS. L'ensemble de ces sorties avait pour but de recouvrir une majorité des besoins possibles pour une analyse de génomique comparative. Cependant, seuls les alignements et les arbres après réconciliation ont été utilisés. Il faudrait peut-être que je réorganise les sorties et que j'améliore la documentation.

— Quels sont les temps d'exécution sur des jeux de données variés ?

La création d'un jeu de données avec CAARS, au vu des retours que j'ai eus, est de quelques jours sur une machine assez conséquente (16 CPU) pour un jeu de données de 20 espèces dont 2 nouvelles. Je n'ai actuellement pas de moyen de connaître le temps d'exécution a priori. Cela est un frein pour l'utilisation de CAARS. En effet, un utilisateur ne peut pas se permettre de lancer un processus sans savoir si le temps d'exécution sera de 5 jours ou 30 jours. Il faudrait que je lance des tests pour connaître comment certains paramètres (nombre d'espèces, nombre de lectures dans les échantillons de RNA-Seq, nombre de gènes, etc.) font augmenter le temps de calculs (augmentation linéaire, quadratique, etc.). Cependant je ne peux pas lancer la création de gros jeux de données juste pour connaître leur temps de création. Il faut que je fasse des tests à petite échelle puis que j'extrapole pour des jeux de données complets et enfin que je demande aux utilisateurs le temps d'exécution de leur processus pour vérifier que mes extrapolations sont bonnes.

Et d'un point de vue biologique :

— Est-ce que les paralogues sont bien annotés ?

Le jeu de données que j'ai réalisé pour mon équipe a été analysé par un étudiant, Timothée Kastylevsky, qui a utilisé les annotations d'homologie et récupéré les orthologues 1:1. Il a vérifié manuellement un certain nombre de familles et montré que dans l'ensemble les données étaient de bonne qualité. Cependant, un petit nombre de séquences annotées comme des paralogues (une dizaine) sont en fait des orthologues après vérification manuelle. La réconciliation pourrait donc sans doute être encore améliorée.

— Est-ce que CAARS peut être utilisé sur des jeux de données très divers ? Notamment chez les plantes connues pour leurs nombreuses duplications ?

Je n'ai pas encore de recul sur cette question. L'application sur les plantes est en cours et j'ai hâte d'avoir les résultats.

2.4 Conclusion et Perspectives

Les retours positifs au sujet de CAARS montrent que c'est un outil prometteur et attendu par la communauté. Cependant, les cas d'utilisation de CAARS ont montré certaines de ces limites : son temps d'exécution, qui augmente de manière considérable si apytram est utilisé, et la réconciliation des arbres, qui peut être encore améliorée.

Pour remédier au problème du temps d'exécution, je conseille de n'utiliser apytram que dans des cas spécifiques, par exemple sur certaines familles de gènes mais pas pour l'ensemble du génome. Dans l'article, nous avons montré qu'apytram améliore considérablement les assemblages des gènes les moins exprimés. Une alternative serait d'automatiser l'utilisation d'apytram uniquement pour certains gènes. On peut imaginer de sélectionner ces gènes en fonction de leur niveau d'expression ou de la taille de leur séquence par rapport à la taille des séquences des autres gènes de sa famille. Ce n'était pas possible au moment où CAARS a été publié, mais maintenant le gestionnaire de pipelines que CAARS utilise rend ce nouveau développement possible.

Une autre piste pour accélérer la création de jeux de données avec CAARS est de changer le système de containerisation de CAARS. Lors de la création de CAARS, j'ai utilisé Docker qui était la solution la plus adéquate à ce moment là, mais depuis, la technologie Singularity est devenue disponible. Singularity ne nécessite pas de droits administrateurs. Cela signifie que l'on pourrait utiliser plus facilement CAARS sur des fermes de calculs, ce qui permettrait de paralléliser de manière intensive les différentes tâches de CAARS et donc réduire le temps global d'exécution. Cependant, cela ne doit pas être la seule solution à retenir mais plutôt une solution complémentaire à l'optimisation de l'utilisation d'apytram.

La seconde partie de CAARS qui pourrait être améliorée est l'étape de réconciliation. Nous avons déjà identifié deux logiciels qui pourraient améliorer les performances de CAARS : Treerecs (https://gitlab.inria.fr/Phylophile/Treerecs) et Generax (https://github.com/BenoitMorel/GeneRax). Ces deux logiciels de réconciliation sont plus récents que la solution retenue dans CAARS et semblent être plus efficaces. Il faudrait donc les tester par rapport à la solution actuelle et voir le gain engendré.

Enfin, bien que les utilisateurs ont souligné l'utilité du tutoriel, certaines de leurs questions montrent qu'il pourrait être encore plus détaillé. Certaines fonctionnalités de CAARS, très utiles, semblent pouvoir être mises davantage en avant. C'est le cas, par exemple, de la désactivation de l'utilisation d'apytram. D'autre part, les sorties de CAARS semblent être trop nombreuses, elles gagneraient à être simplifiées.

En conclusion, le développement de CAARS a apporté un nouvel outil à la communauté qui est utilisé mais qui pourrait être encore amélioré pour faciliter son utilisation.

Ces premiers retours d'utilisateurs sont une étape très importante pour CAARS. En effet, la vie d'un outil de bioinformatique ne s'arrête pas lors de sa publication mais plutôt lors de l'arrêt de son développement. La publication de l'outil peut être vue comme sa naissance et ce sont ses utilisateurs qui le feront grandir. L'outil restera en vie tant qu'il sera en perpétuel développement en prenant en compte les retours de ses utilisateurs et les avancées des autres outils disponibles. L'aboutissement du développement d'un outil bioinformatique n'est pas sa publication mais plutôt obtenir un nombre croissant d'utilisateurs.

Personnellement, le développement de CAARS m'a appris la rigueur, l'importance de la reproductibilité et a permis d'améliorer mes compétences en bioinformatique, notamment avec la collaboration de Philippe Veber. J'espère pouvoir consacrer à nouveau du temps à cet outil qui, je suis sûre, sera utile à la communauté mais qui pour le moment n'est pas complètement satisfaisant et qui a besoin de grandir.

3 Comment détecter de la convergence génomique ?

Sommaire

3.1 Introduction

La deuxième partie de ma thèse a été consacrée à l'étude de la détection de la convergence génomique. Cette partie se décompose en deux projets, le premier a été le développement d'un outil, PCOC, permettant la détection de la convergence génomique au niveau des séquences codantes et le second a été l'étude de l'impact de facteurs confondants sur la capacité à détecter de la convergence génomique.

3.1.1 Contexte historique sur la détection de la convergence à l'échelle génomique

Lorsque j'ai débuté ce projet, deux analyses majeures existaient sur l'étude de l'évolution convergente au niveau génomique. La première était l'étude de l'acquisition indépendante de l'écholocation chez les mammifères (PARKER et collab., 2013) et la seconde, celle de la transition à la vie en milieu marin chez les mammifères (FOOTE et collab., 2015). Ces deux études ont cherché à détecter de la convergence liée au phénotype convergent étudié au niveau des sites des séquences codantes. En effet, de nombreux sites convergents ont été détectés dans le cas de gènes candidats (ex : la prestine) mais la quantité de sites convergents à l'échelle du génome était inconnue. De plus, l'étude des séquences codantes peut permettre d'identifier des sites qui pourraient être fonctionnellement liés à l'acquisition du phénotype convergent étudié.

Les conclusions de ces deux études ont été bien différentes. La première trouvait que la convergence était globale au niveau du génome alors que la seconde trouvait très peu de convergence génomique liée au phénotype convergent. Il est possible que les bases génétiques de ces deux adaptations soient de nature très différente (par exemple touchant préférentiellement l'évolution protéique dans un cas ou l'évolution cis-régulatrice dans l'autre cas ; ou passant par un grand nombre versus un petit nombre de gènes), expliquant cette divergence des résultats. Cependant, les résultats de ces deux études ont été remis en question par des études postérieures (THOMAS et HAHN, 2015; THOMAS et collab., 2017; ZOU et ZHANG, 2015) qui ont soulevé des problèmes méthodologiques.

Dans le cas de l'étude sur l'écholocation, il a été montré que la grande quantité de convergence génomique n'était pas spécifique aux paires d'espèces partageant le phénotype convergent. En effet, des paires d'espèces ne partageant pas de phénotype convergent évident montraient également des quantités de convergences équivalentes ou supérieures (THOMAS et HAHN, 2015; ZOU et ZHANG, 2015). Dans le cas de l'étude sur les mammifères marins, il a été montré que les résultats étaient très sensibles aux nombres d'espèces non-convergentes présentes dans le jeu de données (THOMAS et collab., 2017).

Les résultats des études initiales peuvent donc être affectés d'une part par les méthodes utilisées, et d'autre part, par des facteurs confondants qui pourraient induire de la convergence génétique et auxquels les auteurs ne se sont pas intéressés.

3.1.1.1 Limites techniques des méthodes initiales et proposition d'une méthode alternative

Dans (PARKER et collab., 2013), la méthode utilisée, que l'on nommera "topologique", teste la concordance entre chacun des sites et deux topologies, la première étant celle de l'arbre des espèces et la seconde celle d'un arbre des espèces remanié tel que les espèces convergentes sont regroupées (Figure 3.1). La méthode établit un score de convergence pour chaque site d'un alignement de séquences. Le score de convergence de ce site est d'autant plus élevé que la topologie préférée est la convergente. Dans (FOOTE et collab., 2015), la méthode utilisée, que l'on nommera "identique", est basée sur l'identification de substitutions identiques ayant eu lieu de manière indépendante lors de toutes les transitions convergentes, c'est à dire lors des acquisitions du phénotype convergent (Figure 3.2, substitutions convergentes).

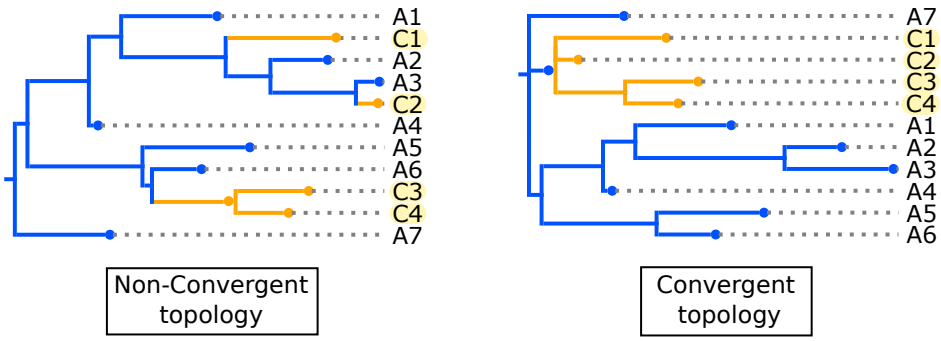

FIGURE 3.1 – Principe de la méthode topologique. Pour chacun des sites, on teste la concordance avec deux topologies, la première étant celle de l'arbre des espèces (gauche) et la seconde celle d'un arbre des espèces remanié tel que les espèces convergentes sont regroupées (droites). Les couleurs des branches indiquent les espèces convergentes (orange) et non-convergentes (bleu).

Aucune de ces deux méthodes ne nous semblait adéquate pour détecter de la convergence à l'échelle génomique, en effet chacune souffre de limites conceptuelles.

D'une part, la méthode topologique n'est pas une méthode mécanistique, c'est à dire qu'elle exploite une caractéristique des modèles d'inférence d'arbre phylogénétique. En effet, plus des séquences sont similaires, comme c'est le cas lors d'une convergence, plus les modèles d'inférence d'arbres ont tendance à les regrouper. Cependant, ce signal peut aussi être dû à d'autres processus biologiques comme par exemple des taux d'évolution élevés, qui sont partagés entre des lignées non apparentées.

D'autre part la méthode identique est très stricte car elle nécessite des substitutions vers le même acide aminé dans l'ensemble des transitions. Or, des acides aminés différents peuvent avoir les mêmes propriétés physico-chimiques et donc la même fonction (Figure 3.2). De plus, la probabilité d'observer le même acide aminé dans les espèces convergentes et un autre acide aminé dans les espèces non-convergentes diminue fortement quand le nombre d'espèces augmente.

C'est pour cela que nous avons proposé de relâcher cette définition en autorisant une substitution vers plusieurs acides aminés (Figure 3.2). Nous avons donc proposé de définir la convergence génomique par la combinaison de deux composantes (Figure 3.2). La première correspond à la présence de substitutions aux niveaux des transitions convergentes qui seraient associées à l'acquisition du phénotype convergent. Nous avons appelé cette composante OC pour "One Change". La seconde correspond au changement de la nature des acides aminés d'un ensemble d'acides aminés vers un autre ensemble d'acides aminés résultant de ces substitutions. Ce changement de nature des acides aminés serait associé au changement de la fonction de ce site chez les espèces convergentes. Nous avons appelé ces ensembles d'acides aminés, profils d'acides aminés et cette composante PC pour "Profile Change".

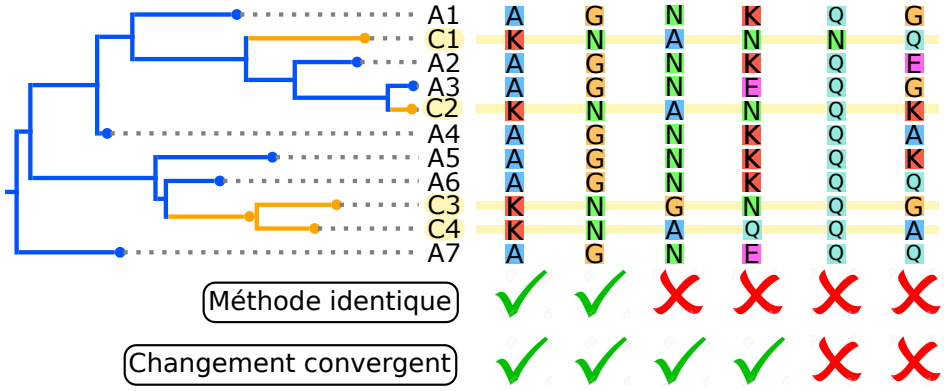

FIGURE 3.2 – Comparaison entre les sites inclus dans les définitions historiques de substitutions convergentes et une nouvelle définition moins stricte où l'on autorise des substitutions vers plusieurs acides aminés. Les crochets verts et les croix rouges indiquent pour chacun des sites s'il est défini convergent par la méthode.

3.1.1.2 Les facteurs confondants potentiels ne sont pas pris en compte

Les sites détectés par les méthodes sont corrélés à des changements phénotypiques convergents, mais nous ne pouvons pas savoir s'il y a un lien de causalité entre la substitution et le changement convergent. En effet, d'autres processus biologiques pourraient provoquer de la convergence génomique sans être liés à une convergence phénotypique (cf section 1.3).

Par exemple, une augmentation de la taille efficace dans les espèces convergentes peut augmenter l'efficacité de la sélection et faciliter la fixation de certaines mutations indépendamment de l'adaptation convergente. Cela signifie que des substitutions se produisant dans les espèces convergentes peuvent être corrélées au phénotype convergent mais sans être fonctionnellement liées au phénotype convergent.

Dans cette partie de ma thèse, j'ai donc travaillé sur ces deux aspects en développant une nouvelle méthode, appelée PCOC, basée sur une définition plus relâchée de la convergence et en caractérisant les méthodes vis-à-vis de facteurs confondants potentiels.

3.2 Résultats

3.2.1 PCOC, un nouvel outil de détection de la convergence basé sur des changements de profils d'acides aminés

3.2.1.1 Avant-propos

J'ai implémenté la définition de la convergence génomique que j'ai présentée précédemment dans un nouvel outil de détection de la convergence qui est donc basé sur des changements convergents de profils d'acides aminés associés à des substitutions au niveau des transitions convergentes. Nous avons appelé cet outil PCOC, pour "Profile Change" and "One Change" (Figure 3.3).

Avant d'utiliser PCOC sur des données réelles, nous avons voulu tester ses performances sur des simulations et ainsi comparer ses performances avec les autres méthodes. Ces travaux sont présentés dans l'article ci-après et sont précédés d'une présentation de PCOC.

3.2.1.1.1 Présentation du modèle implémenté dans PCOC

Le but de cette présentation est de cerner le principe du modèle implémenté dans PCOC sans rentrer dans les détails techniques de son implémentation. PCOC traite chacun des sites de manière

indépendante et définit pour chacun d'entre eux quel modèle d'évolution lui convient le mieux, soit un modèle d'évolution convergente soit un modèle d'évolution non-convergente. Le modèle d'évolution non-convergente considère que l'évolution du site est indépendante du phénotype de l'espèce, alors que le modèle d'évolution convergente considère que le phénotype interagit avec l'évolution du site.

Un modèle d'évolution décrit l'évolution de séquences le long de chacune des branches d'un arbre phylogénétique et dépend de deux éléments pour chacune des branches :

— une matrice qui représente le taux d'obtention d'un nouvel acide aminé à partir d'un acide aminé de départ par unité de temps et qui est appelé la matrice des taux instantanés de substitution.

— la longueur de la branche qui représente le temps écoulé entre les deux extrémités, appelées noeuds, de la branche.

Pour chacune des branches, la probabilité d'obtention d'une substitution est donc le produit de la longueur de cette branche et de la matrice des taux instantanés associée à la branche. La matrice des taux instantanés de substitution est définie à partir des taux d'échangeabilité entre acides aminés et des fréquences d'équilibre des acides aminés, c'est à dire la composition en acides aminés attendue si les longueurs de branches étaient infinies. Cela signifie que plus une branche est longue, plus la probabilité de substitutions est grande mais aussi que plus la composition en acides aminés à la base de la branche est en inadéquation avec le vecteur de fréquences d'équilibre des acides aminés, plus la probabilité de substitutions est grande sur cette branche.

Dans PCOC, nous avons utilisé la matrice de substitutions pour modéliser l'évolution convergente en modifiant cette matrice lors du changement de phénotype. Notre modèle d'évolution convergente est concrètement un modèle hétérogène où l'on combine deux modèles d'évolution. Les branches soutenant des espèces convergentes se partagent un même vecteur de fréquences d'acides aminés, que l'on appellera **profil convergent d'acides aminés** (en orange sur la figure), et les branches soutenant des espèces non convergentes se partagent un profil d'acides aminés différents, que l'on appellera **profil ancestral d'acides aminés** (en bleu sur la figure). De plus, d'après la définition que nous avons choisie, nous imposons la présence d'une substitution au niveau des transitions convergentes. C'est pour cela que nous avons également ajouté à notre modèle global d'évolution convergente, la présence obligatoire d'au moins une substitution au niveau du changement de modèle non-convergent vers le modèle convergent, c'est à dire au niveau de la transition convergente (carré noir sur la figure). Le modèle d'évolution convergente est contrasté dans PCOC avec un modèle non-convergent. Ce modèle non-convergent est quant à lui défini avec un seul vecteur de fréquences d'acides aminés le long de l'arbre, qui correspond au profil d'acides aminés ancestral.

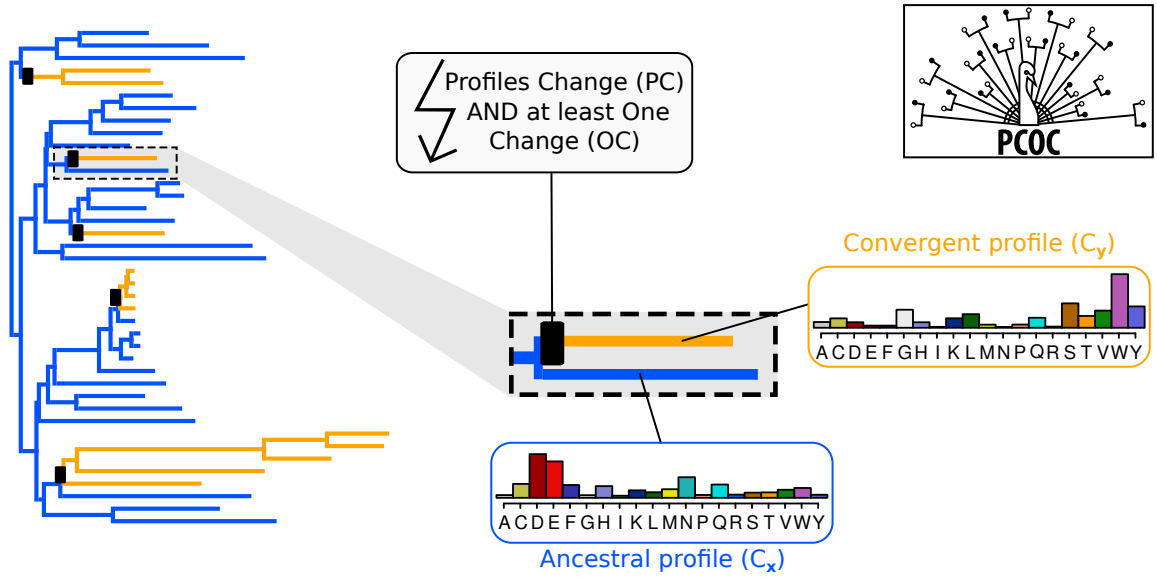

FIGURE 3.3 – Principe général de PCOC. Les couleurs des branches représentent les profils d'acides aminés associés à chacune des branches, en orange le convergent et en bleu le non-convergent. Les encadrés orange et bleu présentent des exemples de profils d'acides aminés où une fréquence est associée à chacun des acides aminés. Le zoom au centre de la figure montre le principe de PCOC, avec un changement de profils (Profile Change, PC) d'acides aminés au niveau des transitions convergentes (carré noir) associé avec un changement d'acides aminés (One Change, OC).

Pour mettre en place notre modèle, nous avions besoin de profils d'acides aminés représentatifs de ce que l'on retrouve dans des données réelles. Nous nous sommes tournés vers des profils déterminés par (LE et collab., 2008) dans le cadre de l'implémentation du modèle CAT (LARTILLOT et PHILIPPE, 2004) avec des profils prédéfinis en maximum de vraisemblance. Ces profils d'acides aminés ont été définis à partir de l'exploitation d'une base de données contenant des séquences de protéines dont la structure était connue (SANDER et SCHNEIDER, 1991). A partir des alignements de ces séquences, les auteurs ont défini des classes de sites dont les compositions en acides aminés étaient similaires (Figure 3.4). Ils ont ensuite définis six ensembles de profils les C10, C20, C30, C40, C50 et C60 comprenant respectivement 10, 20, 30, 40, 50 et 60 profils d'acides aminés.

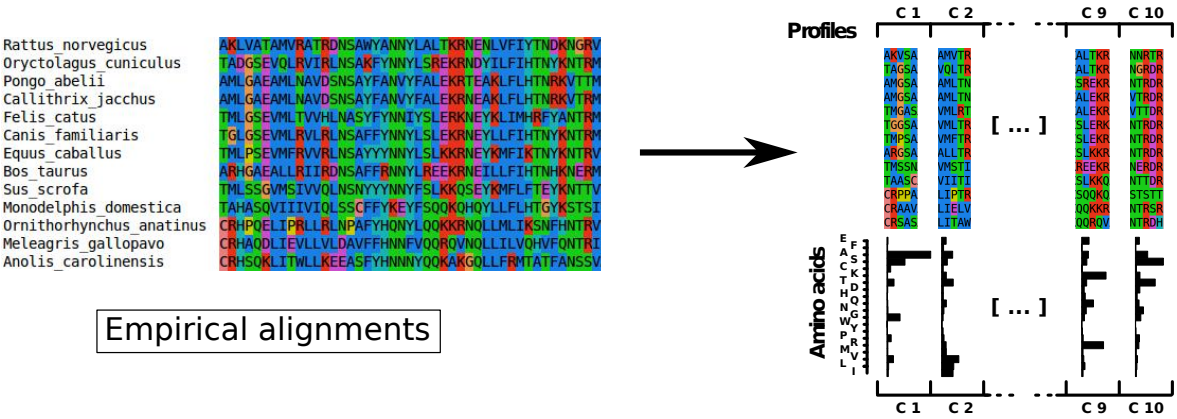

FIGURE 3.4 – Principe simplifié de la définition des profils prédéfinis par (LE et collab., 2008).

3.2.1.1.2 Présentation du logiciel PCOC

Concrètement, PCOC est une boîte à outils composée d'un outil de simulation (PCOC_SIM) et d'un outil de détection (PCOC_DET). Ces outils sont codés en python et font des appels à bppseqgen

et bppml, des utilitaires provenant de la bibliothèque Bio++ (GUÉGUEN et collab., 2013) dans laquelle a été implémenté le modèle PCOC par Laurent Guéguen. Les détails de l'implémentation sont disponibles dans le matériel supplémentaire de la publication. Nous avons également inclus dans la partie détection de PCOC des implémentations des méthodes topologique et identique pour lesquelles aucune implémentation n'était disponible.

En plus d'avoir été utilisé pour tester les performances de PCOC et les comparer aux méthodes existantes, l'outil de simulation a un second intérêt. En effet, il permet de tester les performances théoriques de ces méthodes sur un jeu de données réelles. Par exemple, on ne peut pas savoir à l'avance si un jeu de données est propice à la détection de la convergence, notamment en fonction des longueurs de branches, du nombre d'espèces ou du nombre de transitions convergentes. Avec cet outil, les utilisateurs peuvent anticiper les performances attendues de ces méthodes sur un jeu de données et en modifier la construction *a priori* en ajoutant, par exemple, des espèces afin d'améliorer les performances des méthodes de détection.

3.2.1.2 Article : Accurate Detection of Convergent Amino-Acid Evolution with PCOC

PCOC a été publié dans Molecular Biology and Evolution le 07 Juillet 2018 (https://doi.org/10.1093/molbev/msy114). Tout comme CAARS, il s'agit d'un logiciel libre hébergé et disponible sur github (https://github.com/CarineRey/pcoc). Un tutoriel ainsi que des précisions sur son implémentation sont disponibles dans un wiki (https://github.com/CarineRey/pcoc/wiki).

Accurate Detection of Convergent Amino-Acid Evolution with PCOC

Carine Rey,[1,2] Laurent Guéguen,[2] Marie Sémon,[1] and Bastien Boussau*[,2]

[1]UnivLyon, ENS de Lyon, Univ Claude Bernard, CNRS UMR 5239, INSERM U1210, Laboratoire de Biologie et Modélisation de la Cellule, Lyon, France

[2]Laboratoire de Biométrie et Biologie Évolutive (LBBE), Université de Lyon, Université Lyon 1, CNRS, Villeurbanne, France

***Corresponding author:** E-mail: bastien.boussau@univ-lyon1.fr.

Associate editor: Tal Pupko

Abstract

In the history of life, some phenotypes have been acquired several times independently, through convergent evolution. Recently, lots of genome-scale studies have been devoted to identify nucleotides or amino acids that changed in a convergent manner when the convergent phenotypes evolved. These efforts have had mixed results, probably because of differences in the detection methods, and because of conceptual differences about the definition of a convergent substitution. Some methods contend that substitutions are convergent only if they occur on all branches where the phenotype changed toward the exact same state at a given nucleotide or amino acid position. Others are much looser in their requirements and define a convergent substitution as one that leads the site at which they occur to prefer a phylogeny in which species with the convergent phenotype group together. Here, we suggest to look for convergent shifts in amino acid preferences instead of convergent substitutions to the exact same amino acid. We define as convergent shifts substitutions that occur on all branches where the phenotype changed and such that they correspond to a change in the type of amino acid preferred at this position. We implement the corresponding model into a method named PCOC. We show on simulations that PCOC better recovers convergent shifts than existing methods in terms of sensitivity and specificity. We test it on a plant protein alignment where convergent evolution has been studied in detail and find that our method recovers several previously identified convergent substitutions and proposes credible new candidates.

Key words: **convergent evolution, genomics, bioinformatics, echolocation, C4 metabolism, sequence evolution.**

Article

Introduction

Convergent phenotypic evolution provides unique opportunities for studying how genomes encode phenotypes, and for quantifying the repeatability of evolution. These questions are typically addressed by sequencing genes or genomes belonging to a sample of species sharing a convergent phenotype, along with those of closely related species sharing a different ancestral phenotype. Then, nucleotide or amino acid positions that are inferred to have changed specifically on those branches where the phenotypes convergently changed may be assumed to be involved in the convergent evolution of those phenotypes. Such an approach has been used on spectacular cases of convergent evolution such as the C4 metabolism in grasses (Besnard et al., 2009), the ability to consume a toxic plant compound in insects (Zhen et al., 2012), echolocation in whales and bats (Parker et al., 2013), or the ability to live in an aquatic environment in mammals (Foote et al., 2015). These studies have found different levels of convergent evolution. In particular Parker et al. (2013) investigated convergent substitutions associated with the evolution of echolocation in mammals, which has evolved once in whales and once or twice in bats. They focused on amino acid sequences rather than on nucleotide sequences, assuming that it is where most selective effects would be observed. Using a topology-based method, they found a large number of convergent substitutions in close to 200 genes. However when these protein data were reanalyzed using another method, it was concluded that many of those convergent changes were likely false positives (Thomas and Hahn, 2015; Zou and Zhang, 2015b).

These strong disagreements come from differences in the bioinformatic methods that were used to detect convergent substitutions, and the underlying definition of what makes a substitution convergent. If we put aside studies of individual genes that involved manual analyses of alignments and detailed investigations of the rate of sequence evolution and patterns of selection along gene sequences (Besnard et al., 2009; Zhen et al., 2012), genomic studies have relied on two different methods. In Zhang and Kumar (1997), and later in Foote et al. (2015), Zou and Zhang (2015b), and Thomas and Hahn (2015), convergent sites are defined as those that converged to the exact same amino acid in all convergent species. Instead, in Parker et al. (2013), a more operational definition is used: a convergent site is one that prefers to the species

phylogeny a phylogeny in which species with the convergent phenotype group together. In doing so, they have no explicit requirement over the type of amino acid change that occurred in the species with the convergent phenotype because their method is remote from the actual mechanism of substitutions. With a more relaxed definition than in Zou and Zhang (2015b) and Thomas and Hahn (2015), it is not surprising that they recover more instances of convergent amino-acid evolution.

From Convergent Substitutions to Convergent Shifts

We believe that these two definitions have several shortcomings. First, the historical definition of Zhang and Kumar (1997) seems very strict. Selecting only sites that converged to the exact same amino acid in all species with a convergent phenotype is bound to capture only a subset of the substitutions associated with the convergent phenotypic change. This will capture only those sites where a unique amino acid is much more fit in the convergent phenotype than all other amino acids. In many other cases, there may be more than one amino acid that is fit at a particular position, given the convergent phenotype. For instance, it may be that several amino acids with similar biochemical properties have roughly the same fitness at that site. In such circumstances, we do not expect that identical amino acids will be found in all species with the convergent phenotype, but that several amino acids with similar biochemical properties will be found in all species with the convergent phenotype. Such convergent shifts in the amino acid preference at a given site are not considered under the definition of Zhang and Kumar (1997) and Foote et al. (2015). Second, Parker et al.'s (2013) definition may be too loose, as it is entirely disconnected from the substitution process.

We propose to consider shifts in amino acid preference instead of convergent substitutions. To us, a substitution is convergent if it occurred toward the same amino acid preference on every branch where the phenotype also changed toward the convergent phenotype. We model the amino acid preference at a position and on a branch by a vector of amino acid frequencies, which we call a profile. The amino acid profile used in species with the convergent phenotype needs to be different from the profile used in species with the ancestral phenotype. This definition conveys the idea that a convergent substitution is necessary to a convergent phenotype, that is, every time the phenotype changes to the convergent state, the position must change toward the convergent phenotype. It is thus equivalent to Zhang and Kumar's (1997) definition in its positioning of changes on the branches where the phenotypic change occurred, but it seems less restrictive from a biochemical point of view. It extends previous works (Tamuri et al., 2009; Studer et al., 2014; Parto and Lartillot, 2017, 2018) that also modeled changes in amino acid profiles, but did not require that there should be a change on the branch where the phenotype changed from ancestral to convergent.

Detecting Convergent Shifts

In this manuscript, we evaluate our proposed definition by comparing a method that uses our definition to two other methods proposed in the literature to detect convergent substitutions.

The power of a method is usually analyzed in terms of specificity and sensitivity. Specificity is critical for methods that detect convergent substitutions. Specifity is inversely correlated to the false positive rate. A low false positive rate is necessary because we expect that most differences found in a group of genomes will not be directly related to the convergent phenotypic change, but may come from neutral processes or be selected for reasons unrelated to the convergent phenotype (Bazykin et al., 2007; Rokas and Carroll, 2008; Zou and Zhang, 2015a). Therefore, among a large number of changes, only a small number will be associated with convergent phenotypic evolution. There will be very few positives to find, and a large number of negatives, which provides many opportunities for methods to predict false positives. To illustrate this point, we can use the numbers of substitutions inferred on terminal branches of the species tree provided in Thomas and Hahn (2015), based on transcriptome-wide analyses. If we take the example of microbats and dolphins, species that both evolved the ability to echolocate, Thomas and Hahn (2015) report roughly 4,000 substitutions to different amino acids, which they call divergent, and 2,000 substitutions to the exact same amino acid, which they call convergent, that is, 6,000 substitutions total. These numbers are in proportion with those reported in pairs of non-echolocating species, which was taken as evidence that the majority of the 2,000 convergent substitutions detected by Parker et al. (2013) are not linked to the convergent evolution of echolocation. Instead they find that $<7\%$ of genes with convergent substitutions are also associated with positive selection, a number they choose as the true number of convergent substitutions. Based on these considerations, among the 6,000 substitutions, 140 are truly convergent, and 5,860 are not. If we were to apply a test that has a very respectable sensitivity of 98% and an equally good specificity of 98%, we would detect $0.98 \times 140 = 137$ true positives, and $0.02 \times 5{,}860 = 117$ false positives. So, we would have a false discovery rate of $117/(117 + 137) = 46\%$, despite a test with excellent properties. We use these simple calculations later in the manuscript when presenting the results obtained with different methods.

The three methods to detect convergent evolution are as follow. The first method used in Parker et al. (2013) is based on the comparison of two topologies, one for convergent sites, and the other for nonconvergent sites. It is derived from earlier efforts by Castoe et al. (2009). Here, we named this method "Topological." The second method used in Zou and Zhang (2015b), Thomas and Hahn (2015), and Foote et al. (2015) proposes to detect convergent changes related to a phenotypic change by focusing on substitutions to the exact same amino acid in each species with the convergent phenotype. We named this method "Identical." Both methods can be used on rooted or unrooted trees, since they do not explicitly consider changes in the substitution models. Finally, the third method fleshes out our own definition of convergent shifts and is based on a modification of usual models of site evolution (fig. 1). Under those models, any

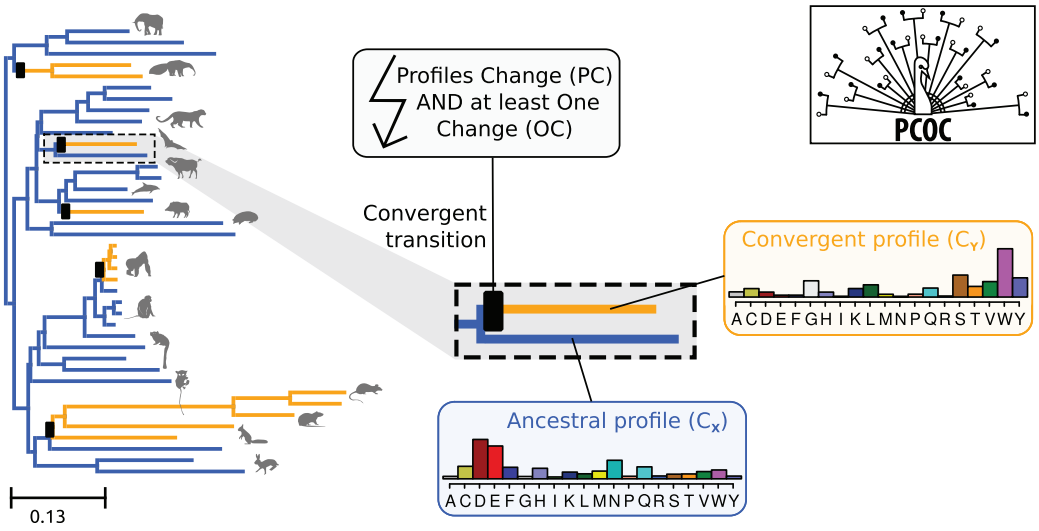

Fig. 1. PCOC attempts to detect sites that are linked to the repeated evolution of a convergent phenotype. On the left, the Ensembl Mammalian phylogeny has been represented, and five transitions have been randomly placed on its branches (black boxes). On the branches with the boxes, PCOC imposes an amino acid profile change and the use of the OC model. The convergent profile is used in subsequent branches.

number of substitutions (including zero) can occur on a branch. To impose that convergent substitutions should occur on the branches where the phenotype changes, we introduce the OneChange model, shortened into OC, which imposes at least one substitution per site on the branch where it is applied. In addition to OC, we consider that convergent sites evolve according to different amino acid equilibrium frequencies (i.e., different profiles) in species with the ancestral or convergent phenotypes. Here, amino acid profiles are defined as profiles from (Quang et al., 2008) (see supplementary fig. S1, Supplementary Material online), but other profiles could in principle be used. We named this model PCOC, for "Profile Change with One Change," and also because it is the name of a beautiful bird.

PCOC therefore combines two models, OC, which is new, and changes in amino acid profiles (PC), an idea that has been used before on single genes. In particular, it has been used to study changes in selective constraints in the Influenza virus (Tamuri et al., 2009), or convergent evolution of a particular enzyme in C3/C4 plants (Studer et al., 2014). Recently such profile changing models have been extended into a Bayesian framework by Parto and Lartillot (2017, 2018) for a gene-wise analysis of convergent evolution. In PCOC, it is possible to use only OC, or only PC, and in the manuscript, we explore the properties of these two submodels PC and OC. PCOC detects convergent sites by comparing the fit of two models.

Under the convergent model, a site evolves under a commonly used model of protein evolution on most branches. Then, in clades with the convergent phenotype, it evolves under a model with a different vector of amino acid equilibrium frequencies. Further, we apply OC on branches where the phenotype has changed from ancestral to convergent, imposing that the model shift occurs at the beginning of the branch (but the substitution event can occur anywhere on this branch). As the PCOC model is by definition nonstationary, it requires a rooted tree. Under the nonconvergent (null) model, a site evolves under a single amino acid profile throughout the phylogeny. We can thus compare the fit of the two models, the convergent and the nonconvergent ones, on a given site of an alignment in terms of their likelihood to classify this site as convergent or nonconvergent. We implemented these models to perform sequence simulation as well as probabilistic inference in the Maximum Likelihood framework. Mathematical details are provided in Materials and Methods as well as in Supplementary Material online.

In this manuscript, we implement the PCOC model for simulation and estimation. We compare its efficiency to that of two existing methods for detecting convergent evolution and investigate its behavior in a variety of conditions, changing the parameters of the simulation model, varying the number of convergent events, or introducing discrepancies between the simulation and inference conditions. Then we apply PCOC to a previously analyzed alignment of plant proteins where many convergent sites have been proposed. We find that although PCOC uses a different definition, it recovers many of the previously proposed convergent sites and conclude that this new model can be used on real data.

Results

Comparison of the Three Methods to Detect Convergent Changes

We compared the performance of the Topological, Identical, and PCOC approaches on simulations. We used empirical phylogenies, where a number of convergent transitions were placed at the beginning of random branches (from two to seven events). We also performed simulations with five convergent events, on the same empirical topologies, but varying branch lengths from small to large (fig. 2). To compare the methods fairly, we have chosen thresholds that maximize their individual performance (see Materials and Methods).

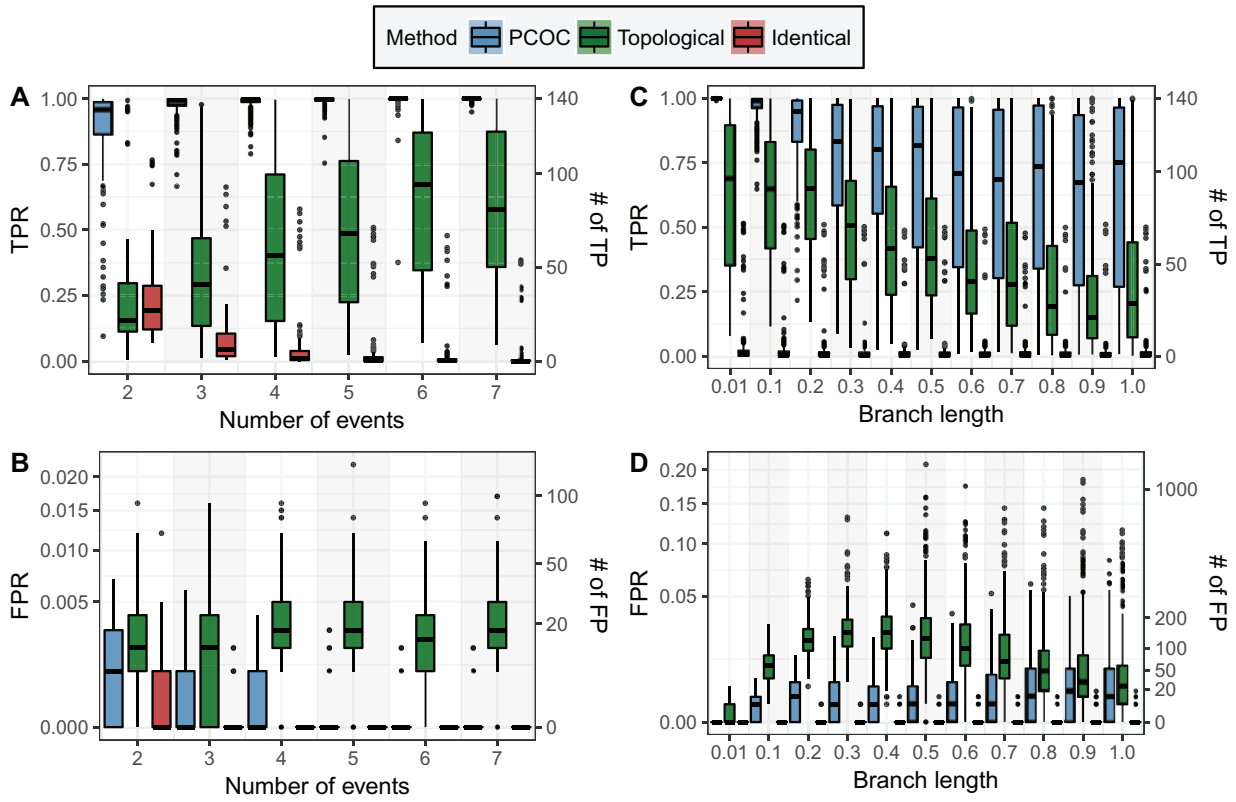

FIG. 2. Comparison between the topological, identical, and PCOC approaches to detect convergent substitutions. In (*A*) and (*B*), we vary the number of convergent events from two to seven. In (*C*) and (*D*), we set all branch lengths in the tree to a single value, ranging between 0.01 and 1.0 expected substitutions per site. The True Positive Rate (TPR) is the rate of TP among positives, that is, the *sensitivity*, and the False Positive Rate (FPR) is the rate of FP among the negatives, that is, 1−specificity. The right axes provide the numbers of true and false positives in the context of the example of the Introduction.

However, the simulations are performed under our definition of convergent substitutions, which could advantage our method, designed to fit this definition, compared with the Topological and Identical methods. It is unclear how we could have avoided this bias. We expect that the Topological approach, with its operational definition, should be able to capture shifts in amino acid profiles, and could obtain very good results. The Identical approach is expected to have a much worse sensitivity, and can only capture convergent changes only when the convergent profile is very centered on a single amino acid. We will see that the results recover these broad tendencies. We used the mammalian subtree of the Ensembl Compara phylogeny, but similar results were obtained on other phylogenies (a phylogeny of birds from Jarvis et al., 2014, a phylogeny of Rodents from Schenk et al., 2013, and a phylogeny of the PEPC gene in sedges; supplementary figs. S17, S25, and S33, Supplementary Material online). PCOC outperforms the other approaches in the vast majority of conditions, by recovering higher proportions of true positives and lower proportions of false positives. Expectedly, PCOC and the Topological approaches both improve as the number of convergent changes increases (fig. 2*A* and *B*). However, the performance of the Identical method degrades as the number of changes increases, because it is rare that the exact same amino acid is found in, for example, seven clades. As expected, the efficiency of all the methods increases as the distance between the simulated ancestral and convergent profiles increases (supplementary fig. S4, Supplementary Material online). We also investigated the impact of the convergent profile itself, using a measure of its entropy. A profile with high entropy has similar frequencies for all 20 amino acids, whereas a profile with low entropy only has a few amino acids containing most of the probability mass. We find that PCOC is nearly insensitive to the entropy of the convergent profile, because its OC component itself is insensitive. However, both the identical and topological approaches have better results on convergent profiles with low entropy (supplementary fig. S16, Supplementary Material online). This result is expected for the Identical method, which should be best in cases where the probability mass of the convergent profile is all contained in one single amino acid.

The performance of all methods tends to decrease as branch lengths become longer (fig. 2*C* and *D*). The Topological approach however predicts fewer false positives for branches nearing 1.0 expected substitution per site than for branches of length 0.5, but always performs worse than PCOC.

To ensure that PCOC was not unfairly favored in those tests, the above simulations have been performed using the C60 set of amino acid profile, while inference was performed using a different set of profiles, C10. We also tried to further

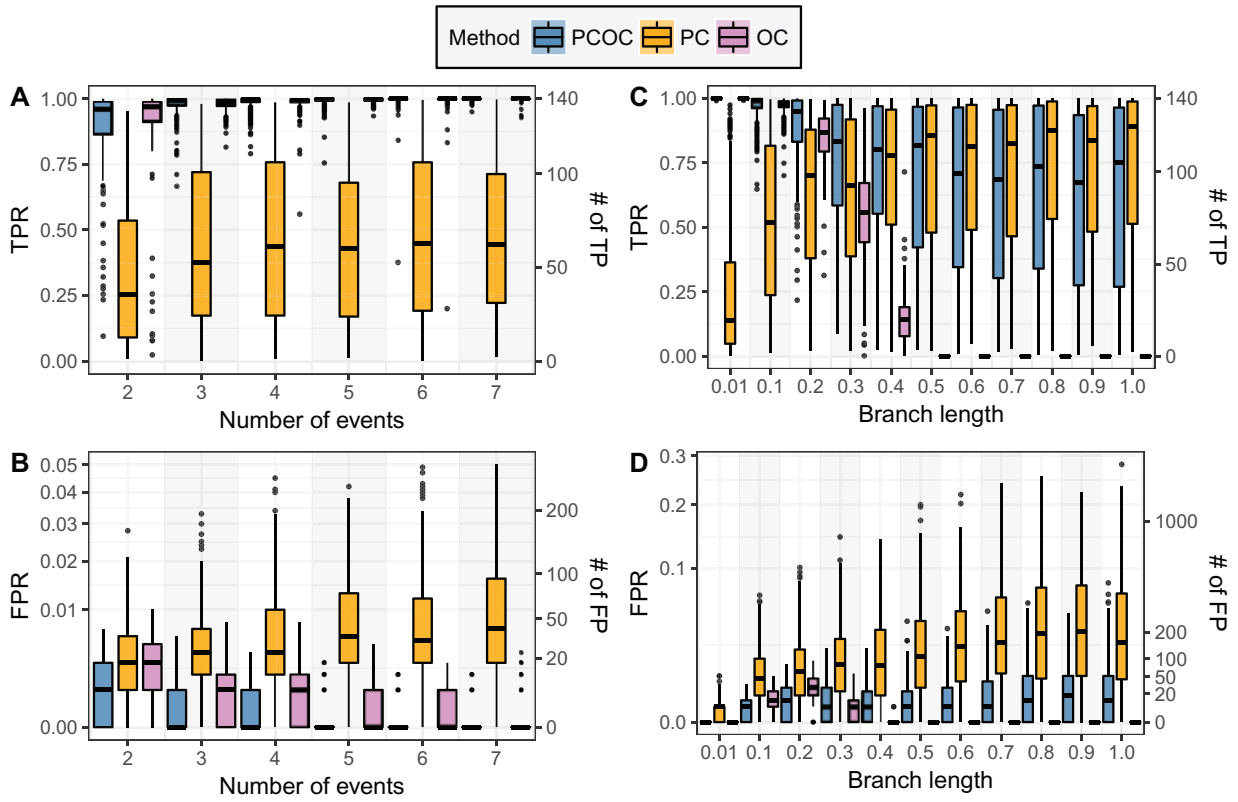

FIG. 3. The power of PCOC draws upon its submodels PC and OC. See figure 2 for legend.

complexify the simulations to make them harder for PCOC to analyze and evaluate how PCOC fares when some of its assumptions are violated. In particular, we used more than one amino-acid profile on the branches with the ancestral phenotype. To achieve this, we picked at random a few branches with the ancestral phenotype, and applied a different amino acid profile to those branches and the subsequent branches (supplementary fig. S8, Supplementary Material online). We observed that PCOC's performance did not change (supplementary figs. S9 and S10, Supplementary Material online). We also tested the performance of PCOC with misestimated branch lengths. To this end, we performed inferences on the trees used for simulation but after altering their branch lengths (see Materials and Methods). The results did not seem to be affected by the amount of error introduced (supplementary figs. S11 and S12, Supplementary Material online).

We also assessed how PCOC was affected by misplacements of the events of convergent evolution. Supplementary figure S13, Supplementary Material online, shows that PCOC is more sensitive to the inclusion of a spurious event of convergent evolution than to the removal of an event of convergent evolution. However, PCOC still obtains better results than the topological or the identical approaches.

We also investigated how PCOC was affected by errors in the root of the tree by moving the root to neighboring branches of the root. Incorrect rooting did not seem to have much of an impact on PCOC (supplementary fig. S15, Supplementary Material online).

Finally, analyzing our set of random positioning of convergent transitions, we did not observe an influence of the proportion of leaves in convergent clades on the performance of the three methods (supplementary fig. S7, Supplementary Material online). This differs from results obtained with the Identical method in Thomas et al. (2017) which showed that fewer convergent sites were detected when more taxa with the convergent phenotype were used. However their experimental setup differs from ours in that we operate under a fixed total number of taxa whereas they changed the total number of taxa.

PCOC's Performance Draws on the PC and OC Submodels

Figure 3 shows the contributions of the PC and OC submodels to the performance of PCOC on the simulations with a single amino acid profile on ancestral branches. PCOC shows a much better performance than both its submodels. In most conditions, on those simulations, OC seems to perform better than PC. However, we find that PC and OC perform best in different conditions. OC is most useful when branch lengths are short: in such conditions, encountering a substitution on a site provides a strong support for the OC model (fig. 3C and D). As soon as the expected number of substitutions approaches 0.5, the performance of OC drops markedly, because when a branch is longer than 0.5, a substitution is more likely than none, and then forcing one change on this branch has a minor impact on the transition probabilities. On the contrary, PC becomes more powerful as branch lengths

increase, because PC can then exploit a larger number of substitutions both on branches with the ancestral profile and on branches with the convergent profile to identify a site as convergent. Similar results were obtained on three other phylogenies (supplementary figs. S18–S39, Supplementary Material online).

Detection of Convergent Substitutions during Repeated Evolution of C4 Metabolism in Plants

Figure 4 represents sites with predicted convergent substitutions in the PEPC protein occurring jointly with the transition toward C4 metabolism in sedges (Besnard et al., 2009). Sites are represented if they have been found convergent in Besnard et al. (2009) (highlighted by a star), and/or by PCOC, using a threshold of 0.8. To detect convergent sites, Besnard et al. (2009) performed analyses of positive selection on the alignment, as well as comparative analyses with PEPC sequences from other plants. They proposed a set of 16 sites under positive selection (stars in fig. 4). In addition to our analysis of the empirical alignment, we inferred convergent substitutions on simulations performed on the same topology, placing convergent transitions on the same branches, and using the C60 set of profiles to evaluate the numbers of false positives and negatives we should expect when running PCOC. In these simulations, with the same proportion of convergent sites as defined in the Introduction, we found that PCOC should produce neither false positives nor false negatives for an alignment of the same size as the empirical alignment. Accordingly, there is an important overlap between PCOC and the set of convergent sites proposed in Besnard et al. (2009).

Their intersection contains eight sites (both with a star and in red, orange, or yellow on the top of fig. 4), and their union 20 sites. Only four sites predicted by PCOC have not been proposed in Besnard et al. (2009). Further, manual inspection of the two new sites with the best posterior probabilities (positions 584, 620) suggests that they have undergone substitutions inside each of the C4 clades, possibly on the branch ancestral to those clades, and toward amino acids that are seldom found in the gene sequences from C3 species. To better understand why PCOC detects these two sites, we looked at the separate posterior probability of the PC and OC models for each of those two sites. In both cases, the very high posterior probability of PCOC is due in large part to the support for OC (pp $>$0.99), but the support for PC is also superior to 0.5 (0.82 and 0.66 for positions 584 and 620, respectively). The two other sites with lower posterior probabilities (611 and 852) are not as convincing, and are identified only thanks to the OC component of PCOC. In addition, there are eight positions classified only by Besnard et al. (2009) as convergent and not predicted as convergent by PCOC, because they each underwent substitutions only in a subset of the C4 clades out of five: four for position 505, three for position 761, 839, two for positions 749, 770, 810, and 906, and one for position 733. For all those sites, there is no support for OC and at best weak support for PC, because those sites do not fit PCOC's definition of a convergent site.

We also performed analyses by using only the OC and PC submodels. PC only predicts seven sites as convergent (supplementary fig. S41, Supplementary Material online), and none of them are predicted in Besnard et al. (2009). Among the 14 sites it predicts as convergent (supplementary fig. S42, Supplementary Material online), OC finds eight sites also predicted by Besnard et al. (2009), like PCOC. The similarity between the sites selected by OC and those selected by PCOC is large, but two sites, sites 518 and 579, are predicted as convergent by OC but not by PCOC, and are not found in Besnard et al. (2009). Overall, PCOC's predictions appear to be derived mostly from the OC submodel rather than from the PC submodel, and are consistent with a previously published detailed analysis of an amino acid alignment. New positions suggested by PCOC represent candidates for convergent substitutions.

Discussion

Defining Convergent Amino-Acid Evolution

In this work, we have used a new definition of convergent events of genomic evolution, focusing on events that involve single amino acid substitutions that occur simultaneously (at the scale of single branches) with convergent phenotypic changes. This definition fits causative changes, or changes so intimately associated to the convergent phenotype that they occur very shortly after the phenotype has changed. We developed PCOC to simulate and detect changes according to this definition.

PCOC Accurately Detects Events of Convergent Amino-Acid Evolution

Compared with two previously proposed methods to detect convergent substitutions, PCOC has best power to detect changes that fit its definition. However, because PCOC relies on two submodels PC and OC, in principle, it can also capture convergent changes that do not perfectly fit the definition above (fig. 3). For instance, it may be able to detect substitutions that occur systematically on branches where the phenotype changed, irrespective of whether this was associated to a profile change, thanks to the OC component of PCOC. OC may thus recover sites detected by methods that look for accelerations on branches where the phenotypes changed (Partha et al., 2017). Similarly, thanks to its PC component, it may be able to detect sites that have not undergone substitutions on the branches where the phenotype changed, but that have undergone substitutions in underlying branches according to the convergent amino acid profile.

In practice, the PC submodel does not seem to contribute as much as the OC submodel, as seen from the C4 convergence example (fig. 4 and supplementary figs. S41 and S42, Supplementary Material online). It is unclear whether this is an inherent limitation of the data set, where branch lengths are at most 0.217, of the PC approach, or if better fitting profiles could improve PC's performance. Regarding branch lengths, PC could indeed contribute more than OC to PCOC on data sets where branch lengths are long (supplementary fig. S6,

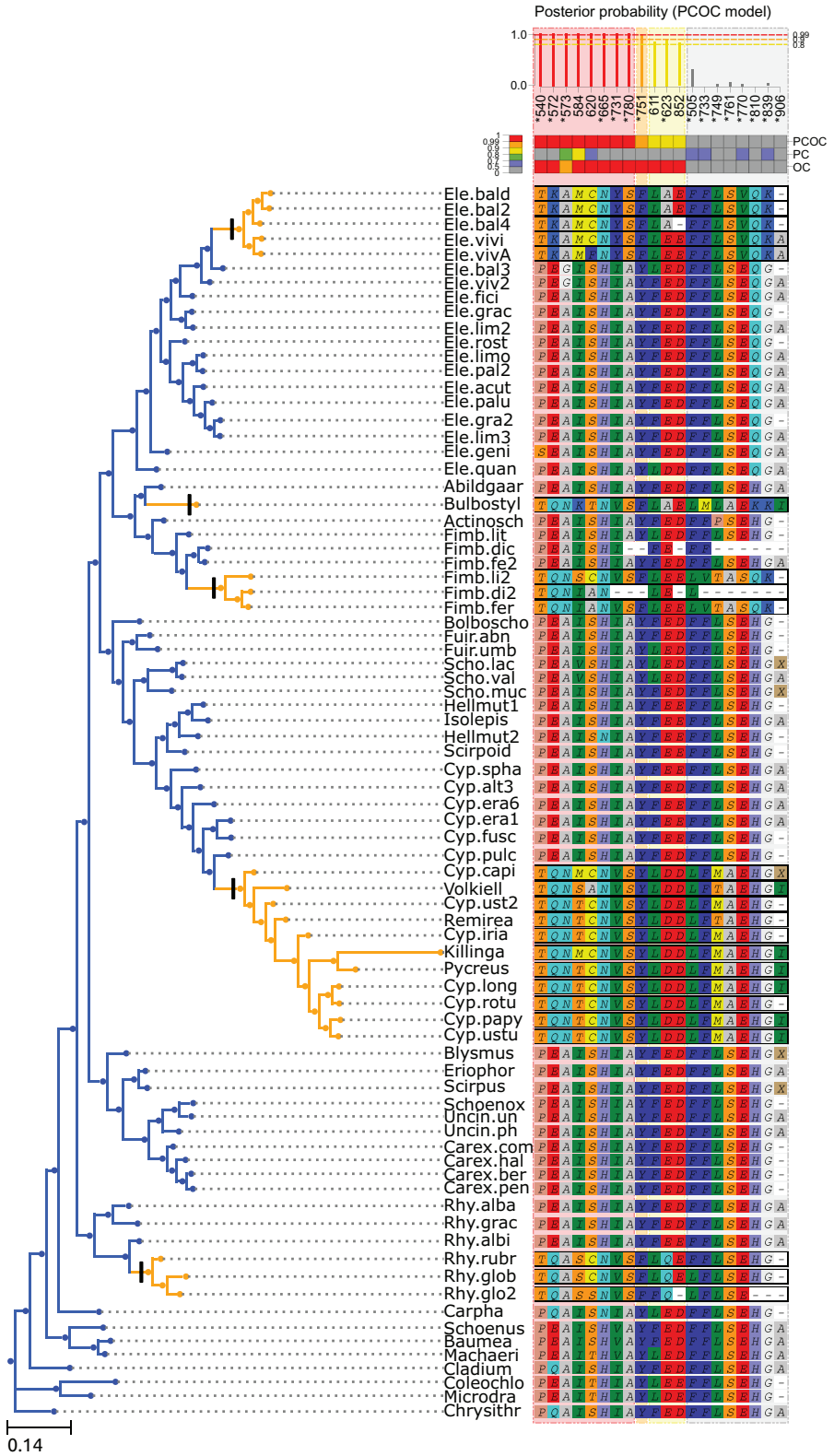

Fig. 4. Detection of convergent substitutions using the PCOC toolkit in the PEPC protein in sedges. Sites are ordered by their posterior probability of being convergent according to the PCOC model. Only sites with a posterior probability (pp) according to the PCOC model above a given threshold (here, 0.8) or sites detected in Besnard et al. (2009) (highlighted by a star) are represented. Sites are numbered according to Zea mays sequence (CAA33317) as in Besnard et al. (2009). Posterior probabilities for the PCOC, PC, and OC models are summarized by colors, red for pp$\geq$0.99, orange for pp$\geq$0.9, yellow for pp$\geq$0.8, and gray for pp $<$ 0.8.

Supplementary Material online). Regarding better fitting profiles, inferences performed under the same C60 model as that used for simulation show that the PC component is still minor compared with the OC component (supplementary fig. S5, Supplementary Material online), even when the profiles perfectly fit the simulation. However, more pointy profiles, where only a few amino acids have nonzero probability, may fit the data better. Such profiles are uncommon in the C60 and C10 sets, but they would better correspond to the particular subset of amino acids found at a given site in the convergent species.

Comparison between PCOC and Mutation-Selection Models

Parto and Lartillot (2017, 2018) have used a mutation-selection model to detect convergent evolution in single gene sequences. Mutation-selection models are codon models that attempt to distinguish the contribution of the mutational process at the DNA level from the contribution of the selection process, typically at the amino acid level. PCOC is a model of amino acid sequence evolution and therefore ignores phenomena that happen at the DNA level. In both PCOC and mutation-selection models, convergence is expected to be linked to changes in amino acid profiles; in fact, the PC submodel of PCOC can be thought of as an approximation of Parto and Lartillot's model, in the Maximum Likelihood framework, with a fixed set of profiles. However PCOC further adds the OC submodel, which enables it to detect repeated accelerations of the evolution of a site on the branches where the phenotype changed, even in the absence of a profile change. Further, PCOC benefits from a speed advantage over mutation-selection models as implemented in Parto and Lartillot (2017, 2018) for two reasons. First, because it works with protein sequences instead of codon sequences, which reduces the time required to compute the likelihood of a model. Second, because PCOC does not attempt to estimate amino acid profiles: instead it draws from profiles that have been estimated from large numbers of alignments. For these reasons PCOC can be used easily at the scale of whole genomes (see below).

PCOC Is a Tool to Simulate and Detect Convergent Genomic Evolution

We developed PCOC as a set of tools to perform simulation and detection of convergent evolution in sequences. These tools are user-friendly and require a gene tree provided by the user. It takes ~40 s to run the detection tool on a laptop for a data set with 79 leaves and 458 sites with the C10 set of profiles, and up to 20 min with the C60 set of profiles. The PCOC tool-kit is open source and available on GitHub https://github.com/CarineRey/pcoc with a tutorial. Simulations can be used to test the capacity of PCOC or other methods to detect convergent evolution on a specific data set, with its idiosyncratic characteristics. We have observed that the power of the methods depends on the number of independent convergent phenotypic changes, on branch lengths, and on the tree topology. These simulations can also be used to choose thresholds for controlling the amounts of false positives and false negatives. It is also easy to simulate sites with and without convergent evolution, for testing other methods.

Using PCOC with Genomic Data

We have not attempted to work at the level of entire gene sequences or even functional groups of genes, whereby the evidence obtained at the level of individual sites would be used collectively over the entire gene length or over several genes with a particular function to classify a gene or group of genes as convergent or not. However, other works have developed methods to work above the level of single sites (Chabrol et al., 2017; Marcovitz et al., 2017), and our method is compatible with these. Both these approaches detect convergent substitutions that fit the definition of Zhang and Kumar (1997) and Foote et al. (2015), but use different approaches to classify genes as convergent or not. Chabrol et al. (2017) combine their site-wise analysis with a procedure involving simulations according to a null model to classify genes as convergent or not. This simulation procedure is easy to perform with the PCOC toolkit. In particular, to investigate convergent evolution in a gene, we suggest that first convergent sites are identified using PCOC. Then, using the same tree and same parameters that were used for detection, one would perform simulations of a large number of sites with convergent evolution, and of sites without convergent evolution. PCOC would then be run on those simulated sites, which would provide the amount of true positives and false negatives. Such an approach can be used to assess the false discovery rate associated with the selection of candidate convergent sites in the empirical data. We applied this approach in our study of the C3/C4 alignment and described the procedure in the PCOC tutorial.

Possible Improvements to PCOC

PCOC relies on a set of profiles empirically built from a large number of alignments (Quang et al., 2008). These profiles were constructed to accurately model protein evolution in a time-homogeneous manner, and may be suboptimal for describing the evolution of sites that switch between two distinct profiles. Other profiles could be used although this has not yet been implemented in PCOC.

PCOC relies on a more general definition of convergent genomic events than the usual definition involving substitutions to a specific amino acid, but still does not account for other types of convergent events. For instance, PCOC has not been designed to deal with convergent relaxations of selection, which may contribute false positives. To filter out candidate genes that may be under convergent relaxations of selection, Marcovitz et al. (2017) used the numbers of divergent substitutions, that is, substitutions to different amino acids in the convergent species. PCOC does not rely on the definition of Zhang and Kumar (1997) and Foote et al. (2015), and therefore it is difficult to define such divergent substitutions. In our case, to identify convergent relaxations, we would rely on the fact that such phenomena should be associated with an accumulation of substitutions in the convergent branches, but with weaker preference for particular

amino acids. This would correspond to a shift from a pointy to a broad amino-acid profile. Detecting this requires to access the scores for all profiles in PCOC, and contrast their pointedness. This is not yet implemented in PCOC. To detect potential cases of convergent relaxations, we could also filter candidate genes based on branch lengths in convergent species: genes under relaxed selection specifically in lineages with the convergent phenotype are expected to have longer branches in those lineages.

Finally, the requirement linked to the OC submodel that convergent sites should undergo substitutions simultaneously with each convergent transition may be too strict: in some cases, it will be sufficient to consider a site as convergent if it undergoes substitutions on a large subset of those transitions. PCOC could be modified to fit such situations by using a mixture model, so that according to a probability P, the OC submodel would be used on the branches subtending convergent clades, and according to $1 - P$, the OC submodel would not be used. The estimation of this single parameter P would probably not incur an important computational cost.

Materials and Methods

A New Probabilistic Model of Convergent Evolution

We adopt a biochemical point of view and consider that adaptive convergence drives the preference at a given site toward amino acids that share specific properties. We do not define those properties *a priori*, but instead consider a set of amino acid profiles, empirically built from a large number of alignments (Quang et al., 2008). These profiles serve as a proxy to amino acid fitnesses at a given site: more frequent amino acids in the profiles have higher stationary frequencies, as in mutation-selection models (Parto and Lartillot, 2017). Following this Profile Change (PC) model, a convergent site will exhibit a preference in all convergent clades toward a specific profile, different from an ancestral profile, whereas a nonconvergent site will remain with the same profile in all the tree. In our simulations, we also consider the possibility that a nonconvergent site alternates randomly between a few different profiles along the phylogeny on branches with the ancestral phenotype, but switches to a particular single profile on branches with the convergent phenotype. In addition, we consider that a substitution must occur when a convergent site switches from the ancestral profile to the convergent profile, and to this end, we implemented the OneChange (OC) model. The combination of PC and OC into PCOC models the situation where the convergent phenotype is tightly linked to a given type of amino acid at a certain position, so much so that it can be considered necessary or at least highly advantageous for the phenotype to have one of the fittest amino acids from the convergent profile at this position. Our approach therefore does not attempt to model positions that change to a convergent amino acid profile after the switch from the ancestral to the convergent phenotype has occurred, and which would be noncausative substitutions. Such sites would be appropriately modeled by PC alone, but not quite as well by PCOC.

PCOC Tool-Kit: A Tool for Simulation and Inference of Convergent Substitutions

Simulation Process

To evaluate the ability of detection methods to detect convergent sites, we performed two types of simulation. In one type, we simulate under convergent evolution, varying the parameters of the evolutionary model (e.g., varying the number of convergent transitions). This allows us to estimate the sensitivity of the methods. In the other type, we simulate without any event of convergent evolution. This allows us to assess the specificity of the methods. In each case, we simulated 1,000 sites. To simulate convergent evolution, we aimed at placing events of convergent evolution uniformly on a species tree, irrespective of branch length. We were interested in the impact of the number of events of convergent evolution on our power to detect it and placed between two and seven events. To avoid any bias in the location of these events, in all cases, we drew uniformly exactly seven potential events, so that all events were in independent clades. From these seven events, we then subsampled the desired number of events of convergence. All branches in the clades below those events were labeled "convergent," and all other branches (above these events and in the nonconvergent clades) labeled "ancestral." A particular amino acid fitness profile c_x was used for ancestral branches, another c_y for convergent branches and we applied the OneChange model with the c_y profile on the branch where the switch to the convergent phenotype was positioned. The switch was placed at the very beginning of the branch. We randomly drew amino acid profiles from the C60 model (Quang et al., 2008) (supplementary fig. S1, Supplementary Material online) and did not attempt to test all pairs of C60 profiles in order to save computation time and slightly reduce our carbon footprint. We also performed additional simulations where more than one profile was used on branches with the ancestral phenotype (supplementary figs. S8–S10, Supplementary Material online). Although C60 was built to describe amino acid sequence evolution in a time-homogeneous manner, we assume that this limited set of profiles provides a rough approximation to the set of possible amino acid profiles. In addition to the simulations with convergent events that we used to measure the proportion of True Positives (TP) and False Negatives (FN) of the methods, we performed similar simulations (i.e., using the same trees) where the ancestral profile is used for all branches of the phylogeny, to measure their proportion of True Negative (TN) and False Positive (FP).

Sequence evolution was simulated along the phylogenetic tree using the model associated to each branch, with rate heterogeneity across sites according to a Gamma distribution discretized in four classes (Yang, 1994) with the α parameter set to 1.0, using bppseqgen (Dutheil and Boussau, 2008).

Inference Methods

For each of the three compared approaches, we have to infer if a site is convergent.

For the PCOC, PC, OC, and the Topological methods, the decision is controlled by a threshold on the a posteriori probability of the convergent model versus the null model, using a uniform prior. We used bppml (Dutheil and Boussau, 2008) to measure the likelihood of each model.

To compare the studied methods fairly, we tuned this threshold for each method to reach its optimal performance. We use the Matthews correlation coefficient (MCC) (Matthews, 1975) as a measure of the performance because the MCC takes into account the proportions of positives and negatives which are expected to be heavily biased in our case as we saw in the Introduction. Therefore, we chose the threshold so as to maximize the MCC of each method using the proportions of the Introduction example (supplementary fig. S2, Supplementary Material online).

Below, we describe the procedure we adopted to call a site as convergent for each of the three compared approaches.

PCOC Approach. In accordance with our definition of convergence and our simulation procedure, we used a model-based inference to detect convergent substitutions. We used the branch lengths that had been used for simulation for inference, but we checked that the impact of errors in branch lengths on inference was minimal (supplementary figs. S11 and S12, Supplementary Material online). We used the C10 set of profiles from the CAT model (Quang et al., 2008), containing 10 profiles, to be in a more realistic scenario where the CAT profiles used in the simulation (C60) are not those used for inference. However, we checked that using the same C60 set of profiles for inference and simulation yielded very similar results (supplementary fig. S5, Supplementary Material online). For each i in {1...10} and for each j in {1...10} such as $i \neq j$, we calculated the likelihood of two models: one, $M0_i$, in which the same profile c_i is used on all branches, and another model, $M1_{i/j}$in which the profile c_i is used only on "ancestral" branches, and the profile c_j on "convergent" branches. We explain in details how one can compute the likelihood under $M1$ in supplementary section 2, Supplementary Material online. Then, we compared the likelihoods of two average models, $M0$ and $M1$. The likelihood of $M0$ is computed as the mean of the likelihoods of the $M0_i$ models and the likelihood of $M1$ as the mean of the likelihoods of the $M1_{i/j}$ models.

We classified each site as a positive or a negative using an Empirical Bayes approach. A positive is a site predicted to have evolved according to the heterogeneous model $M1$, and a negative according to the homogeneous model $M0$. For each site i, we computed the likelihood of the $M1$ model $P(s_i|M1)$ and of $M0$ $P(s_i|M0)$. We computed the empirical posterior probability of $M1$ with a uniform prior on each model: $P(M1|s_i) = P(s_i|M1)/(P(s_i|M1) + P(s_i|M0))$. A positive is defined such that $P(M1|s_i) > 0.99$ for the PCOC and the OC models and 0.9 for the PC model.

Topological Approach. We also performed comparisons of likelihoods with two different topologies, as in Parker et al. (2013). The rationale of this approach is that, for sites showing convergence, the phylogenetic signal would prefer to cluster together convergent branches. So, for these sites, the true tree should be less likely than the tree for which the convergent branches are together, named "convergent tree." We present in Supplementary Material, the algorithm we used to construct convergent trees and an example of such a "convergent tree" (supplementary fig. S3, Supplementary Material online).

We computed for each site the mean of the likelihoods with the ancestral model c_i applied on all branches for each i in {1...10} for the true and the convergent trees. And, as in the method based on heterogeneous models, we considered a site as convergent when the empirical posterior probability of the convergent tree was >0.9.

Approach Based on Ancestral Reconstruction. To detect convergent substitutions as in Zou and Zhang (2015b), Thomas and Hahn (2015), and Foote et al. (2015), we considered the branches ancestral to convergent clades.

We declared a substitution on a given site as convergent if all substitutions on the ancestral branches were toward the exact same amino acid.

Statistical Measures of the Performance

Finally, we measured the power of the three methods of detection on simulations using their specificity, sensitivity, and MCC (supplementary figs. S4, S6, S7, S9–S12, S18–S24, S26–S32, and S34–S40, Supplementary Material online).

Simulations to Assess the Impact of the Number of Convergent Transitions

We used the simulator and benchmark tool of the PCOC toolkit to produce the data used in the panels A and B of figures 2 and 3. We extracted the subtree containing mammals only from the Ensembl Compara tree (Herrero et al., 2016; Yates et al., 2016), and used it to position a random number X of convergent events between two and seven. We repeated this procedure 160 times. For each random assignment of convergent events, we sampled 10 pairs of C60 profiles and for each pair simulated 1,000 convergent sites using both profiles and 1,000 nonconvergent sites using only the ancestral profile.

Simulations to Assess the Impact of Branch Lengths

We used the simulator and benchmark tool of the PCOC toolkit to produce the data used in the panels C and D of figures 2 and 3. We used the same tree as above, and set all its branch lengths to values between 0.01 and 1. For each branch length value, we performed 32 replicates by randomly placing five events of convergent evolution in the phylogeny. For each random assignment of convergent events, we simulated alignments with 10 pairs of C60 profiles and for each pair simulated 1,000 convergent sites using both profiles and 1,000 nonconvergent sites using only the ancestral profile.

PCOC Tool-Kit: Detector Tool, Test on Real Data

We used the detector tool of the PCOC toolkit to build figure 4. It takes ~40 s to run on a laptop for a data set with 79 leaves and 458 sites with the C10 set of profiles, and up to 20 min with the C60 set of profiles. The nucleotide alignment and tree topology come from Besnard et al. (2009). As the detector tool of the PCOC toolkit needs a tree and an

amino-acid alignment, we inferred branch lengths on the fixed topology using phyml (Guindon et al., 2010) with the GTR model using the nucleotide alignment and obtained the amino-acid alignment by translating the nucleotide sequences. For clarity, we only showed sites if they had a posterior probability >0.8 according to the PCOC model (see supplementary figs. S41 and S42, Supplementary Material online, for the PC and OC models).

Conclusion

We have proposed a new definition of convergent substitutions that contains and relaxes the commonly used definition from Zhang and Kumar (1997). We have implemented a model embodying this definition into simulation and inference methods, and find that our method has better power to detect convergent changes than previously proposed approaches. It is sufficiently fast to be applied on large data sets, and should be useful to detect traces of convergent sequence evolution on genome-scale data sets.

Supplementary Material

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

Acknowledgments

This work was performed using the computing facilities of the CC LBBE PRABI. We thank Thibault Lorin, Vincent Lanore, Gilles Didier, Philippe Veber, Nicolas Lartillot, and Vincent Daubin for fruitful discussions. Fundings: ANR-15-CE32-0005 "Convergenomix," ANR-10-BINF-01-01 "Ancestrome," ANR-11-JSV6-00501 "Convergdent." We thank Pauline Sémon for the PCOC logo. The work presented in this manuscript involved >400 computer.days.

References

Bazykin GA, Kondrashov FA, Brudno M, Poliakov A, Dubchak I, Kondrashov AS. 2007. Extensive parallelism in protein evolution. *Biol Dir*. 2(1):20.

Besnard G, Muasya AM, Russier F, Roalson EH, Salamin N, Christin P-A. 2009. Phylogenomics of C4 photosynthesis in sedges (Cyperaceae): multiple appearances and genetic convergence. *Mol Biol Evol*. 26(8):1909–1919.

Castoe TA, de Koning APJ, Kim H-MM, Gu W, Noonan BP, Naylor G, Jiang ZJ, Parkinson CL, Pollock DD. 2009. Evidence for an ancient adaptive episode of convergent molecular evolution. *Proc Natl Acad Sci U S A*. 106(22):8986–8991.

Chabrol O, Royer-Carenzi M, Pontarotti P, Didier G. 2017. Detecting molecular basis of phenotypic convergence. *bioRxiv* doi: 10.1101/137174.

Dutheil J, Boussau B. 2008. Non-homogeneous models of sequence evolution in the Bio++ suite of libraries and programs. *BMC Evol Biol*. 8(1):255.

Foote AD, Liu Y, Thomas GWC, Vinař T, Alföldi J, Deng J, Dugan S, van Elk CE, Hunter ME, Joshi V, et al. 2015. Convergent evolution of the genomes of marine mammals. *Nat Genet*. 47(3):272–275.

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.

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* 2016:bav096.

Jarvis E, Mirarab S, Aberer A, Li B, Houde P. 2014. Whole-genome analyses resolve early branches in the tree of life of modern birds. *Science* 346(6215):1126–1138.

Marcovitz A, Turakhia Y, Gloudemans M, Braun BA, Chen HI, Bejerano G. 2017. A novel unbiased test for molecular convergent evolution and discoveries in echolocating, aquatic and high-altitude mammals. *bioRxiv* doi: 10.1101/170985.

Matthews B. 1975. Comparison of the predicted and observed secondary structure of t4 phage lysozyme. *Biochim Biophys Acta Protein Struct*. 405(2):442–451.

Parker J, Tsagkogeorga G, Cotton JA, Liu Y, Provero P, Stupka E, Rossiter SJ. 2013. Genome-wide signatures of convergent evolution in echolocating mammals. *Nature* 502(7470):228–231.

Partha R, Chauhan BK, Ferreira Z, Robinson JD, Lathrop K, Nischal KK, Chikina M, Clark NL. 2017. Subterranean mammals show convergent regression in ocular genes and enhancers, along with adaptation to tunneling. *eLife* 6.

Parto S, Lartillot N. 2017. Detecting consistent patterns of directional adaptation using differential selection codon models. *BMC Evol Biol*. 17(1):147. https://elifesciences.org/articles/25884.

Parto S, Lartillot N. 2018. Molecular adaptation in rubisco: discriminating between convergent evolution and positive selection using mechanistic and classical codon models. *PLoS One* 13(2):e0192697.

Quang Le S, Gascuel O, Lartillot N. 2008. Empirical profile mixture models for phylogenetic reconstruction. *Bioinformatics* 24(20):2317–2323.

Rokas A, Carroll SB. 2008. Frequent and widespread parallel evolution of protein sequences. *Mol Biol Evol*. 25(9):1943–1953.

Schenk JJ, Rowe KC, Steppan SJ. 2013. Ecological opportunity and incumbency in the diversification of repeated continental colonizations by muroid rodents. *Syst Biol*. 62(6):837–864.

Studer RA, Christin P-A, Williams MA, Orengo CA. 2014. Stability-activity tradeoffs constrain the adaptive evolution of rubisco. *Proc Natl Acad Sci U S A*. 111(6):2223–2228.

Tamuri AU, dos Reis M, Hay AJ, Goldstein RA. 2009. Identifying changes in selective constraints: host shifts in influenza. *PLoS Comput Biol*. 5(11):e1000564–e1000514.

Thomas GW, Hahn MW. 2015. Determining the null model for detecting adaptive convergence from genomic data: a case study using echolocating mammals. *Mol Biol Evol*. 32(5):1232–1236.

Thomas GW, Hahn MW, Hahn Y. 2017. The effects of increasing the number of taxa on inferences of molecular convergence. *Genome Biol Evol*. 9(1):213–221.

Yang Z. 1994. Estimating the pattern of nucleotide substitution. *J Mol Evol*. 105–111.

Yates A, Akanni W, Amode MR, Barrell D, Billis K, Carvalho-Silva D, Cummins C, Clapham P, Fitzgerald S, Gil L, et al. 2016. Ensembl 2016. *Nucleic Acids Res*. 44(D1):D710–D716.

Zhang J, Kumar S. 1997. Detection of convergent and parallel evolution at the amino acid sequence level. *Mol Biol Evol*. 14(5):527–536.

Zhen Y, Aardema ML, Medina EM, Schumer M, Andolfatto P. 2012. Parallel molecular evolution in an herbivore community. *Science* 337(6102):1634–1637.

Zou Z, Zhang J. 2015a. Are convergent and parallel amino acid substitutions in protein evolution more prevalent than neutral expectations? *Mol Biol Evol*. 32(8):2085–2096.

Zou Z, Zhang J. 2015b. No genome-wide protein sequence convergence for echolocation. *Mol Biol Evol*. 32(5):1237–1241.

3.2.2 Capacité des méthodes existantes à détecter de la convergence dans des situations réelles

3.2.2.1 Avant-propos

De manière concomitante à la fin de la réalisation de PCOC, Bastien Boussau, mon co-encadrant de thèse, s'est vu proposer par Nathan Clark, un scientifique très investi dans le champ de la détection de la convergence, la réalisation d'une revue pour comparer les différentes méthodes de détection de la convergence. Cette revue devait s'inscrire dans une suite d'articles consacrés à la convergence dans Philosophical Transactions B.

En collaboration avec Vincent Lanore, Philippe Veber, Laurent Gueguen, Nicolas Lartillot, Marie Sémon et Bastien Boussau, nous avons répondu favorablement à cette proposition. Cependant, nous nous sommes rapidement rendu compte que si l'on voulait comparer les méthodes de détection de la convergence, nous devions les tester sur des données et qu'il ne s'agirait pas d'une revue mais plutôt d'une analyse complète.

L'article est composé de quatre parties : une présentation des facteurs pouvant mener à la convergence et notamment les facteurs confondants, une description des méthodes permettant de détecter la convergence, une comparaison de leurs performances sur des données simulées incluant des facteurs confondants et enfin une comparaison des résultats de ces méthodes sur des données réelles. Par rapport à l'article présentant PCOC, nous avons pris en compte quatre autres méthodes supplémentaires : diffsel (PARTO et LARTILLOT, 2018), msd (CHABROL et collab., 2018), Tdg09 (TAMURI et collab., 2009) et multinomial qui n'étaient pas existantes ou pas destinées explicitement à la détection de la convergence. On peut noter que diffsel est développé au sein de l'équipe par Nicolas, Vincent, Philippe et Bastien.

La manière de simuler les données a été une grande problématique lors de la réalisation de cette analyse afin de ne pas avantager de méthodes et d'être au plus proche de données réelles. Nous avons repris le même utilitaire de simulation de séquences (BPPSEQGEN de Bio++) que lors de l'article sur PCOC mais nous avons utilisé des profils d'acides aminés réels tirés de (BLOOM, 2016). Nous avons remplacé la composante OC (One Change) du simulateur qui permettait de favoriser l'apparition de substitutions convergentes par un paramètre mécanistique, Ns, qui permet de moduler l'efficacité de la sélection et que nous avons réglé à l'aide de données réelles. Il est important de noter que cette modification a pour conséquence que les sites convergents simulés ne contiennent plus nécessairement de substitutions au niveau de toutes les transitions convergentes comme c'était le cas avec le simulateur de PCOC.

Pour finir, l'ensemble des analyses présentes dans ce papier est produit par un unique pipeline permettant la reproductibilité de l'analyse. De plus, nous l'avons prévu modulable afin de pouvoir ajouter facilement des méthodes de simulation ou de détection et donc qu'il puisse être réutilisé dans d'autres cas.

3.2.2.2 Article : Detecting adaptive convergent amino acid evolution

Comme précisé précédemment, cet article s'inscrit dans une série d'articles traitant de la convergence et a été publié dans Philosophical Transactions B le 03 juin 2019 (http://dx.doi.org/10.1098/rstb.2018.0234).

Le code permettant de reproduire les analyses est disponible sur le gitlab de l'in2p3 (https://gitlab.in2p3.fr/pveber/reviewphiltrans). Les données intermédiaires pour éviter de refaire tourner complètement le pipeline sont disponibles sur Dryad https://doi.org/10.5061/dryad.57hr00q/1.

PHILOSOPHICAL TRANSACTIONS B

royalsocietypublishing.org/journal/rstb

Research

Cite this article: Rey C, Lanore V, Veber P, Guéguen L, Lartillot N, Sémon M, Boussau B. 2019 Detecting adaptive convergent amino acid evolution. *Phil. Trans. R. Soc. B* **374**: 20180234.
http://dx.doi.org/10.1098/rstb.2018.0234

Accepted: 25 February 2019

One contribution of 16 to a theme issue 'Convergent evolution in the genomics era: new insights and directions'.

Subject Areas:
genomics, bioinformatics, evolution, computational biology

Keywords:
convergent evolution, genomics, molecular evolution, C3/C4, phylogenetics, probabilistic models

Author for correspondence:
Bastien Boussau
e-mail: boussau@gmail.com

†These authors contributed equally to the study.
‡These authors contributed equally to the study.

Electronic supplementary material is available online at https://doi.org/10.6084/m9.figshare.c.4467500.

THE ROYAL SOCIETY PUBLISHING

Detecting adaptive convergent amino acid evolution

Carine Rey[1,†], Vincent Lanore[2,†], Philippe Veber[2], Laurent Guéguen[2], Nicolas Lartillot[2], Marie Sémon[1,‡] and Bastien Boussau[2,‡]

[1]ENS de Lyon, CNRS UMR 5239, INSERM U1210, LBMC, Univ Lyon, Université Claude Bernard Lyon 1, F-69007 Lyon, France
[2]CNRS UMR 5558, LBBE, Univ Lyon, Université Claude Bernard Lyon 1, F-69100 Villeurbanne, France

NL, 0000-0002-9973-7760; BB, 0000-0003-0776-4460

In evolutionary genomics, researchers have taken an interest in identifying substitutions that subtend convergent phenotypic adaptations. This is a difficult question that requires distinguishing *foreground* convergent substitutions that are involved in the convergent phenotype from *background* convergent substitutions. Those may be linked to other adaptations, may be neutral or may be the consequence of mutational biases. Furthermore, there is no generally accepted definition of convergent substitutions. Various methods that use different definitions have been proposed in the literature, resulting in different sets of candidate foreground convergent substitutions. In this article, we first describe the processes that can generate foreground convergent substitutions in coding sequences, separating adaptive from non-adaptive processes. Second, we review methods that have been proposed to detect foreground convergent substitutions in coding sequences and expose the assumptions that underlie them. Finally, we examine their power on simulations of convergent changes—including in the presence of a change in the efficacy of selection—and on empirical alignments.

This article is part of the theme issue 'Convergent evolution in the genomics era: new insights and directions'.

1. Introduction

It is difficult to replicate experiments when we study evolutionary biology. However, one can benefit from natural replicates that have arisen through time and across taxa. Indeed, lineages that have adapted independently to a given environmental constraint can be seen as having been subjected independently to the same 'experimental' conditions. When lineages subjected to the same conditions evolve similar phenotypes, they are said to have *converged* in their phenotypes. In the rest of the article, we call 'convergent lineages' lineages that have undergone such convergent phenotypic evolution. In evolutionary genomics, researchers have taken an interest in identifying substitutions that subtend those convergent phenotypes.

We call these causative substitutions 'foreground convergent substitutions'. We distinguish them from 'background convergent substitutions' that include substitutions that may be confused with 'foreground convergent substitutions' but that have no phenotypic consequences on the studied convergent phenotype.

Foreground convergent substitutions may be adaptive, i.e. they fixed through positive selection, or non-adaptive, i.e. they fixed through a relaxation of selection (electronic supplementary material, figure S1). The latter may, for instance, occur in cases of regressive evolution, where a gene is no longer needed in a particular environment. Being able to distinguish these two types of convergent substitutions provides information about the underlying evolutionary process.

To identify foreground convergent substitutions, many methods search for substitutions that are correlated with the phenotype. These are substitutions that have occurred repeatedly in convergent lineages, towards the same derived state, or towards similar derived states (e.g. towards amino acids with similar biochemical profiles). Methods vary in how they quantify the similarity between substitutions, resulting in different sets of candidate substitutions [1–3]. Finding foreground convergent (causative) substitutions among many substitutions in genomes containing billions of sites is a challenge in modern bioinformatics.

Several processes at work in genome evolution may affect the number of convergent substitutions in a given dataset. For instance, mutational biases, changes in recombination rates, biased gene conversion (bGC) or changes in population size may inflate or diminish the number of convergent substitutions. They may also affect differently the numbers of foreground and background convergent substitutions. Although this has not been studied yet, one may assume that these complex processes make it harder for methods to distinguish foreground from background convergent substitutions.

In this article, we first describe some processes contributing adaptive and non-adaptive foreground convergent substitutions in coding sequences. Second, we review existing methods to detect foreground convergent amino acid substitutions and expose the assumptions that underlie them. Third, we examine their power on simulations of convergent changes—including in the presence of variations in selection efficacy—and on two empirical alignments.

(a) Defining adaptive convergent amino acid evolution

In this section, we examine how foreground convergent substitutions can arise through adaptive processes. To this end, it is useful to first discuss adaptive genomic evolution in general. Adaptive genomic evolution is expected to occur when constraints on the phenotype change, which alters the selective pressure at some sites in the genome. Individuals with mutations that provide an increased fitness in the new environment have a reproductive advantage. Such mutations then increase in frequency and can eventually fix. The fixation of one or more of these mutations can, in turn, change the selective pressure on other sites of the genome through epistatic interactions [4].

The characteristics of the *fitness landscape* have an impact on how likely adaptive convergent evolution is. The fitness landscape describes the mapping between genotypes and fitness in a species, for a given set of constraints on the phenotype. Because it treats the genotype as a whole, it naturally considers all the sites of the genome and their interactions at once. If it is highly peaked, it means that only one genotype can provide the largest fitness. In that case, one can expect that several related species under the same constraints on their phenotype may adaptively converge towards the same genotype, i.e. adaptive convergent substitutions are likely. Instead, if the fitness landscape is very flat, different genotypes can provide similar fitnesses, so that several related species under the same constraints on their phenotype may move towards different genotypes, making adaptive convergent substitutions less likely.

These intuitive considerations should make it clear that a good mechanistic model of convergent evolution needs to consider the entire genome at the same time, along with the fitness landscape, to take into account all the dependencies between sites. For computational reasons, and because fitness landscapes are only rarely studied experimentally [4], such a model is currently out of reach. Instead, sites share some general parameters ruling their evolution (e.g. branch lengths, some parameters of the substitution matrices) but, conditionally on those parameters, each site is typically modelled independently of the others. In addition, many models make simplifying assumptions; for instance, fitness landscapes only depend on the phenotype, and not on the lineage under consideration. All models that have been developed to detect convergent genomic evolution assume such site-independent models.

In this article, we propose to define convergent evolution through the comparison of coding sequences across species. Coding sequences offer a window into where the mutation process and the selective process meet. Indeed, non-synonymous mutations (i.e. mutations that change the encoded amino acid) might be strongly selected, while synonymous mutations (i.e. mutations that do not change the encoded amino acid) should be neutral or weakly counter-selected. Natural models to study coding sequences are codon models, in which one site is made of three nucleotides encoding a particular amino acid. The simplest codon models consider one site at a time, independently of other sites, and distinguish between synonymous and non-synonymous substitutions. They assume that synonymous substitutions provide a proxy for the rate of fixation of neutral substitutions, while all non-synonymous substitutions have the same rate of fixation, which depends on selection efficacy [5,6]. More sophisticated codon models distinguish between different amino acid changing substitutions and assume that different amino acids provide different fitnesses. Such models use amino acid fitness profiles—which we simply call *amino acid profiles* in the rest of the article (figure 1) [7]. Some of the richest models allow individual fitness profiles for different sites [8,9]. Overall, codon models provide a convenient framework to define adaptive convergent amino acid evolution.

In a simple model that considers one codon at a time, adaptive convergent evolution can result from an increase in the selective pressure or from a change in its nature. Increases in the pressure would mean that, in the amino acid profile at a given codon, the amino acids that provided high fitness before the environmental change provide even higher fitness, while amino acids providing low fitness before the change now provide even lower fitness (figure 1*a*, 'Scaling of Selection Efficacy'). It has become more important for the organism to have particular amino acids at this position. For example, this could be associated with a lifestyle where the function of the protein has become more important than it was. Changes in the nature of the selective pressure manifest themselves by a change between two amino acid fitness profiles, referred to as 'ancestral' and 'convergent' from now on (figure 1*a* 'profile change'). As opposed to increases in the pressure, profile changes alter which amino acids are the most fit at a given position.

In the rest of the article, we distinguish between two types of adaptive convergent substitutions (figure 1*b*), because detection methods vary in their ability to detect each type. We call 'type 1 substitutions' the early substitutions that

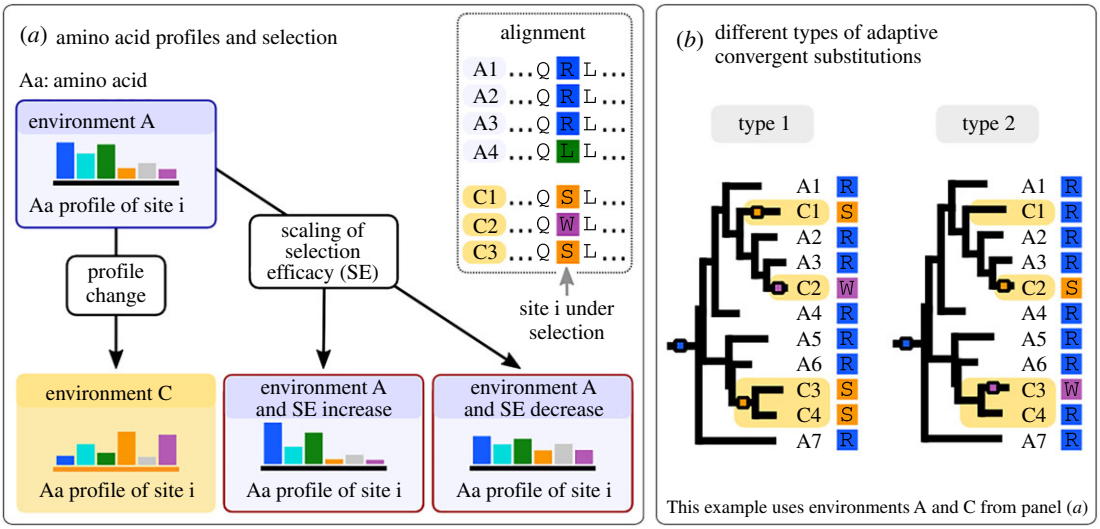

Figure 1. Categories of adaptive and non-adaptive convergent amino acid evolution. (*a*) At a particular position in a protein, some amino acids provide better fitness than others. This is represented by coloured bars for six amino acids, the bigger the bar the higher the fitness. In the ancestral environment A, amino acids blue and green provide the highest fitness, whereas in the convergent environment C, amino acids orange and purple provide the highest fitness. Increasing the selection efficacy makes the profiles more pointed, while decreasing it makes them more flat, but the amino acid relative rank does not change. Decreases of the selection efficacy are not adaptive, while the two other types of changes are. (*b*) Species with the convergent phenotype are named C* and species with the ancestral phenotype are named A*. Substitutions are represented by small boxes on the branches. We distinguish two types of adaptive convergent substitutions. Type 1 are substitutions that occur systematically on the branch where the phenotype changes, at the transition between Ancestral and Convergent environments (A–C). Type 2 are substitutions that occur on later branches (e.g. in the branch leading to C3).

occur on the branch where the phenotype changed, and 'type 2 substitutions' those that fix after type 1 substitutions, on subsequent branches.

(b) Non-adaptive background convergent amino acid evolution

Background convergent amino acid substitutions may be linked to convergent phenotypes that have not been detected, or not linked to convergent phenotypes, and possibly have no phenotypic consequences. In this latter case, they arise non-adaptively. Some number of such non-adaptive background convergent substitutions is expected, if only because there are only 20 possible amino acids. Further, the structure of the genetic code and the characteristics of the mutation process (e.g. that transitions are more frequent than transversions) all contribute to making some amino acid substitutions more likely than others and therefore increase the probability that they will be convergent.

In addition, fixation and mutation biases could create patterns resembling adaptive convergent evolution, and possibly adaptive foreground convergent evolution. In particular, GC-bGC is a fixation bias that favours G or C alleles over A or T alleles and is widespread across the tree of life [10,11], and CpG hypermutability is a well-known mutation bias. bGC is most intense in regions of the genome that recombine frequently and has a stronger effect over time in species with large effective population sizes and short generation times. Those two characteristics have appeared independently several times in the tree of life. Because of bGC, one can expect to detect similar changes to GC alleles in the species sharing these characteristics, even without any adaptive value to having GC alleles instead of AT at those positions. This phenomenon seems to be strong enough to affect single gene phylogenies in birds [12,13] and may be an important driver of background convergent sequence evolution. CpG hypermutability results from a higher rate of mutations of methylated CG dinucleotides and could also contribute to background convergent sequence evolution. It has also been shown to promote foreground convergent evolution, with recurrent changes at the same CpG site in passerine bird haemoglobin [14].

Repeated and independent global relaxations of selection could also create background convergent evolution. If the phenotypic change is linked to a genome-wide decrease in selection efficacy, e.g. through a decrease in the effective population size [15], mutations that used to be counter-selected become tolerated. Because of the structure of the genetic code, those mutations could result in similar amino acid substitutions in lineages undergoing the decrease in selection efficacy.

Finally, epistatic interactions between sites in the genome or within a protein can create non-adaptive convergent amino acid evolution [16,17]. The same mutation at a particular site can occur in independent lineages simply because by chance sites that are in epistatic interactions with it happen to be in the same state in those lineages. The mutation therefore fixes not because of an adaptation to a new environment, but because of the states of interacting sites. Such non-adaptive convergent evolution is more likely in closely related lineages than in distant lineages [16]. It is difficult to know how frequent such events are, in part because most of the models used to study protein evolution have ignored epistatic interactions. We do not study such phenomena in this article but acknowledge that they may be an important confounding process in the search for adaptive convergent amino acid evolution.

(c) Detecting adaptive foreground convergent amino acid substitutions

Several methods have been designed to detect adaptive convergent amino acid evolution. We list them below and attempt to predict their relative strengths and weaknesses,

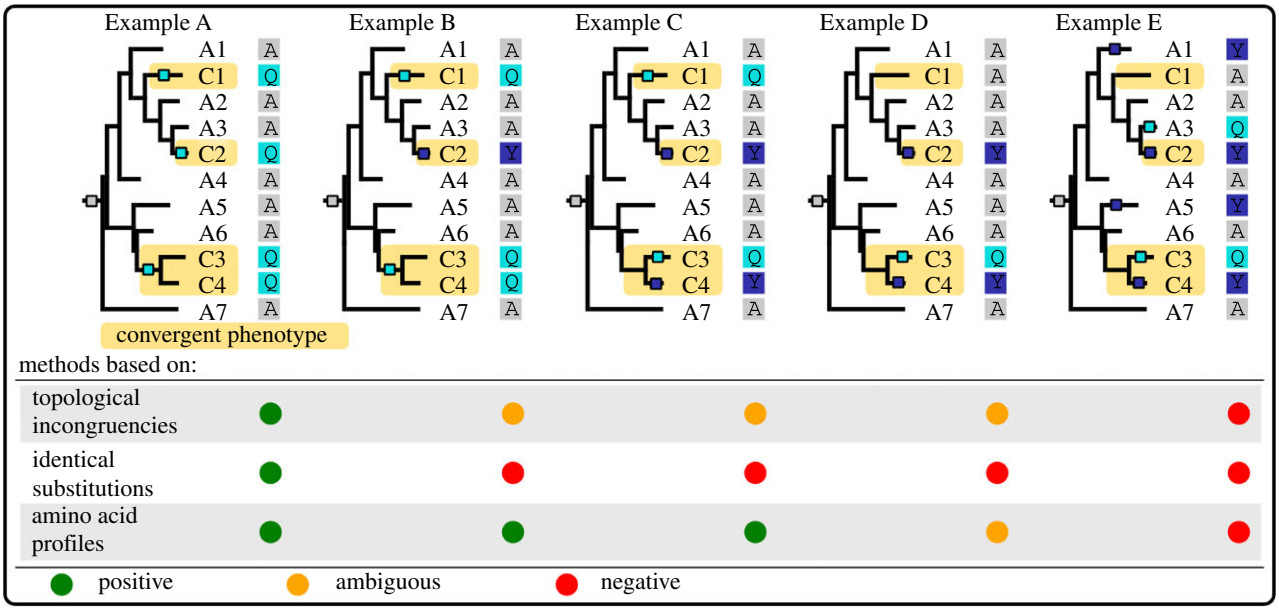

Figure 2. Cartoon examples of the types of sites targeted by each type of method. The tree topologies and species are the same in all examples. Species with the convergent phenotype are named C*, those with the ancestral phenotype A*; the transitions between ancestral and convergent phenotype occur where the subtrees become shaded in yellow. Coloured squares on the branches of the phylogeny indicate substitution events, with the colour corresponding to the new amino acid. In Example A, every time the phenotype changes, a substitution occurs towards amino acid Q (type 1 substitutions to a single amino acid). This is an ideal case for the methods based on identical substitutions and should be detectable by all methods. Example B shows a site that has undergone a profile change, whereby two different amino acids, Q and Y, have good fitness in the convergent case. All methods but the identical may detect such changes, although this depends on how different the ancestral and the convergent profiles are [18]. Example C is similar to Example B except that some substitutions occurred after the phenotype has changed (type 2 substitutions), not simultaneously with the phenotype change. Example D is similar to Example C except that the amino acid change only occurred three times out of four: this makes it more controversial and harder to detect. But if the change in profile is strong enough, profile methods should be able to detect it. Example E shows a case where the evolution of the site does not seem to correlate with the convergent/ancestral state of the species. We do not expect the methods to detect such a site, but some such sites will nevertheless come out as false positives.

in particular, their capacity to detect type 1 and type 2 adaptive convergent substitutions. Figure 2 presents in cartoon format the type of convergent sites that each type of method should be able to detect. None of the following methods have been designed to detect convergent increases or decreases in selection efficacy, so we expect they should do much better at detecting convergent profile changes than convergent changes in selection efficacy. All methods except one (msd, [19]) assume that convergent lineages are known without uncertainty and that corresponding clades are given as input.

(i) Methods looking for independent substitutions to the same amino acid

The most intuitive method, the 'identical' method, looks for independent substitutions to the exact same amino acid in all clades with the convergent phenotype [20,21]. It therefore assumes that a particular amino acid has a much better fitness than all other amino acids at this particular position of a protein. In practice, it relies on ancestral sequence reconstruction to infer the amino acids present before each convergent transition and make sure that the transition of interest occurred on the branch where the phenotypic transition occurred. By design, it is very conservative because it aims to detect only sites where a single particular amino acid is much more fit than others, which fixed with a type 1 substitution (figure 2).

An extension of this method, the 'expectation' method of Chabrol *et al.* [19]—also called msd—looks for sites with a high 'convergence index'. This convergence index is the expected number of substitutions to a particular amino acid in convergent lineages. Interestingly, and contrary to the other methods presented here, this method does not assume that convergent lineages must be known. Instead, it is enough to have phenotypic annotations for extant species only. It is unclear whether this method is very conservative or not: on one hand, it detects only sites where a particular amino acid is found in most species with the convergent phenotype (as in the 'identical' method), but on the other hand, this convergence could apply to only a subset of the species with the convergent phenotype, an advantage compared to methods based on amino acid profiles (see below, §1c(iii)). Both type 1 and type 2 substitutions can be detected by this method, but type 2 substitutions get a higher convergence index than type 1 substitutions and may therefore be better detected.

(ii) Method based on topological incongruencies

The 'topological' method is an early attempt to look for an indirect effect of convergent sequence evolution, based on an observation first made on the prestin gene [22] and later systematized in genome-scale studies [1–3]. When a particular site has evolved convergently in several lineages, it displays the same or similar amino acids in those lineages, and not in lineages with a different phenotype. As a result, for this site, a phylogeny in which convergent lineages are grouped together is more likely than the true species phylogeny. This approach involves constructing the species

topology and a 'convergent' topology where species with the convergent phenotype are grouped together. Then, each site can be tested for which topology it prefers—the true species phylogeny or the convergent phylogeny—by comparing the likelihoods of the two trees for this site. This method is capturing a byproduct of convergent evolution, and not its mechanism, hence it is difficult to know precisely what type of substitution this method can work with. Presumably, both type 1 and type 2 substitutions can be detected.

(iii) Methods based on amino acid profiles

'Profile methods' are methods aiming to detect selection pressure changes, whereby different amino acids provide the highest fitnesses in the ancestral and convergent phenotypes. The simplest of these methods is the 'multinomial' approach, which performs a simple χ^2 test for multinomial distributions [23] between two vectors of amino acid frequencies. One vector is based on the amino acids found in extant species with the ancestral phenotype, and the other vector is based on the amino acids found in extant species with the convergent phenotype. This approach has not previously been used in the literature to our knowledge and suffers from a major drawback in that it fails to account for the phylogenetic structure of the data. However, we chose to include it in our tests as it provides a baseline against which the other more sophisticated methods can be tested. Both type 1 and type 2 substitutions can be detected by this method.

Other profile methods include profile change with one change (PCOC) [18], diffsel [24] and TDG09 [25], which belong to a family that we loosely call 'mechanistic methods' because they combine a phylogenetic approach with amino acid fitness profiles.

The 'PCOC' method [18] models convergent evolution at the amino acid level, without taking into account the codon level. It combines the 'profile' idea—by attributing to the 20 amino acids different equilibrium frequencies before and after the phenotypic changes—with the One Change (OC) model. OC assumes that convergent sites must have undergone a substitution on the branches where the adaptation took place. Detection of convergent sites is obtained by comparing the likelihoods of two nested models. In the first model, both the profile change and OC models are used—this means that profiles change on branches where the phenotype changes and that at least one substitution must occur on each of these branches. In the second model, evolution is homogeneous across all branches. Amino acid profiles are not estimated but are drawn from pre-existing distributions that have been estimated on large collections of alignments [26]. Both type 1 and type 2 substitutions can be detected by PCOC, but with a different power: the OC component of PCOC expects only type 1 substitutions, but the PC component can accommodate both type 1 and type 2 substitutions.

The TDG09 model [25] is similar to PCOC in that it works at the amino acid level, but it focuses on profile changes and does not include the OC component. In addition, it estimates the profiles separately for each site of the alignment. To do so, it builds two profiles, one for the species with the ancestral phenotype, and one for the species with the convergent phenotype. Amino acids with a count of 1 or less are considered absent, and all absent amino acids are assigned a 0.0 frequency in the profile vector. To detect sites undergoing adaptive convergent evolution, a likelihood ratio test is performed between a model that assumes a single profile across the entire tree, or two profiles for the ancestral and convergent parts of the tree. Both type 1 and type 2 substitutions can be detected by this method.

Finally, diffsel [24] is similar in spirit to TDG09 but works at the codon level and uses an MCMC algorithm to perform inference in the Bayesian framework. In this codon model, mutations occur at the DNA level, and selection occurs at the amino acid level. Selection is modelled as site-wise fitness profiles of 20 amino acid fitnesses. Convergent sites are characterized by a systematic change from an ancestral amino acid fitness profile to a different amino acid profile on all branches where the phenotype changed. Both type 1 and type 2 substitutions can be detected by this method.

2. Results and discussion

Some of the methods presented above have been implemented in several software packages (table 1). In this article, we evaluate these software packages on simulated and empirical data along with methods we have reimplemented ourselves. Regarding empirical data, we focus on sites that had been identified as convergent in previous publications and look at how the methods rank those sites. Regarding simulations, we evaluate the power of the methods in three cases: (1) a convergent profile change; (2) a convergent increase or decrease in selection efficacy; and (3) a combination of the above two, whereby a convergent profile change occurs simultaneously with a scaling of selection efficacy. To achieve this scaling, we set a selection efficacy parameter that is the product of two parameters, the population size (Ne) and the selective pressure (S) (also called scaled selection coefficient). In the following, we refer to this value by NeS, a composite parameter whose variations can be interpreted as e.g. a genome-wide variation of population size, or a site-wise variation of selective pressure. We choose to use $\mathrm{NeS} = 4$ as the reference value, because it produces alignments similar to empirical alignments according to a range of statistics (electronic supplementary material, figures S4–S7).

We ran the methods on four empirical phylogenies with different size, depth and number of transitions [20,28,29] (electronic supplementary material, figure S2).

In case 1, 'Convergent profile change', selection efficacy remains constant but the amino acid profile is different in convergent lineages compared to the rest of the tree. To simulate this case (figure 3*a* and figure 4*a*), we change the amino acid profile in the convergent clades and we keep the same global NeS along the tree. The results are presented in figure 3 for $\mathrm{NeS} = 4$, and for the four empirical phylogenies.

Profile methods perform better than the other methods in the four phylogenies, and among them, diffsel dominates the benchmark according to AUC values (figure 3). The sensitivity at 90% precision is not as easily interpretable as AUC because the curves are very rugged; TDG09, PCOC and diffsel seem to dominate this metric, with a different order depending on the tree. Surprisingly, the simplistic multinomial method performs well on the Cyperaceae tree, competing with the TDG09 and PCOC in terms of its sensitivity at 90% precision. The relative ranks of PCOC, multinomial and TDG09 vary depending on the tree, which may be attributable to differences in the number of

Table 1. Summary of the methods used in the pipeline.

name	original method publication	level	executable or source available	median computing time on 2000 sites
identical	[20]	site	no, reimplemented (Python and C++)	55 s
topological	[1,27]	site	no, reimplemented (Python and C++)	5 s
TDG09	[19,25]	site	yes. Used a modified version. https://github.com/tamuri/tdg09	1648 s (27 min)
diffsel	[24]	site	yes: https://github.com/vlanore/diffsel	141084 s (39 h)
PCOC	[18]	site	yes: https://github.com/CarineRey/pcoc	181 s (3 min)
multinomial	—	site	no, implemented de novo (Python)	21 s
msd	[19]	gene	yes. Used a modified version. https://github.com/gilles-didier/Convergence	70 s

convergent transitions and in the relative size of the convergent clades. For instance, we suspect that PCOC's performance is degraded when the number of convergent transitions increases, because by design it looks for sites with convergent changes in all the convergent clades, not just a subset of them. TDG09 shows the opposite trend, with better performance when the number of transitions increases. The topological, identical and msd approaches typically perform worse, but the AUC rank of msd is volatile. The low sensitivity of identical and msd is expected as those methods can only detect convergent substitutions to a particular amino acid, not to an amino acid profile. Overall, these results are qualitatively congruent with previously published simulations obtained with simpler settings and fewer methods [18]. However, the precisions and sensitivities observed here are much worse than those reported in [18], because simulations do not use the PCOC model, which enforces substitutions on all transition branches.

Note that diffsel, which performs well in our experiments, is also the most expensive method computationally by several orders of magnitude (table 1). Other methods may be preferable for large datasets unless extensive computing resources are available. The better performance of profile methods may be owing to their fitting the simulation conditions better. However, it is unclear how we could have simulated convergent evolution realistically without using mutation-selection models that use profiles of amino acid frequencies. In the end, this indicates that profile methods may perform better on empirical data as well; apart from diffsel, which always comes out first, the variability of the AUC ranks among trees, however, indicates that using several methods on a dataset is recommended.

We then studied the performance of the methods for a wider range of genome-wide selection efficacies, focusing on the Cyperaceae tree (see electronic supplementary material, figure S8 for the three other trees). Figure 4*a* represents AUC values for the Cyperaceae tree, for NeS = 1, 4 and 8, corresponding to values for weak, medium and high selection efficacy respectively, all of which produce alignments with realistic properties (electronic supplementary material, figures S4–S7). As expected, the methods are most accurate when NeS is high (NeS = 8) and the performance collapses when selection is not efficient (NeS = 1). In other words, it should be extremely difficult to detect convergent molecular evolution in species with small Ne, or for sites under weak selective pressure.

In case 2, 'Convergent scaling of selection efficacy' (figure 4*b*), the same amino acid fitness profile is used along the whole tree for a given site, but NeS is changed in convergent clades (from $\mathrm{NeS_A}$ to $\mathrm{NeS_C}$) in Ha simulations. It is important to note that an NeS variation implies the modification of the fitness of each amino acid in the profile but not of its rank (figure 1*a*). We made 3 runs, two with an increase and another with a decrease of NeS in convergent clades. Overall, methods perform poorly at detecting selection efficacy scaling, with the exception of the $\mathrm{NeS_A} = 1$ to $\mathrm{NeS_C} = 4$ cases where PCOC and diffsel detect a small number of sites.

By the two previous cases, we saw that methods can detect adaptive convergent sites under two conditions: they have undergone a profile change and they are under moderate to high selective pressure. But the methods cannot detect profile changes when selection efficacy is low and also fail to detect scalings in selection efficacy alone.

Finally, case 3 introduces a confounding factor. Here we assume a genome-wide scaling of selection efficacy on top of which convergent sites undergo profile changes (figure 4*c*), and we try to detect those latter sites. This is modelled by a selection efficacy scaling from $\mathrm{NeS_G}$ to $\mathrm{NeS_C}$ in both convergent (Ha) and non-convergent (H0) sites, plus an amino acid profile change in Ha. We tried both to decrease (figure 4*c*(i)) or increase (figure 4*c*(ii)) the selection efficacy in the convergent clades and compared the results to the situation obtained when selection efficacy is constant. With a decreased selection efficacy in convergent clades, the methods' performances deteriorate compared to the reference simulation. With an increased selection efficacy in convergent clades, the performances remain roughly the same. In other words, a decrease in selection efficacy (for instance, owing to a decrease in Ne) coinciding with convergent transitions has a negative impact on the detection of convergent profile changes, but an increase has very little impact.

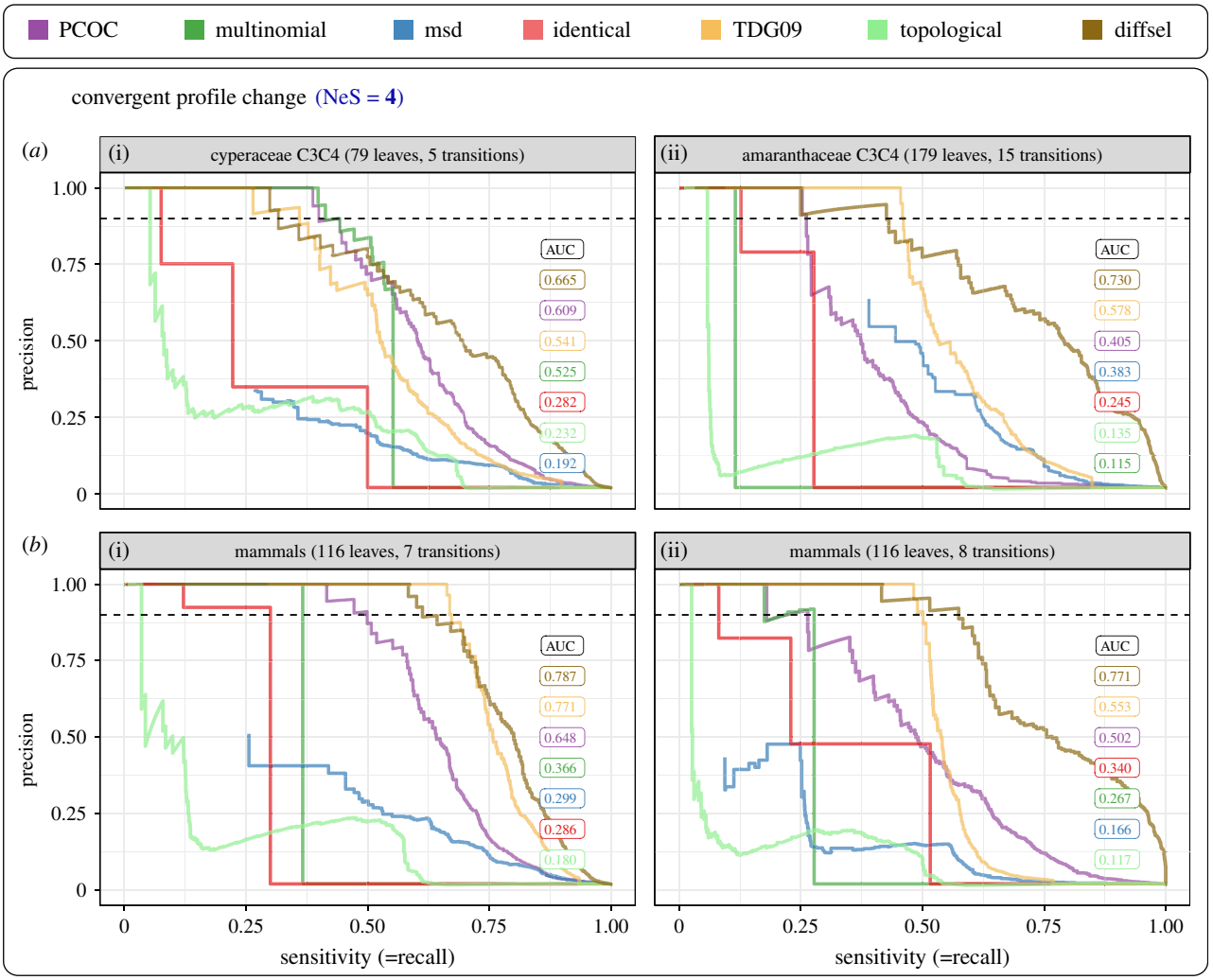

Figure 3. Detection of sites undergoing convergent profile change by different methods. Simulations are performed with constant selection efficacy (NeS = 4). Each panel corresponds to one empirical phylogeny, with convergent transitions placed as in electronic supplementary material, figure S2. The trade-off between sensitivity and precision is presented for each method, assuming that 2% of the sites are convergent in the sequences (colour code indicated on the top of the figure). The dashed lines highlight 90% precision. Area under the curves (AUC) ranked from best to worst are presented on the right-hand sides of each panel, with the same colour code as the precision-recall curves.

We then ran the seven methods on two previously published datasets, where a list of convergent sites had been proposed [20,24,28]. In these articles, the detection of convergent sites was performed by a version of the identical method for Cyperaceae, and by either diffsel or a dN/dS analysis for Amaranthaceae. Note that in the original diffsel article, the method was run with slightly different settings: it evaluated fitness profiles separately for the sister clades of convergent clades. In this article, diffsel was instead set up to only evaluate one profile per site for convergent clades and one profile per site for the rest of the tree. This change was done to make diffsel results more comparable with other methods and explains the differences with the original results. We compared the ranks of these previously reported convergent sites across methods. The alignments and the ranks are available in electronic supplementary material, figures S10 and S11, along with further discussion. Overall, the methods tend to agree with each other and rank the previously reported convergent sites among their best candidates (figure 5). In particular, most profile methods are in strong agreement with the publications; this is especially true for TDG09. Some methods fail to find any or nearly any convergent evolution on the Amaranthaceae alignment (identical, multinomial, topological), which is consistent with our results on data simulated on the same tree (figure 3*a*(ii)). These methods have a more consistent behaviour on the Cyperaceae dataset. For both datasets, most of the sites that have low ranks have low ranks across methods. For instance, this is the case for sites 733 and 770 in Cyperaceae, and 143 and 439 in Amaranthaceae. Overall, methods that produce the best results on simulated data also recover convergent sites identified in previous studies. Those sites had been identified with either diffsel or identical, so it is not surprising that these methods performed well in our study on the alignment on which they had been used; nonetheless, the general agreement between the methods is reassuring.

3. Material and methods

(a) Simulation of alignments of coding sequences

We simulated coding sequences using bppseqgen [30] under (heterogeneous) mutation-selection models, which belong to the 'mechanistic' family of methods tested in this work. Mutation-selection models are codon models that combine mutations at the DNA level with amino acid fitness vectors, so

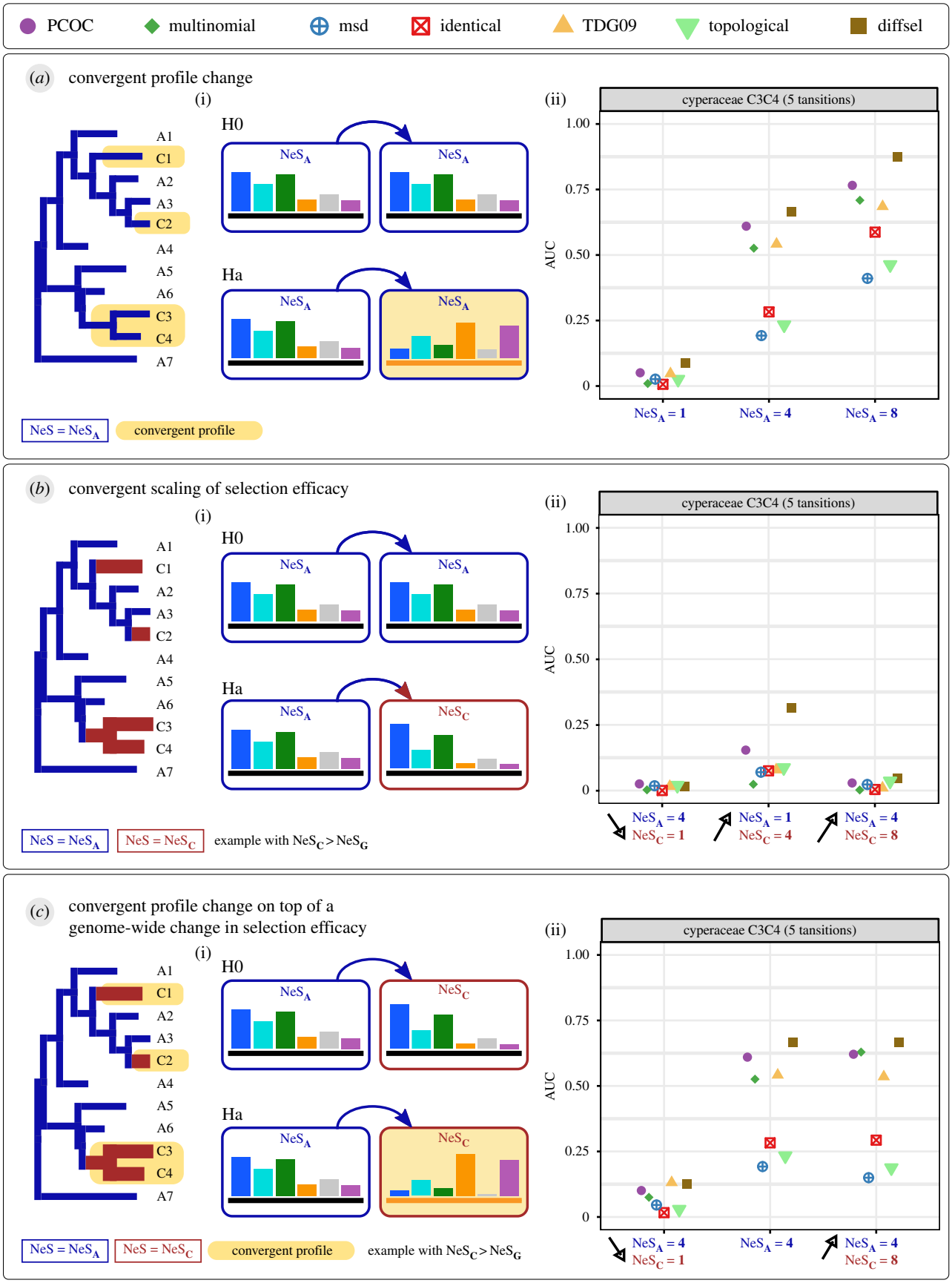

Figure 4. Overview of the simulations and AUC values for the Cyperaceae tree. The trees, convergent clades and symbols are as in figure 1. Three kinds of adaptive convergent cases have been simulated: (*a*) a convergent profile change, (*b*) a convergent scaling of selection efficacy and (*c*) a convergent profile change combined with a selection efficacy scaling. The genome-wide selection efficacy (NeS_A) remains the same in (*a*) and is changed to a convergent selection efficacy (NeS_C) in Ha (*b*) and Ha (*c*). The black arrows (*b*(ii) and *c*(ii)) indicate if selection efficacy increases or decreases in convergent clades. AUC values are calculated based on precision-recall curves such as presented in figure 3 ((*a*) AUC values for $NeS_A = 4$ in case 1 correspond to figure 3*a*(i)).

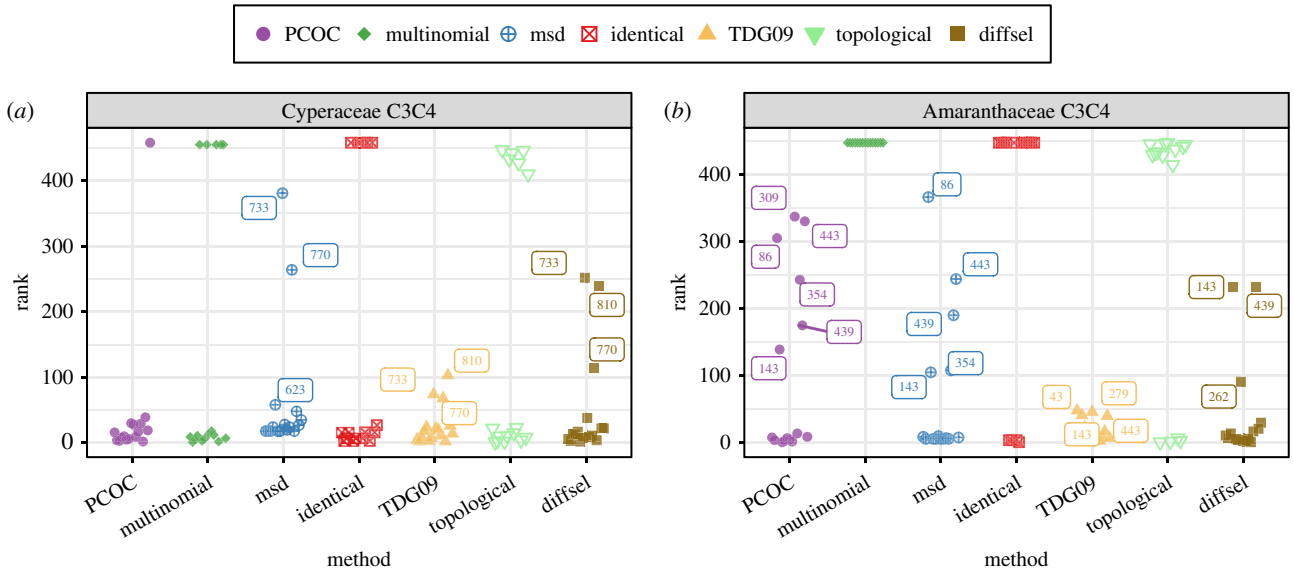

Figure 5. Ability of the different methods to recover published convergent sites in two empirical alignments. Those alignments had been used to study convergent transitions from C3 to C4 metabolisms in plants ((*a*) Besnard *et al.* [20] found 16 convergent sites in Cyperaceae, (*b*) Parto & Lartillot [24] found 15 convergent sites in Amaranthaceae). For each method, the scores were obtained for each site of the alignment. The sites were then ranked according to their scores, and only the ranks of previously published convergent sites are reported on the figure.

that selection operates only at the amino acid level. Our mutation-selection models were complemented by a parameter indicating the efficacy of selection, NeS. In our mutation-selection model, NeS controls the flatness of the amino acid profiles (electronic supplementary material, §2). With a high NeS, the profiles are very peaked, and with a low NeS, very flat. We investigated the impact of different NeS values, in homogeneous models, where the same NeS is applied to all the branches (figure 4*a*), and in heterogeneous models, where different NeS are used for the branches in the ancestral and convergent parts of the tree.

We performed several types of simulations. Simulation settings are described in the results section (figure 4). For each simulated codon position, one or two profiles are selected randomly in our set of 263 non-redundant profiles and one or two NeS values are chosen. One profile and one NeS value are used for the ancestral branches, and the others for convergent subtrees.

(b) Methods to detect foreground adaptive convergent substitutions

In order to compare results across methods, it was necessary to standardize their output. See the electronic supplementary material for details.

(c) Pipeline and implementation of the methods

The results in this article were obtained using an all-in-one pipeline that encompasses simulations, detection and post-simulation analysis, including the generation of the plots used for figures 3 and 4. The pipeline itself was implemented in OCaml using bistro (https://github.com/pveber/bistro), a library to build statically typed reproducible workflows. Special attention was paid to reproducibility, in particular, by following the guidelines given in [31]. Instructions to reproduce our results are given in the electronic supplementary material.

The implementations of the methods used in the pipeline are as follows:

— The multinomial method has been implemented *de novo* in Python as well as the identical and topological methods which additionally use executables from the bppsuite [30]. They are available via the pipeline.
— The TDG09 implementation we used is a slightly modified version of the one available on github (table 1) where multithreading has been removed to avoid multithreading-related problems. Results should be identical to the github version. In addition, a script available in the pipeline repository was written to adapt input alignments and trees to TDG09 expected formats.
— For diffsel, we used an optimized version of the original implementation that is faster but implements the same model. The implementation we used is available on github (table 1). In addition, we use a different approach to establish MCMC convergence. The original method compares two MCMC chains using the tracecomp program from the PhyloBayes suite [32]. Instead, we run only one chain, use the Raftery and Lewis's Diagnostic implemented in the R package coda (v0.19-1) [33] after 200 iterations to estimate the number of necessary iterations, then run as many iterations as 120% of the estimated number and finally perform the same diagnostic to check convergence.
— We used the github version of PCOC (table 1) as is.
— Regarding msd, we used a version modified by the author so as to output a *p*-value for all sites, which we needed to compute scores.

The experiments performed for this article—i.e. the whole pipeline with 2000 sites for each hypothesis times 12 hypotheses times four trees—took 5 days to run on a 24-core virtual machine. Computation times observed during this run for individual detection methods are given in table 1. Note that most of the computing time for the whole pipeline is spent in diffsel tasks, which are a lot more costly to compute than other methods.

(d) Using the methods on real alignments

We ran the methods on two previously published alignments: the Amaranthaceae alignment (447 sites, 15 published convergent sites) [24,28] and the Cyperaceae alignment (458 sites, 16 published convergent sites) [20]. The sites displayed in figure 5 are the sites proposed as possibly convergent in the original publications. Scores were obtained for each method and the sites were ranked (tied elements get the highest rank).

4. Conclusion

Our simulation results reveal the performance of existing methods to detect two different types of convergent amino acid evolution on simulated data, in isolation or combined with each other. The simulations have been performed with complex models of sequence evolution, parametrized so as to generate datasets that resemble empirical data on a few test statistics. However, some key assumptions underlying those models are clearly unrealistic: first, each site is simulated independently of the others. It would be useful to incorporate epistatic constraints in our simulations as those increase the number of background convergent substitutions [16]. Such a model has been proposed [16,17], but the current implementation can only work one branch at a time, not along a tree topology.

Second, although it is an important part of the model, the phenotype is here considered in an extremely naive fashion. In particular, we have made no effort to incorporate a distribution of fitness effects, whereby different sites would contribute differently to the phenotype under consideration, and therefore to the fitness [34]. Using such a distribution would be key to understanding why some sites, those of large effect, undergo convergent evolution while others, with smaller effects, do not. It could also indicate to users what effect sizes are large enough to be detected in a given experimental setting, and what effect sizes are just too small to be detected.

Third, several known confounding factors have not been simulated. In particular, we have not incorporated bGC in our simulations, and we have not incorporated population-level processes that would allow polymorphisms to cross speciation events (incomplete lineage sorting, ILS) and would increase the levels of polymorphisms present at the tips of the trees. We have not investigated several factors that are likely to affect the ability of the methods to detect convergent amino acid evolution such as tree size, tree shape and branch lengths (but see [18]). Our simulation pipeline can, however, be used to study such parameters.

With these caveats in mind, our simulations show that existing methods are much better at detecting convergent profile changes rather than convergent selection efficacy rescalings. Further, detection of convergent profile changes is improved when selection efficacy is high, possibly because this increases the frequency of type 1 substitutions. They also show that model-based methods, which explicitly rely on profiles, perform better than other methods.

Moving forward, we can think of three complementary directions for improving methods aiming to detect adaptive convergent evolution in amino acid sequences. In all cases, they will be based on profile methods anchored in a mechanistic modelling of sequence evolution. As a first direction, we need to complement models of sequence evolution so that, in addition to profile changes, we can also accurately detect changes in selection efficacy and distinguish those adaptive processes from confounding factors such as bGC and ILS. Further anchoring the model in population genetics theory may allow the interpreting of detected sites in terms of the fitness advantage they provide. As a second direction, we need to improve the computational efficiency of model-based inference. This should be a major concern here, because datasets are getting larger every year; algorithmic or mathematical developments will probably be necessary to fit such complex models onto large datasets. In this respect, one intriguing result of this study is the performance of the multinomial method. This simplistic method ignores nearly everything of the complexities of codon models of sequence evolution and yet achieves a performance that rivals them in some conditions. Correcting the multinomial method for phylogenetic inertia could provide even better performances, and it may be possible to improve it further while keeping its excellent speed. Finally, we have only tested the methods' ability to detect individual convergent sites; some methods (e.g. msd) can also employ a statistical procedure to detect convergent genes by combining site-wise evidence. Alternatively, TDG09 has a procedure to control its false positive rate, and diffsel estimates parameters based on entire alignments, not single sites. None of those features have been tested but are crucial for application to real data, in particular, for application to genome-wide datasets. Future analyses will have to investigate these aspects.

Data accessibility. Our pipeline's code is available at https://gitlab.in2p3.fr/pveber/reviewphiltrans. It contains everything required to reproduce our results. Detailed reproduction instructions are given in electronic supplementary material, §9. All intermediate data used to produce our results (approx. 20 Go) are available from the Dryad Digital Repository: https://doi.org/10.5061/dryad.57hr00q/1 [35].

Authors' contributions. C.R., V.L. and P.V. built the pipeline for testing the methods. C.R., V.L., L.G., N.L. and B.B. implemented some of the methods. N.L. estimated empirical profiles. C.R., M.S. and B.B. performed statistical analyses. C.R., V.L., M.S. and B.B. wrote the article. All authors read, commented on and approved the article for publication.

Competing interests. We have no competing interests.

Funding. The research presented here was funded by the Convergenomix project (ANR-15-CE32-0005). C.R. was supported by a PhD fellowship (CDSN) from the Ecole Normale Supérieure of Lyon.

Acknowledgements. We would like to thank G. Didier for providing a modified version of msd that can work on individual sites and T. Latrille for providing the profiles and fruitful discussions. This work was performed using the computing facilities of the CC LBBE/PRABI. We would like to thank the French Institute of Bioinformatics—IFB CNRS UMS 3601—(funded as part of Investissement d'avenir program managed by Agence Nationale pour la Recherche, contract ANR-11-INBS-0013) for providing life science data and tools, storage and computing resources on the IFB national service infrastructure in bioinformatics.

References

1. Parker J, Tsagkogeorga G, Cotton JA, Liu Y, Provero P, Stupka E, Rossiter SJ. 2013 Genome-wide signatures of convergent evolution in echolocating mammals. *Nature* **502**, 228–231. (doi:10.1038/nature12511)
2. Zou Z, Zhang J. 2015 No genome-wide protein sequence convergence for echolocation. *Mol. Biol. Evol.* **32**, 1237–1241. (doi:10.1093/molbev/msv014)
3. Thomas GWC, Hahn MW. 2015 Determining the null model for detecting adaptive convergence from genomic data: a case study using echolocating mammals. *Mol. Biol. Evol.* **32**, 1232–1236. (doi:10.1093/molbev/msv013)
4. Khan AI, Dinh DM, Schneider D, Lenski RE, Cooper TF. 2011 Negative epistasis between beneficial mutations

in an evolving bacterial population. *Science* **332**, 1193–1196. (doi:10.1126/science.1203801)

5. Goldman N, Yang Z. 1994 A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Mol. Biol. Evol.* **11**, 725–736. (doi:10.1093/oxfordjournals.molbev.a040153)
6. Muse SV, Gaut BS. 1994 A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. *Mol. Biol. Evol.* **11**, 715–724. (doi:10.1093/oxfordjournals.molbev.a040152)
7. Yang Z, Nielsen R. 2008 Mutation-selection models of codon substitution and their use to estimate selective strengths on codon usage. *Mol. Biol. Evol.* **25**, 568–579. (doi:10.1093/molbev/msm284)
8. Halpern AL, Bruno WJ. 1998 Evolutionary distances for protein-coding sequences: modeling site-specific residue frequencies. *Mol. Biol. Evol.* **15**, 910–917. (doi:10.1093/oxfordjournals.molbev.a025995)
9. Rodrigue N, Philippe H, Lartillot N. 2010 Mutation-selection models of coding sequence evolution with site-heterogeneous amino acid fitness profiles. *Proc. Natl Acad. Sci. USA* **107**, 4629–4634. (doi:10.1073/pnas.0910915107)
10. Pessia E, Popa A, Mousset S, Rezvoy C, Duret L, Marais GAB. 2012 Evidence for widespread GC-biased gene conversion in eukaryotes. *Genome Biol. Evol.* **4**, 675–682. (doi:10.1093/gbe/evs052)
11. Lassalle F, Périan S, Bataillon T, Nesme X, Duret L, Daubin V. 2015 GC-Content evolution in bacterial genomes: the biased gene conversion hypothesis expands. *PLoS Genet.* **11**, e1004941. (doi:10.1371/journal.pgen.1004941)
12. Jarvis ED *et al*. 2014 Whole-genome analyses resolve early branches in the tree of life of modern birds. *Science* **346**, 1320–1331. (doi:10.1126/science.1253451)
13. Weber CC, Boussau B, Romiguier J, Jarvis ED, Ellegren H. 2014 Evidence for GC-biased gene conversion as a driver of between-lineage differences in avian base composition. *Genome Biol.* **15**, 549. (doi:10.1186/s13059-014-0549-1)
14. Zhu X *et al*. 2018 Divergent and parallel routes of biochemical adaptation in high-altitude passerine birds from the Qinghai-Tibet Plateau. *Proc. Natl Acad. Sci. USA* **115**, 1865–1870. (doi:10.1073/pnas.1720487115)
15. Lefébure T *et al*. 2017 Less effective selection leads to larger genomes. *Genome Res.* **27**, 1016–1028. (doi:10.1101/gr.212589.116)
16. Goldstein RA, Pollard ST, Shah SD, Pollock DD. 2015 Nonadaptive amino acid convergence rates decrease over time. *Mol. Biol. Evol.* **32**, 1373–1381. (doi:10.1093/molbev/msv041)
17. Pollock DD, Thiltgen G, Goldstein RA. 2012 Amino acid coevolution induces an evolutionary Stokes shift. *Proc. Natl Acad. Sci. USA* **109**, E1352–E1359. (doi:10.1073/pnas.1120084109)
18. Rey C, Guéguen L, Sémon M, Boussau B. 2018 Accurate detection of convergent amino-acid evolution with PCOC. *Mol. Biol. Evol.* **35**, 2296–2306. (doi:10.1093/molbev/msy114)
19. Chabrol O, Royer-Carenzi M, Pontarotti P, Didier G. 2018 Detecting the molecular basis of phenotypic convergence. *Methods Ecol Evol.* **9**, 2170–2180. (doi:10.1111/2041-210x.13071)
20. Besnard G, Muasya AM, Russier F, Roalson EH, Salamin N, Christin P. 2009 Phylogenomics of C4 photosynthesis in sedges (Cyperaceae): multiple appearances and genetic convergence. *Mol. Biol. Evol.* **26**, 1909–1919. (doi:10.1093/molbev/msp103)
21. Zhang J, Kumar S. 1997 Detection of convergent and parallel evolution at the amino acid sequence level. *Mol. Biol. Evol.* **14**, 527–536. (doi:10.1093/oxfordjournals.molbev.a025789)
22. Li G, Wang J, Rossiter SJ, Jones G, Cotton JA, Zhang S. 2008 The hearing gene *Prestin* reunites echolocating bats. *Proc. Natl Acad. Sci. USA* **105**, 13 959–13 964. (doi:10.1073/pnas.0802097105)
23. Pearson KX. 1900 On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Lond. Edinb. Dublin Phil. Mag. J. Sci.* **50**, 157–175. (doi:10.1080/14786440009463897)
24. Parto S, Lartillot N. 2018 Correction: molecular adaptation in Rubisco: discriminating between convergent evolution and positive selection using mechanistic and classical codon models. *PLoS One.* **13**, e0196267. (doi:10.1371/journal.pone.0196267)
25. Tamuri AU, Dos Reis M, Hay AJ, Goldstein RA. 2009 Identifying changes in selective constraints: host shifts in influenza. *PLoS Comput. Biol.* **5**, e1000564. (doi:10.1371/journal.pcbi.1000564)
26. Le SQ, Lartillot N, Gascuel O. 2008 Phylogenetic mixture models for proteins. *Phil. Trans. R Soc. B* **363**, 3965–3976. (doi:10.1098/rstb.2008.0180)
27. Castoe TA *et al*. 2009 Evidence for an ancient adaptive episode of convergent molecular evolution. *Proc. Natl Acad. Sci. USA* **106**, 8986–8991. (doi:10.1073/pnas.0900233106)
28. Kapralov MV, Smith JAC, Filatov DA. 2012 Rubisco evolution in C4 Eudicots: an analysis of Amaranthaceae Sensu Lato. *PLoS ONE* **7**, e52974. (doi:10.1371/journal.pone.0052974)
29. Douzery EJP, Scornavacca C, Romiguier J, Belkhir K, Galtier N, Delsuc F, Ranwez V. 2014 OrthoMaM v8: a database of orthologous exons and coding sequences for comparative genomics in mammals. *Mol. Biol. Evol.* **31**, 1923–1928. (doi:10.1093/molbev/msu132)
30. Guéguen L *et al*. 2013 Bio++: efficient extensible libraries and tools for computational molecular evolution. *Mol. Biol. Evol.* **30**, 1745–1750. (doi:10.1093/molbev/mst097)
31. Sandve GK, Nekrutenko A, Taylor J, Hovig E. 2013 Ten simple rules for reproducible computational research. *PLoS Comput. Biol.* **9**, e1003285. (doi:10.1371/journal.pcbi.1003285)
32. Lartillot N, Lepage T, Blanquart S. 2009 PhyloBayes 3: a Bayesian software package for phylogenetic reconstruction and molecular dating. *Bioinformatics* **25**, 2286–2288. (doi:10.1093/bioinformatics/btp368)
33. Plummer M, Best N, Cowles K, Vines K. 2006 CODA: convergence diagnosis and output analysis for MCMC. *R News* **6**, 7–11.
34. Wilson DJ, Hernandez RD, Andolfatto P, Przeworski M. 2011 A population genetics-phylogenetics approach to inferring natural selection in coding sequences. *PLoS Genet.* **7**, e1002395. (doi:10.1371/journal.pgen.1002395)
35. Rey C, Lanore V, Veber P, Guéguen L, Lartillot N, Sémon M, Boussau B. 2019 Data from: Detecting adaptive convergent amino acid evolution. Dryad Digital Repository. (https://doi.org/10.5061/dryad.57hr00q/1)

3.3 Conclusions

Dans cette partie de ma thèse, je me suis intéressée à la détection de la convergence d'un point de vue théorique et notamment en comparant les performances des méthodes actuelles à partir de simulations de convergences.

3.3.1 Aucune méthode de détection de la convergence actuelle ne se démarque complètement

Parmi l'ensemble des méthodes actuelles, les méthodes dites à profils d'acides aminés, Tdg09, PCOC et diffsel se démarquent par rapport aux méthodes initiales (topologique et identique). Cependant, ces trois méthodes ont chacune des défauts différents qui les empêchent de devenir la référence en termes de détection de la convergence. Diffsel a les performances les plus élevées. Cependant, c'est aussi l'outil ayant le temps d'exécution le plus long. Il semble donc impensable de l'utiliser à l'échelle génomique sur des jeux de données pouvant contenir plus de 10 000 gènes. Par contre, il pourrait être utilisé sur des gènes candidats ciblés par une analyse préliminaire réalisée avec PCOC et/ou Tdg09.

Tdg09 est un outil dont le fonctionnement ressemble à PCOC. Pour chacun des sites, Tdg09 teste l'adéquation entre un modèle convergent hétérogène contenant deux profils d'acides aminés et un modèle non-convergent homogène avec un unique profil d'acides aminés. Cependant, il y a deux différences majeures. Premièrement, la composition des profils d'acides aminés est calculée *de novo* pour chacun des sites en utilisant la composition des séquences de ce site aux feuilles de l'arbre. Deuxièmement, la méthode n'exige pas une substitution au niveau de chacune des transitions convergentes. Ce dernier point est à la fois un atout et une faiblesse. En effet, dans le cas de jeux de données avec un grand nombre de transitions, un site pourra être déclaré positif s'il y a un changement de composition en acides aminés dans, par exemple, la moitié des transitions. Cependant, si le nombre de transitions est faible, il se peut que par hasard il y ait eu des changements dans la moitié des transitions indépendamment du phénotype convergent et que le site soit déclaré convergent. Ceci explique que Tdg09 soit très sensible quand le nombre de transitions convergentes augmente mais d'autre part qu'il soit moins spécifique quand le nombre de transitions convergentes diminue.

PCOC est un outil très spécifique mais dont la sensibilité diminue quand le nombre de transitions augmente. Ce comportement est lié à la composante OC (One Change) du modèle PCOC qui en fait sa particularité et sa force par rapport aux autres méthodes de détection. En effet, cela le rend plus stringent qu'un outil comme Tdg09 car il sera capable d'identifier uniquement les sites dont le signal convergent présent dans les séquences est très corrélé au phénotype convergent. Ce critère peut être considéré comme très (voire trop) stringent quand on considère de nombreuses transitions, mais c'est un atout lorsque l'on étudie un jeu de données avec peu de transitions.

Tout cela suggère qu'il n'existe pas encore une méthode complètement efficace pour détecter de la convergence mais plutôt trois qui doivent être utilisées de manière complémentaire. Par exemple, Tdg09 et PCOC pourraient être utilisés en premier pour identifier des sites potentiels, puis on pourrait utiliser diffsel sur les gènes contenant ces sites. Si ces sites ont également un score élevé avec diffsel, ce sont de très bons sites candidats qui pourraient être finalement testés fonctionnellement.

3.3.2 Les résultats des méthodes sont cohérents sur des gènes candidats

Pour tester la performance des méthodes de détection de la convergence, nous avons utilisé des simulations car il n'existe pas de jeu de données réels où nous connaissons les sites convergents à l'échelle du génome. Cependant, il existe des gènes candidats très bien documentés où les sites

convergents sont connus.

Les résultats des méthodes de détection sur deux de ces gènes (Figure 5, second article) sont cohérents avec la littérature et également entre eux. Cela permet de valider que ces méthodes fonctionnent également sur des données réelles et pas uniquement sur des données simulées.

3.3.3 Des facteurs confondants peuvent induire de la convergence ou la masquer

Le second article montre que des augmentations de tailles efficaces dans les espèces convergentes peuvent induire de la convergence, et réciproquement, que des diminutions de tailles efficaces peuvent réduire la capacité de détection des méthodes en diminuant l'efficacité de la sélection (Figure 4.c, second article).

Vincent Lanore, Philippe Veber et Bastien Boussau ont poursuivi cette étude des facteurs confondants possibles en s'intéressant à la conversion génique biaisée (gBGC) [1]. Ils trouvent également que ce processus peut induire de la convergence génomique.

Cela signifie que des facteurs confondants peuvent réduire fortement notre capacité à détecter de la convergence mais aussi mimer de la convergence. L'idéal serait de pouvoir les prendre en compte dans nos analyses. Cependant, nous n'avons pas encore les modèles pour le faire explicitement. Il est toutefois possible d'estimer les tailles efficaces de population le long de l'arbre (EDWARDS et collab., 2007; HELED et DRUMMOND, 2009) pour prendre en compte l'influence des changements de taille de population sur la convergence génomique, ou bien d'estimer la force de la bGC le long de l'arbre (PESSIA et collab., 2012). Mais tout cela ajouterait du temps de calcul dans nos méthodes qui sont déjà coûteuses.

3.4 Perspectives

Ces deux articles ont permis d'étudier le comportement des outils de détection de la convergence génomique sur des données simulées. La suite logique de ces travaux est le passage aux données réelles. Cependant cela pose de nouvelles questions.

3.4.1 Comment définir un seuil pour chacune des méthodes ?

La figure 3 du second article montre que chacune des méthodes a des performances différentes d'un jeu de données à l'autre. Cela signifie qu'il n'existe pas un seuil universel à partir duquel un site peut être déclaré convergent dans un jeu de données réelles. Il faut fixer *de novo* les seuils de chaque méthode pour tout nouveau jeu de données.

Une manière de fixer le seuil d'une méthode est d'utiliser des simulations de sites convergents et non-convergents. Comme on s'attend à ce que le nombre de sites non-convergents soit bien supérieur à celui des sites convergents dans les données réelles, il faut définir le seuil de manière à contrôler la précision de la méthode définie comme le nombre de sites réellement simulés comme convergents parmi les sites détectés comme convergents. Ainsi, la définition du seuil prend en compte, à la fois la sensibilité et la spécificité de la méthode, et l'attendu sur le ratio entre sites non-convergents et convergents.

Pour que cette manière de fixer un seuil fonctionne, il faut être capable de simuler des sites convergents et non-convergents proches de la réalité. En effet, si on simule des sites non-convergents facilement identifiables comme non-convergents, la spécificité théorique de la méthode va être élevée à partir d'un seuil peu stringent. Or, si les données réelles contiennent des sites non-convergents plus difficilement identifiables comme tels, ces sites peuvent être détectés comme convergents.

1. Ce processus induit la fixation préférentielle de guanine (G) et de cytosine (C) lors de la recombinaison méiotique

Cela va augmenter le taux de sites détectés et les performances obtenues seront différentes de celles attendues mais nous n'aurons aucun moyen de le savoir. Il faut donc être plutôt trop stringent dans la fixation du seuil que pas assez.

3.4.2 Comment passer de la détection du site convergent à la détection du gène convergent ?

Après avoir été capable d'identifier un site comme convergent, on veut pouvoir être capable de définir un gène comme convergent. En effet, seule la méthode msd parmi l'ensemble des méthodes présentées prend en compte le passage du site au gène convergent. Cependant, msd utilise de nombreuses simulations pour estimer si le signal de convergence pour chacun des gènes est supérieur à ce qui est attendu par hasard. Ce processus est malheureusement très coûteux en temps de calcul et ne semble donc pas transférable aux autres méthodes.

Une possibilité serait de calculer s'il y a un enrichissement de sites convergents dans un gène par rapport aux autres ou peut-être également au niveau du réseau de gènes. Cependant, je ne pense pas que ce soit la solution idéale. En effet, la fonction d'un gène peut être modifiée par une unique substitution.

Une alternative pour contourner ce problème pourrait être de passer à un modèle hiérarchique travaillant au niveau du gène plutôt que du site. Dans ce modèle hierarchique, on pourrait intégrer l'ensemble des scores des sites pour chacun des gènes plutôt que de considérer chacun des sites de manière indépendante. Cela permettrait d'identifier les gènes avec un site avec un très fort signal de convergence ainsi que les gènes ayant plusieurs sites avec un signal de convergence moyen. On pourrait également prendre en compte le taux d'évolution des sites dans l'environnement proche des sites convergents. Par exemple, il serait intéressant de donner plus de poids à un site ou un petit nombre de sites qui ont un signal de convergence dans un environnement de gène très contraint c'est à dire avec peu de substitutions, par rapport à un ou plusieurs sites qui ont un signal de convergence dans un environnement de gène qui évolue rapidement avec de nombreuses substitutions. Cela permettrait à l'échelle du gène de corriger le taux d'évolution qui va nécessairement apporter du bruit.

3.4.3 Comment prendre en compte les facteurs confondants ?

D'après les résultats présentés dans le second article (Figure 4, second article), des facteurs confondants peuvent induire de la convergence génomique. On s'attend donc à ce que les résultats des méthodes soient un mélange de sites réellement liés à l'acquisition du phénotype convergent et d'autres à des facteurs confondants. Il se peut également que des facteurs confondants puissent faciliter l'apparition de mutations permettant l'acquisition du phénotype convergent. Cependant, nous n'en connaissons pas les proportions. Il faudra en complément des analyses sur la convergence, essayer de quantifier l'intensité de ces facteurs dans les espèces convergentes par rapport aux espèces non-convergentes afin d'envisager la présence d'un biais potentiel. Dans tous les cas, il faudra avoir conscience de la présence potentielle de facteurs confondants lors de l'interprétation des résultats.

En conclusion de ce chapitre, les analyses ont montré que sur des données simulées les outils de détection de convergence fonctionnaient mais le passage sur les données réelles apporte de nouveaux défis qu'il va falloir surmonter.

De plus, bien que nous ayons essayé de rendre nos simulations les plus proches de la réalité possible, nous savons qu'elles sont tout de même artificielles et que de nombreux facteurs auxquels nous n'avons pas pensé vont apporter de nouveaux problèmes aux méthodes étudiées.

4 Détection de la convergence dans un jeu de données réelles : des difficultés techniques mais des résultats prometteurs

Sommaire

4.1 Introduction

Jusqu'à présent, j'ai présenté dans le premier chapitre (Partie 2) une méthode pour créer des jeux de données à partir de données réelles et dans le second chapitre (Partie 3) une méthode pour détecter de la convergence. La suite logique est donc de joindre les deux approches et de chercher de la convergence dans un jeu de données réel dans un test grandeur nature.

Les tests sur cas réels que j'ai présentés se bornaient à des exemples connus comme la prestine. Il existe aussi des études de jeux de données génomiques chez les animaux écholochateurs (PARKER et collab., 2013), les mammifères marins (FOOTE et collab., 2015), etc. Cependant, le phénotype convergent étudié ne permet pas d'avoir plus que deux ou trois transitions vers le phénotype convergent dans l'échantillonnage des espèces. De plus, les méthodes employées pour détecter de la convergence avaient, dans nos mains, des performances inférieures à la méthode que nous avons développée.

C'est pour cela que dans l'équipe, nous nous sommes orientés vers l'étude de l'adaptation convergente à la vie en milieu aride chez les rongeurs. En effet, cette étude a l'avantage de présenter un plus grand nombre de transitions convergentes. Par ailleurs, on peut s'attendre à ce qu'une adaptation physiologique telle que l'adaptation à l'aridité soit liée à des modifications dans la séquence des gènes et aussi dans leur expression. De plus, les rongeurs incluent la souris, un modèle historique dont la physiologie est particulièrement bien connue, de nombreuses données génomiques, une bonne annotation des gènes et un grand nombre de données fonctionnelles apportant une connaissance fine de la fonction de beaucoup de gènes. Outre la souris, de plus en plus de génomes de rongeurs sont séquencés ou en cours de séquençage. Par ailleurs, ce choix a aussi été orienté par le fait que les rongeurs sont le modèle de l'équipe.

Nous avons collecté des données publiques de RNA-Seq de rein (un organe qui est particulièrement adapté chez les espèces vivant en milieu xérique), dans des dizaines d'espèces de rongeurs. L'utilisation du RNA-Seq, nous a permis d'étudier la convergence par deux approches différentes : l'une à partir des séquences codantes des gènes et l'autre à partir du niveau d'expression de ces gènes. Le contraste des résultats provenant de ces deux approches, que ce soient les différences ou les points communs, permettra d'avoir un nouveau regard sur la convergence.

Cette étude est le résultat d'un travail collaboratif avec Domitille Chalopin, Jérémy Ganofsky, Bastien Boussau, Sophie Pantalacci et Marie Sémon. Domitille et Marie ont réalisé les analyses des niveaux d'expression des gènes. Domitille a également apporté son expertise sur la physiologie du rein chez les rongeurs. Jérémy a participé à l'implémentation du pipeline permettant la construction du jeu de données. Marie, Sophie et Bastien ont orienté le projet en apportant des remarques et des propositions au cours du déroulement des analyses. Marie et Sophie sont à l'initiative du projet. J'ai, quant à moi, participé à l'implémentation du pipeline de construction du jeu de données et aux analyses de l'expression des gènes, réalisé l'annotation automatique du milieu de vie des espèces et l'analyse de la convergence dans les séquences des gènes. La rédaction de ce chapitre a été collaborative et plus particulièrement sur les parties où j'ai peu participé aux analyses.

Dans ce chapitre, je présente la construction de ce jeu de données et une analyse préliminaire de la convergence au niveau des séquences et de l'expression. Mon but est de montrer sur des données réelles comment les outils que j'ai développés peuvent être utilisés pour traiter une question biologique, quelles sont les difficultés qui se présentent et d'évoquer quelques perspectives de poursuite d'un tel travail. Tout au long du chapitre, j'insisterai sur les points méthodologiques qui me paraissent importants et sur les améliorations qui pourront être apportées (en italique).

4.1.1 Rapide introduction à l'adaptation à l'aridité chez les rongeurs

Les milieux xériques sont caractérisés par une faible quantité d'eau, quelle que soit la température. L'adaptation à la vie désertique se traduit par un ensemble de traits morphologiques et

physiologiques permettant donc d'économiser l'eau. De nombreux clades, incluant de nombreuses plantes et animaux comme les camélidés, les renards ou les rongeurs, ont développé des adaptations variées pour trouver et économiser de l'eau (JOHNSON et collab., 2016). Certaines espèces trouvent de l'eau dans leur nourriture : Psammomys, le rat des sables, se nourrit exclusivement de plantes succulentes [1] contenant 80% d'eau et en cas de besoin sont capables de fabriquer de l'eau métabolique à partir de lipides. D'autres animaux tels que les chameaux et les oryx peuvent également produire de l'eau métabolique. D'autres espèces, comme la gerbille mongole, se nourrissent de graines sèches contenant 10% d'eau (BANKIR et DE ROUFFIGNAC, 1985). Dans ce cas, il est crucial de ne pas perdre l'eau qui est ingérée.

Nous avons choisi d'étudier le rein, qui est l'organe assurant l'homéostasie hydrique. Il assure en grande partie le nettoyage du sang et l'élimination des déchets du métabolisme via l'urine, entraînant une perte d'eau, d'où son importance en milieu aride. Il a été montré que les espèces adaptées à la vie xérique ont une forte capacité à concentrer leur urine (exemple de la gerbille) (BANKIR et DE ROUFFIGNAC, 1985). La concentration finale dans l'urine dépend principalement des solutés et des échanges hydriques le long des tubules rénaux et de la anse de Henle. Ainsi, d'importantes modifications de structure du rein ont été observées de manière récurrente mais pas systématique chez les animaux adaptés au milieu xérique. L'augmentation de l'épaisseur de la médulla interne et son isolement grâce à un épithélium plus épais pour éviter la dispersion des solutés et favoriser la réabsorption au niveau des tubules est un exemple de modification. L'augmentation du nombre de boucles des tubules rénaux, la diminution du nombre de néphrons, le développement de cônes pelviens spécialisés, l'élongation des papilles (BANKIR et DE ROUFFIGNAC, 1985) ont également été observés chez plusieurs espèces.

Les rongeurs sont un très bon modèle pour s'intéresser à l'adaptation à la vie en milieu aride car ils sont présents dans presque tous les écosystèmes terrestres (https://biodiversitymapping.org/wordpress/index.php/mammals/ ; (FABRE et collab., 2012; NOWAK et WALKER, 1999; SCHENK et collab., 2013)), avec des adaptations indépendantes à la vie xérique. En effet, plus de 250 espèces vivent dans des déserts, ce qui représente environ 11% des rongeurs (243 espèces selon IUCN 2013, 289 espèces selon (ALHAJERI et STEPPAN, 2018)).

Cette variété de milieux, ainsi que la quantité croissante de données génomiques, font des rongeurs un excellent modèle pour l'étude de l'adaptation et la convergence moléculaire (DU et collab., 2015; MANCEAU et collab., 2010). Quelques analyses du transcriptome de rein entier chez des espèces de rongeurs mésiques ou xériques ont été publiées, à l'échelle phylogénétique d'une ou quelques espèces (GIORELLO et collab., 2018, 2014; MACMANES et EISEN, 2014; MARRA et collab., 2012, 2014; PRADERVAND et collab., 2010). Des études sur l'expression des gènes dans le rein sont également disponibles, avec des transcriptomes de rein en cellule unique (CAO et collab., 2018; PARK et collab., 2018) et des expériences de plasticité de l'expression suivant une déshydratation sévère (KIM et SHIN, 2016; KORDONOWY et collab., 2017; MACMANES, 2017). Ces travaux soulignent l'importance de certaines grandes familles géniques directement impliquées dans la physiologie du rein, notamment des aquaporines, des vasopressines, des angiotensines mais également d'autres familles de gènes telles que les collagènes, les gènes liés à l'apoptose ou les gènes liés à la matrice extracellulaire.

4.1.2 Objectif : croiser des analyses de séquences et d'expressions

L'objectif de cette étude était donc de chercher de la convergence à l'échelle génomique dans un jeu de données regroupant des dizaines d'espèces, entre des rongeurs adaptés ou non à la vie en milieu xérique. Les questions que nous nous sommes posées concernent la quantité de convergence présente entre ces espèces. Tout d'abord, est-ce que l'on observe de la convergence entre les espèces ? Ensuite, si l'on observe de la convergence, est-ce qu'elle concerne peu de gènes

1. plantes charnues capables de stocker de l'eau dans leurs feuilles

dans l'ensemble des espèces et/ou est-ce que l'on retrouve plus de convergence entre certaines espèces ? Nous n'avons pas vraiment d'attente sur les résultats car il n'existe pas vraiment d'étude comparable avec autant de transitions.

Nous avons voulu mener deux approches complémentaires en cherchant la convergence dans les séquences codantes mais aussi dans les niveaux d'expression des gènes exprimés dans le rein. Les deux types d'analyses pourraient donner des points de vue complémentaires sur l'adaptation à l'aridité. On peut imaginer en effet que certains gènes contribuent à l'adaptation à l'aridité par des changements d'expression, d'autres par des changements de séquence et d'autres enfin par les deux à la fois. Nous avons donc eu besoin de créer un jeu de données en deux parties, l'une constituée des séquences codantes des espèces étudiées et l'autre des niveaux d'expression de ces gènes. A noter que les données générées ne donnent pas accès aux séquences régulatrices des gènes, nous ne pourrons donc pas espérer trouver de substitutions causant des changements d'expression convergents.

Dans ce chapitre, je détaille la création et l'analyse de ce jeu de données composé de 32 espèces, dont 16 que nous considérons adaptées à la vie xérique. Ce jeu de données a été préparé à partir de données publiques uniquement. Une analyse ultérieure fera suite avec des données générées dans l'équipe et un jeu de données plus conséquent.

4.2 Construction et validation du jeu de données

4.2.1 Construction du jeu de données

La composition de ce jeu de données doit répondre à différentes contraintes. Le premier est de contenir un échantillonnage représentatif des rongeurs, contenant des espèces mésiques et xériques. Le second est d'avoir, pour chacune des espèces, accès à des données de RNA-seq pour, d'une part, obtenir les séquences codantes de leurs gènes et, d'autre part, le niveau d'expression des gènes dans le rein.

Une première recherche des banques de données publiques de RNA-Seq a mis en évidence qu'il y avait peu d'échantillons de reins de rongeurs. Nous avons donc choisi d'utiliser l'ensemble des échantillons disponibles. Nous avons mis en place un pipeline automatisé permettant de sélectionner les échantillons de reins de rongeurs (cf paragraphe ci-après) et enfin de construire les deux sous-jeux de données. Ce pipeline automatisé a permis d'être exhaustif dans la sélection des échantillons.

L'annotation des milieux de vie des espèces sélectionnées par le pipeline a été également un challenge lors de la création de ce jeu de données. Cette information est souvent disponible dans des bases de données sous la forme d'un texte plus ou moins long. Cependant, la catégorisation binaire entre mésique ou xérique peut être arbitraire pour les espèces proches de l'intermédiaire entre les deux conditions et nécessite une intervention humaine. Nous avons mis en place une annotation automatique des milieux de vie de ces espèces sur la base des conditions climatiques de leurs habitats ce qui a permis une annotation basée sur des variables quantitatives et entièrement automatisée.

L'implémentation du pipeline de création du jeu de données et celui de l'annotation des milieux de vie sont détaillés dans les deux prochains paragraphes.

4.2.2 Création des alignements de séquences et des niveaux d'expression par un pipeline automatisé

La construction du jeu de données a été réalisée par un pipeline en trois parties. Ces trois parties sont successives mais elles nécessitent deux interruptions pour des interventions de l'utilisateur. De

plus, les trois parties contiennent des ensembles cohérents qui pourraient être utilisés de manière indépendante.

Nous avons choisi de créer un pipeline permettant de construire notre jeu de données pour que notre analyse soit reproductible mais aussi pour rendre disponible une base de code pour d'autres analyses. Pour assurer la reproductibilité du code nous avons utilisé le gestionnaire de pipeline Nextflow (https://www.nextflow.io/) permettant la containerisation [1] des applications.

A noter : Je présente CAARS dans le premier chapitre (cf Partie 2), un outil qui précisément permet de construire des jeux de données multispécifiques à partir de données de RNAseq. Nous n'avons pas pu l'utiliser pour cette étude, car le jeu de données que nous voulions créer comprend 32 espèces dont 22 nouvelles ce qui est trop pour CAARS dans sa version actuelle, comme discuté dans le premier chapitre. De plus, la création de ce jeu de données a été initiée dans le cadre d'un stage de M2 par Jérémy Ganofsky (que j'ai co-encadré) et qui est maintenant en doctorat dans l'équipe. Nous ne pouvions pas faire reposer la réussite de son stage sur une potentielle amélioration de CAARS et nous avons préféré implémenter la manière standard présentée également dans le chapitre 1 en utilisant un nouveau gestionnaire de pipeline, Nextflow. Ce choix a également été guidé par l'intérêt pédagogique de ce stage. Le pipeline (partie 1,2,3) est donc le fruit d'un travail collaboratif avec Jérémy.

4.2.2.1 Partie 1 : Récolte des données sur le NCBI et pré-analyse de qualité

La première partie du pipeline permet de trier les données d'intérêt parmi l'ensemble des échantillons de RNA-Seq disponibles dans la Short Read Archive du NCBI. Une première recherche permet de récupérer tous les échantillons appartenant au niveau taxonomique "rodent". Puis, une seconde requête cherche le mot "kidney" dans les métadonnées de ces échantillons. Finalement, nous téléchargeons une partie des lectures (50 000) de chacun des échantillons retenus par la double recherche pour avoir une première idée de la qualité des données.

Ce premier tri nous permet d'obtenir une table contenant l'ensemble des échantillons disponibles et leurs métadonnées (espèce, tissus, nombre de lectures, publication correspondante...) ainsi qu'une idée de la qualité des données. Nous avons ensuite vérifié manuellement que les échantillons provenaient de données uniques de rein et non d'un mélange d'organes. Finalement, nous avons récupéré les identifiants des échantillons respectant ces critères. Pour les espèces avec plus de trois échantillons, nous en avons sélectionné trois de manière à essayer d'avoir les échantillons les plus homogènes possibles en termes de nombres de lectures, outils de séquençages (uniquement Illumina, en privilégiant les données paired-end), publications (limiter le nombre de papiers d'origine des échantillons).

4.2.2.2 Partie 2 : Assemblages, annotations et quantifications

La seconde partie du pipeline prend en entrée une table listant les échantillons sélectionnés précédemment avec leur identifiant et leur espèce. Chacun des échantillons est téléchargé à partir de son identifiant avec fastq-dump (https://github.com/ncbi/sra-tools), puis nettoyé en utilisant Trimmomatic (BOLGER et collab., 2014) . Une analyse de qualité des échantillons complets est ensuite réalisée avec fastqc (http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc) et les résultats sont agglomérés avec multiqc (EWELS et collab., 2016). Les lectures nettoyées sont ensuites assemblées *de novo* avec Trinity (GRABHERR et collab., 2011). La qualité des assemblages est analysée avec BUSCO (SIMÃO et collab., 2015) pour connaître le contenu en gènes des assemblages et par TrinityStats (HAAS et collab., 2013) pour calculer les indicateurs communs de qualité des

1. Les containers permettent de fournir un environnement indépendant de la machine hôte et contiennent tous les outils nécessaires pour exécuter différentes tâches.

assemblages, telle que le nombre de contigs, la N50 [1]. Transdecoder (https://github.com/TransDecoder/TransDecoder) est ensuite utilisé sur l'assemblage pour inférer les CDS (Coding DNA Sequence) et les ORFs (Open Reading Frame) dans chacun des contigs afin de récupérer uniquement les séquences codantes pour chacune des espèces et d'enlever les UTRs (Untranslated Transcribed Region).

L'ensemble de ces CDS est ensuite annoté par similarité de séquences en utilisant Blast (CAMACHO et collab., 2009) sur une base de données d'orthologues de rongeurs 1:1, roNOG, qui est un sous ensemble d'EGGNOG 4.5.1 (HUERTA-CEPAS et collab., 2015; POWELL et collab., 2013). Cette base de données a été créée à partir des séquences de cinq rongeurs provenant d'Ensembl. D'autre part, pour chaque espèce, Kallisto (BRAY et collab., 2016) est utilisé pour quantifier l'expression des CDS. A la fin de ce pipeline, nous pouvons récupérer l'ensemble des quantifications de l'expression des gènes ainsi que leur annotation, ce qui nous permet de générer une table de comptes qui sera ensuite utilisée pour les analyses d'expressions différentielles.

4.2.2.3 Partie 3 : Construction des alignements multiples de séquences

Le dernier pipeline a pour but de créer l'arbre d'espèces ainsi que le jeu de données "séquences" qui est composé des alignements multiples de séquences.

Les données d'entrée de ce pipeline sont les assemblages *de novo* du pipeline précédent auxquels on ajoute les transcriptomes complets des espèces provenant d'Ensembl. Cela permet d'augmenter le nombre d'espèces dans notre jeu de données et de ce fait le nombre de transitions. Comme ces derniers étaient absents dans le pipeline précédent, l'étape de recherche des CDS est renouvelée ainsi que l'annotation des CDS pour les nouvelles espèces avec la base de données roNOG comme présenté précédemment. On appellera par la suite "cluster eggnog" l'ensemble des séquences de cette base de données correspondant à un gène. De plus, on est capable d'associer à chacun des clusters un nom de gène que l'on appellera par la suite MGI (pour Mouse Genome Informatics). A ce stade, chaque séquence est associée à un cluster eggnog. Pour chacun des clusters, la séquence ayant le meilleur score pour chaque espèce est récupérée pour ne garder qu'une seule séquence par espèce et par cluster.

Ensuite, chacun des clusters est aligné avec MAFFT (KATOH et STANDLEY, 2013) puis nettoyé avec HMMCleaner (FRANCO et collab., 2019) afin d'enlever les régions mal alignées. Puis, les sites avec un taux d'indels supérieur à 10% sont supprimés par un script *ad hoc* ainsi que les séquences ayant été nettoyées à plus de 50% par HMMcleaner. Les alignements sont réalisés au niveau protéique puis on utilise un outil "transferCleaner.pl" (FRANCO et collab., 2019) pour faire la traduction inverse (reverse traduction) des séquences et ainsi avoir un jeu de données au niveau protéique en acide-aminés mais aussi au niveau des codons en nucléotides.

La dernière étape de ce pipeline consiste à inférer l'arbre d'espèces de l'ensemble des espèces du jeu de données. Comme le jeu de données est très gros, l'inférence de l'arbre d'espèces à partir de l'ensemble des données prendrait énormément de temps. Pour contourner ce problème, nous avons procédé en deux temps, d'abord pour inférer la topologie et ensuite la longueur des branches.

Pour inférer la topologie de l'arbre des espèces, nous avons inféré 10 arbres d'espèces par maximum de vraisemblance en utilisant raxml-ng (KOZLOV et collab., 2019) (modèle LG+G) à partir de 10 sous-échantillonnages indépendants des données (200 sites pris aléatoirement dans 500 alignements de gènes au niveau nucléique, le nombre de sites utilisés a été défini de telle sorte que les topologies soient congruentes).

Puis, nous avons calculé la vraisemblance de chacune des 10 topologies d'arbres en utilisant l'ensemble du jeu de données. Nous avons retenu la topologie la plus vraisemblable. Enfin, nous

1. Taille du contig tel que 50% des bases de l'assemblage sont comprises dans des contigs de taille supérieure à cette taille

avons utilisé l'ensemble des données pour calculer les longueurs de branches de l'arbre d'espèces en fixant la topologie retenue.

4.2.3 Annotation des environnements des espèces

Le milieu de vie de l'ensemble des espèces présentes dans le jeu de données a ensuite été annoté. Pour cela, nous avons utilisé une base de données, GBIF (https://www.gbif.org/en/occurrence/search), qui contient les lieux de piégeage de 4 600 783 rongeurs. Pour chacune des espèces, nous avons d'abord récupéré les coordonnées GPS des individus présents dans la base de données et les avons croisées avec les variables environnementales (récupérées dans worldclim, https://www.worldclim.org/bioclim). Ensuite, nous avons nettoyé le résultat de cette requête pour enlever certaines aberrations, telles que des coordonnées nulles, les lieux correspondant à des zoos, des musées ou des laboratoires de recherche ou des méta données incohérentes (coordonnées GPS ne correspondant pas au pays de capture).

Afin d'annoter les transitions convergentes et les orienter, nous avons également utilisé ce pipeline pour annoter le milieu de vie de nombreuses espèces présentes dans une phylogénie très fournie des rongeurs. En effet, elle contient 2 260 espèces organisées en 474 genres (FABRE et collab., 2012). Sur les 2 260 espèces, nous avons obtenu les variables bioclimatiques pour les habitats de 1 898 espèces. Nous avons ensuite reporté ces valeurs sur la phylogénie et inféré les conditions ancestrales par Brownian Motion, en utilisant le package phytools (REVELL, 2011) avec un modèle ER (pour "Equal-rate model"). Cela nous a permis de sélectionner un critère permettant de définir nos groupes environnementaux mésiques et xériques. Nous avons utilisé le trimestre le plus sec (variable Bio 17 de worldclim) comme critère pour réaliser nos groupes d'environnements plutôt que la pluviométrie totale (Bio 12) ou le mois le plus sec (Bio 14). En effet, Bio17 prend en compte le fait qu'une espèce puisse avoir des saisons très sèches et des saisons avec un accès à l'eau, ou bien qu'une espèce a de l'eau tout au long de l'année, même en faible quantité. Pour illustrer notre choix, prenons les Fukomys. Deux des espèces de Fukomys présentent une moyenne annuelle élevée car la pluviométrie de novembre à février est très élevée, cependant de mai à septembre, l'apport en eau est quasiment nul. Sachant que ces espèces n'hibernent pas, elles subissent directement ce manque d'eau. Avec l'utilisation de la variable Bio12, nous n'aurions pu mettre en évidence cela. Par exemple, Peromyscus maniculatus vit dans un habitat où la moyenne annuelle de pluviométrie est très faible mais où les moyennes mensuelles sont très stables, ce qui signifie que cette espèce a un accès constant à l'eau.

Nous avons utilisé la bibliographie (IUCN) sur les espèces sélectionnées par le pipeline de construction du jeu de données pour fixer le seuil à partir duquel on définit une espèce comme mésique ou xérique. Finalement on a choisi 40 mm d'eau lors du trimestre le plus sec (Bio 17).

A noter : Il est très difficile d'annoter les transitions environnementales sur un arbre phylogénétique. Nous avons choisi de nous appuyer sur l'ensemble des 1 889 espèces pour lesquelles il est possible d'associer une pluviométrie, puis d'inférer les états ancestraux, pour maximiser la représentativité de ce calcul. Cette approche est puissante sur de grands jeux de données, mais peut échouer à classifier les adaptations très locales de certaines espèces (qui, par exemple, vivent dans des rares bosquets humides d'une zone globalement aride).

Par ailleurs, notons qu'il est difficile d'inférer un niveau environnemental (xérique versus mésique) bien représentatif des milieux et de la façon de vivre de nos espèces. En effet, même si une espèce vit dans un milieu dit aride ou xérique selon la pluviométrie, cela ne prend pas en compte les adaptations comportementales. Certaines espèces vivent principalement la nuit et restent cachées la journée pour éviter la chaleur, d'autres ont des périodes d'hibernation ou d'estivation évitant ainsi des saisons entières. Ces adaptations n'ont pas été prises en compte, car cela est un travail titanesque à l'échelle de 2 000 espèces, mais nous avons estimé que la variable Bio17 était probablement la plus représentative. De plus, les adaptations individuelles pourront être discutées a posteriori.

4.2.4 Composition finale de notre jeu de données

A l'issue du pipeline de construction du jeu de données et de l'annotation des habitats de nos espèces, nous avons 32 espèces de rongeurs appartenant à 10 familles différentes dont 16 espèces xériques et 16 mésiques (Figure 4.1). La topologie de l'arbre d'espèces obtenue à l'issue du pipeline est congruente avec la phylogénie des rongeurs (FABRE et collab., 2012).

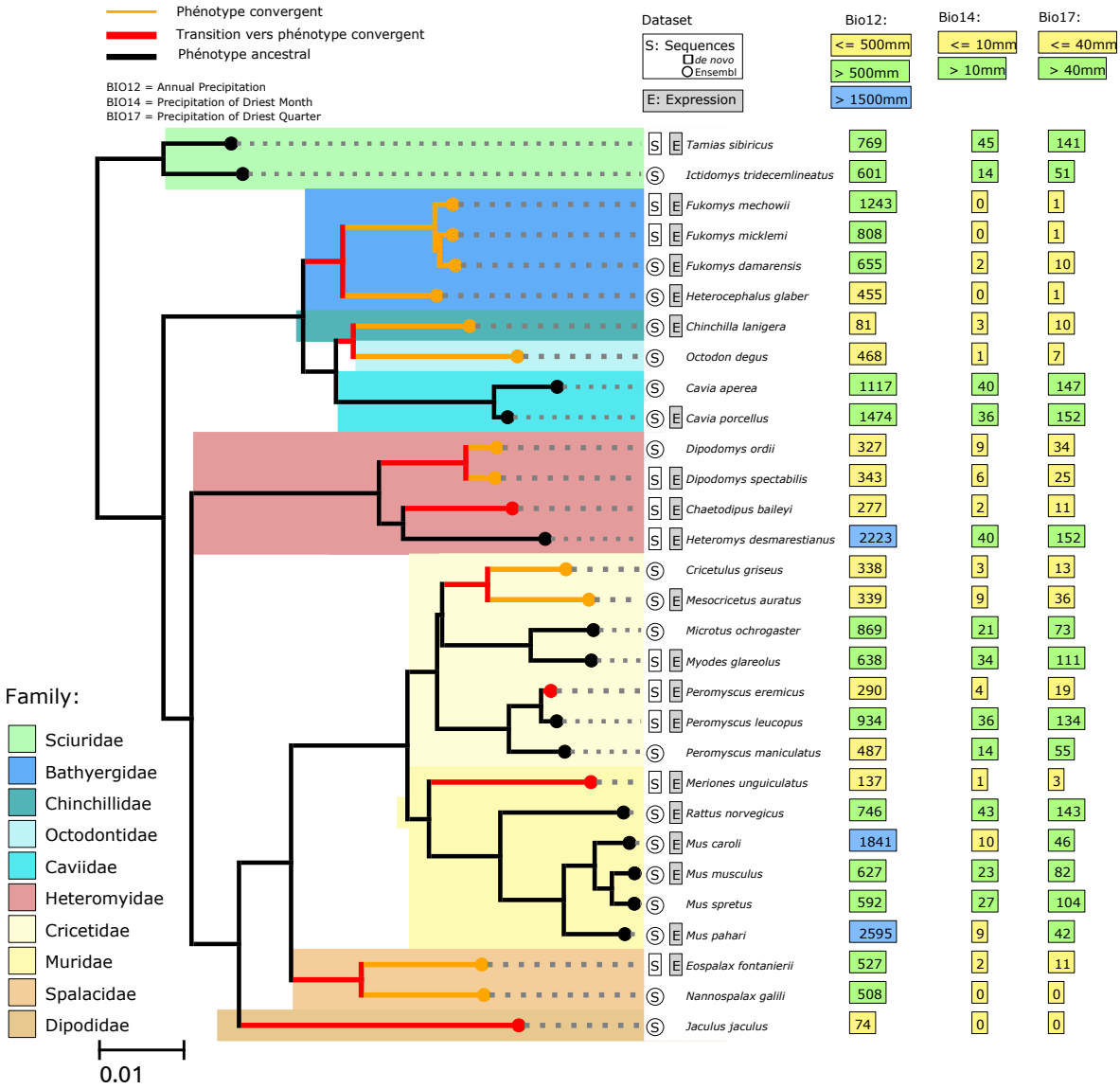

FIGURE 4.1 – Arbre des espèces de notre jeu de données. La présence de chacune des espèces dans l'un des sous jeux de données, "séquences" et "expression" est annotée à droite de l'arbre ainsi que les variables bioclimatiques associées aux habitats de chacune des espèces (Bio 12, Bio 14, et Bio 17). Dans l'arbre, la couleur des branches indique l'habitat des espèces, en noir les espèces mésiques, en orange et en rouge les espèces xériques. Les branches colorées en rouge marquent l'adaptation d'une espèce ancestrale mésique vers un milieu xérique. Les longueurs de branches ainsi que la topologie de l'arbre ont été calculées par maximum de vraisemblance en utilisant un modèle LG+G (cf section 4.2.2.3 pour plus de détails).

Le jeu de données "expression" est composé des 22 espèces, sur les 32, pour lesquelles des données RNA-Seq de rein sont disponibles (données de type Illumina). Parmi ces 22 espèces, 11 sont xériques et 11 sont mésiques. Les 11 espèces xériques se répartissent dans 8 transitions, c'est à dire dans 8 adaptations indépendantes à la vie en milieu xérique.

Le jeu de données "séquences" est composé de 30 espèces, 11 pour lesquelles nous avons uniquement des données RNA-Seq, 10 pour lesquelles on a seulement les transcriptomes complets

provenant d'Ensembl (Version 92) et 9 pour lesquelles on a les deux. Dans ce cas, le transcriptome provenant d'Ensembl est privilégié. Parmi ces 30 espèces, 16 sont définies comme xériques et se répartissent en 9 transitions.

A noter : la composition du jeu de données "séquences" diffère par l'absence des 2 espèces mésiques du genre Abrothrix qui ont été écartées dans le jeu des données "séquences" à cause de longues branches dans l'arbre d'espèces et qui seront possiblement réintroduits dans de futures analyses après des tests complémentaires. Elles ont été conservées dans le jeu de données "expressions" car elles ne présentent pas de positions extrêmes dans l'ACP (Analyse en Composantes Principales).

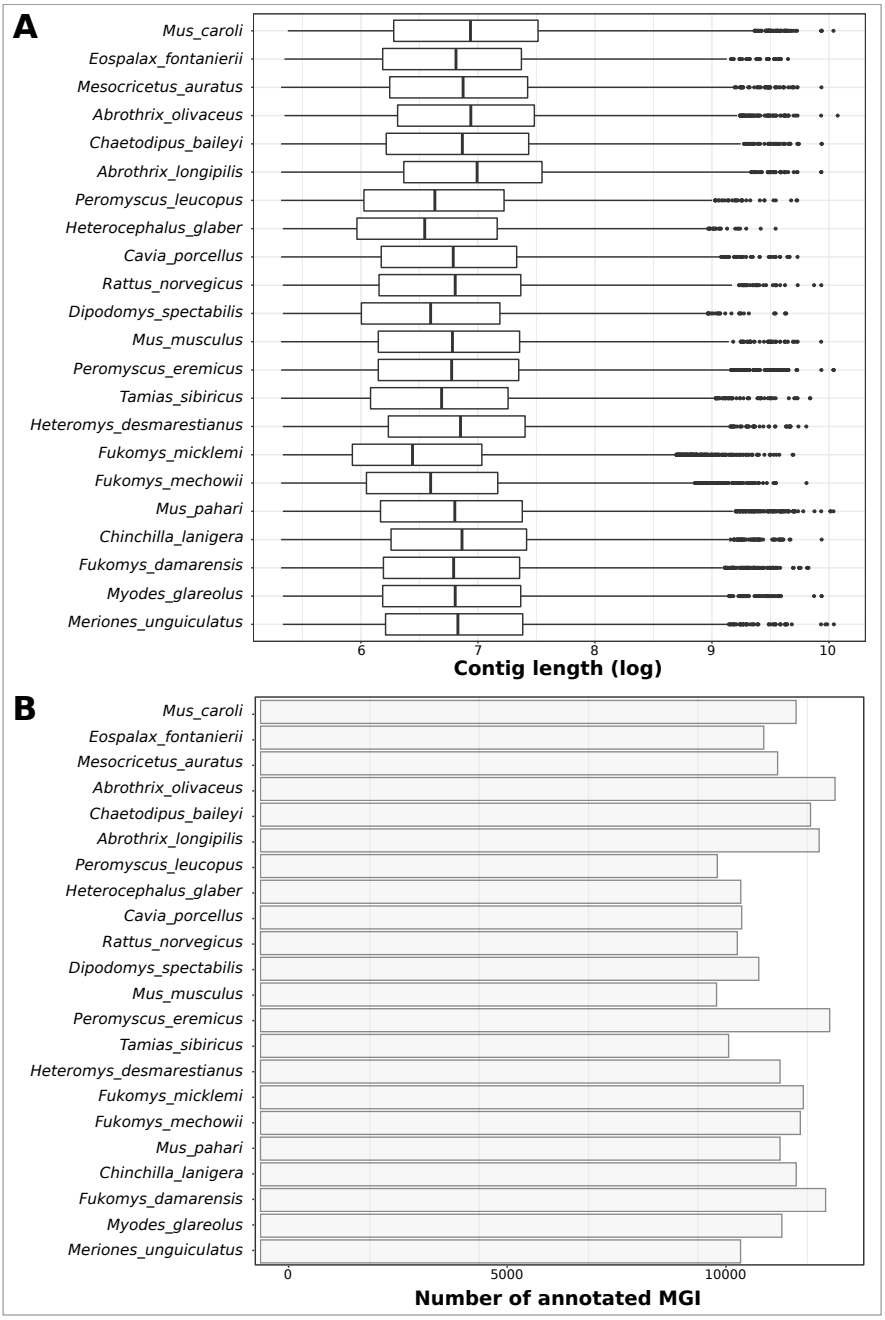

FIGURE 4.2 – Statistiques d'assemblages *de novo* des 22 espèces du jeu de données pour lesquelles on a utilisé des données RNA-Seq. (A) Taille moyenne des contigs dans l'assemblage de chacune des espèces. Les valeurs ont été transformées par le logarithme népérien pour des soucis de représentation. (B) Nombre de gènes annotés pour chacune des espèces.

4.2.5 Validation du jeu de données d'expression et des assemblages

4.2.5.1 Statistiques sur les assemblages

La comparaison des assemblages des transcriptomes des 22 espèces pour lesquelles on ne disposait que de données RNA-Seq montre des qualités assez homogènes, notamment au niveau de l'annotation des transcripts (Figure 4.2 et Table supp. 4.1).

Les assemblages contiennent entre 54 610 et 418 923 transcripts, ce qui représente de forts écarts. Cependant, lorsque l'on regarde la longueur des transcripts, les assemblages semblent homogènes, excepté quelques espèces telles que *Fukomys micklemi* (un seul individu utilisé), *Dipodomys spectabilis* ou *Heterocephalus glaber*. La longueur moyenne des transcripts pour chacune des espèces des transcripts est comprise entre 601 pb et 1 343 pb et la longueur médiane entre 321 pb et 630 pb. L'analyse réalisée avec BUSCO montre que 65% à 88% des gènes de la base de données de référence sont complets. De même, le nombre de gènes annotés se situe entre 10 838 et 13 496. Cela signifie que l'on n'a pas d'assemblage pour une espèce qui est d'une qualité nettement plus mauvaise que les autres. Cela aurait pu induire un biais dans les analyses suivantes.

4.2.5.2 Composition du jeu de données d'expressions

Au terme de ce pipeline, le jeu de données "expressions" est composé de 15 439 familles de gènes orthologues, que l'on appellera par la suite, "gène", pour lesquelles on possède au moins les niveaux d'expression pour une espèce. Pour étudier l'évolution de l'expression, par exemple avec DEseq2 (LOVE et collab., 2014), nous devons conserver uniquement les gènes pour lesquels des comptes sont obtenus pour tous les individus. De plus, les gènes mitochondriaux ont été exclus de la table (cf paragraphe ci-après). Le jeu de données pour l'analyse est réduit à 6 747 gènes, pour 22 espèces et 55 individus. Chaque espèce contient entre 1 et 3 individus.

Nous avons extrait les 20 gènes les plus exprimés, ainsi que les 20 gènes les plus variables entre espèces et nous avons vérifié qu'il n'y avait pas de chimère [1] parmi ces gènes. Par contre, nous avons observé des contaminations entre espèces pour les gènes mitochondriaux, qui sont très fortement exprimés. Ces contaminations sont apparentes chez les rats-kangourous, représentés par trois espèces dans notre jeu de données. A partir de chacun des trois échantillons, nous avons pu reconstruire le *cytochrome B* des trois espèces de rat-kangourou étudiées dans (MARRA et collab., 2012, 2014). Pour contourner ce problème, nous avons décidé de retirer les gènes mitochondriaux de l'analyse.

4.2.5.3 L'ACP pour détecter les individus ou espèces possiblement problématiques

Une analyse multivariée, par exemple, une ACP (Analyse en Composantes Principales), est une étape préliminaire classique avec un jeu de données d'"expressions" (LOVE et collab., 2014). Elle permet, sans à priori, de détecter des effets confondants généraux (par exemple, origine ou type des données, sexe, âge, etc.) ainsi que des échantillons problématiques qui apparaîtraient comme des points aberrants sur la carte de l'analyse. L'ACP est d'autant plus importante, dans notre cas, que nos données sont d'origines très diverses.

L'ACP nous a permis de diagnostiquer des espèces pour lesquelles l'assemblage était perfectible et ainsi d'améliorer le jeu de données. En effet, les librairies correspondant à trois individus avaient été choisies arbitrairement pour chaque espèce, pour faire l'assemblage et l'étude de l'expression (cf section 4.2.2). Nous avons observé, pour *Dipodomys spectabilis* et *Fukomys mechowii* (Dsp et Fme, respectivement), qu'un échantillon de chacune de ces espèces avait une position aberrante sur la carte de l'ACP. Nous avons donc refait les assemblages en remplaçant ces échantillons, ce qui améliore l'homogénéité intra-espèce.

1. fusion de gènes dûe à des problèmes d'assemblages et qui n'ont pas un sens biologique

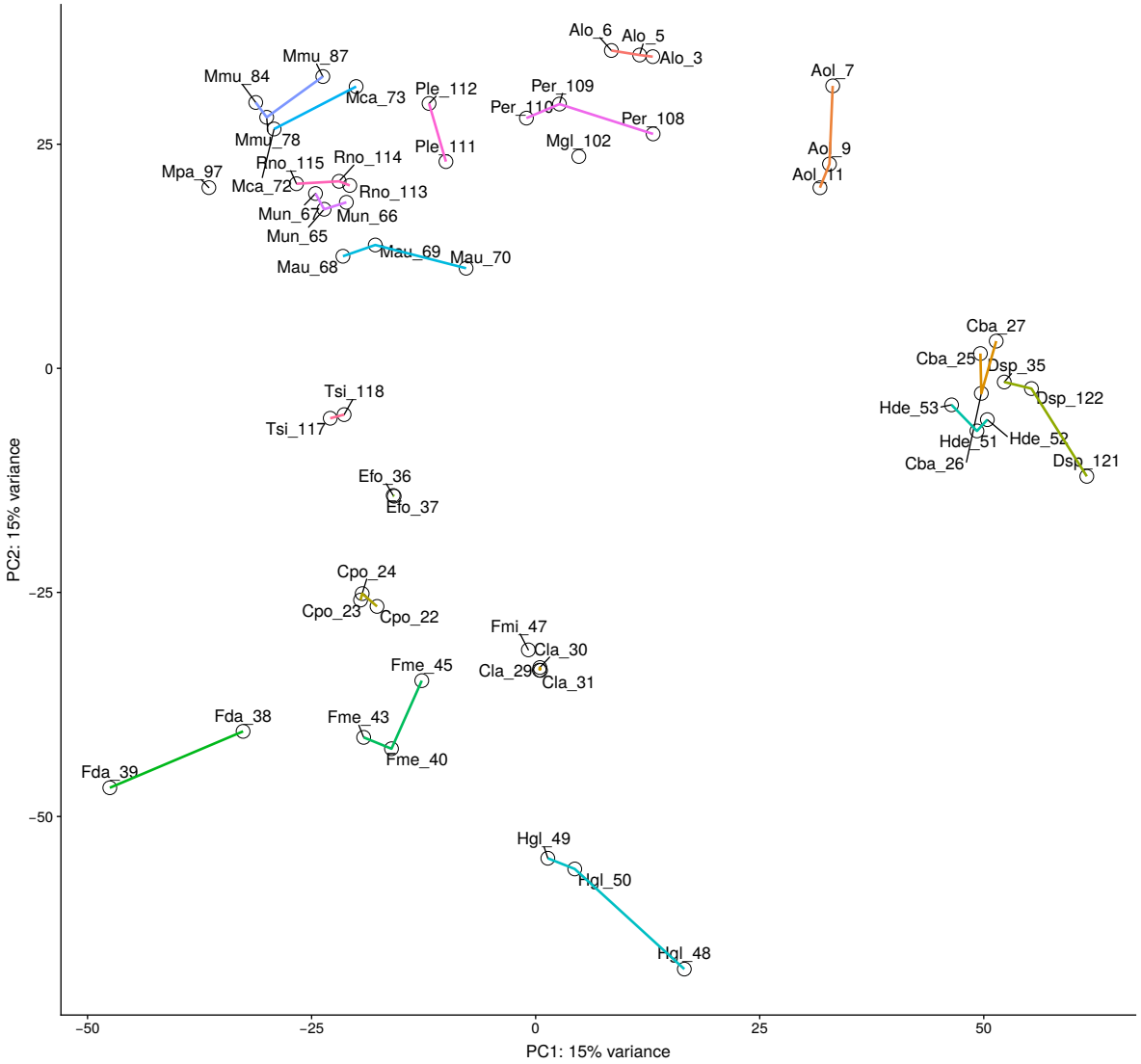

FIGURE 4.3 – Analyse en Composantes Principales sur les niveaux d'expressions normalisés de l'ensemble des gènes. Chaque point est un individu. Les étiquettes correspondent à la première lettre du genre et les deux premières lettres de l'espèce associées à un numéro unique par individu. Les lignes de couleurs relient tous les individus d'une même espèce. Abréviations des espèces : Tsi, *Tamias sibiricus* ; Rno, *Rattus norvegicus* ; Ple, *Peromyscus leucopus* ; Per, *Peromyscus eremicus* ; Mgl, *Myodes glareolus* ; Mpa, *Mus pahari* ; Mmu, *Mus musculus* ; Mca, *Mus caroli* ; Mau, *Mesocricetus auratus* ; Mun, *Meriones unguiculatus* ; Hde, *Heteromys desmarestianus* ; Hgl, *Heterocephalus glaber* ; Fmi,*Fukomys micklemi* ; Fme, *Fukomys mechowii* ; Fda, *Fukomys damarensis* ; Efo, *Eospalax fontanieri* ; Dsp, *Dipodomys spectabilis* ; Cla, *Chinchilla lanigera* ; Cba, *Chaetodipus baileyi* ; Cpo, *Cavia porcellus* ; Aol, *Abrothrix olivaceus* ; Alo, *Abrothrix longipilis.*

Finalement, après le remplacement des assemblages de ces espèces, nous avons obtenu une ACP où les individus de chacune des espèces forment des groupes compacts dont les distances relatives sont cohérentes avec les distances taxonomiques (Figure 4.3).

4.2.5.4 Le signal phylogénétique important dans les données est un gage de qualité

Une ACP permet de définir des axes maximisant la variabilité présente dans les données, et ceci de façon agnostique. Un autre type d'analyse multivariée, la BCA (Between-Class Analysis, package ade4 (DRAY et collab., 2007)), permet de calculer la variance expliquée par des variables externes telles que l'environnement des échantillons ou la famille taxonomique. Par exemple, dans notre jeu de données, 50% de la variance est expliquée par la différence entre familles taxonomiques. L'étude

d'où proviennent les échantillons en explique 54%, mais une partie de cet effet est indistinguable de l'effet famille, car la plupart des analyses publiées ont travaillé sur des espèces proches donc au sein de la même famille de rongeurs. En enlevant l'effet famille au préalable, l'effet "etude" explique toujours 36% de la variance, ce qui n'est pas inattendu, au vu des différences d'année, de protocole, de méthode de séquençage, etc. Nous espérons tout de même des résultats biologiques intéressants à partir de notre jeu de données, mais gardons en tête le manque de contrôle que nous avons sur les caractéristiques des individus échantillonnés (sexe, âge, état physiologique ou pathologique) et les effets techniques confondants.

4.2.6 Validation du jeu de données d'alignements de séquences

4.2.6.1 Composition du jeu de données d'alignements de séquences

A la fin du pipeline de construction du jeu de données, nous avons récupéré 16 090 familles de gènes avec un nombre variable d'espèces dans chacune d'elles (Figure 4.4). Nous avons retenu seulement les 14 554 gènes avec au moins une séquence pour la moitié des espèces et au moins deux transitions convergentes afin de pouvoir utiliser les logiciels de détection de la convergence.

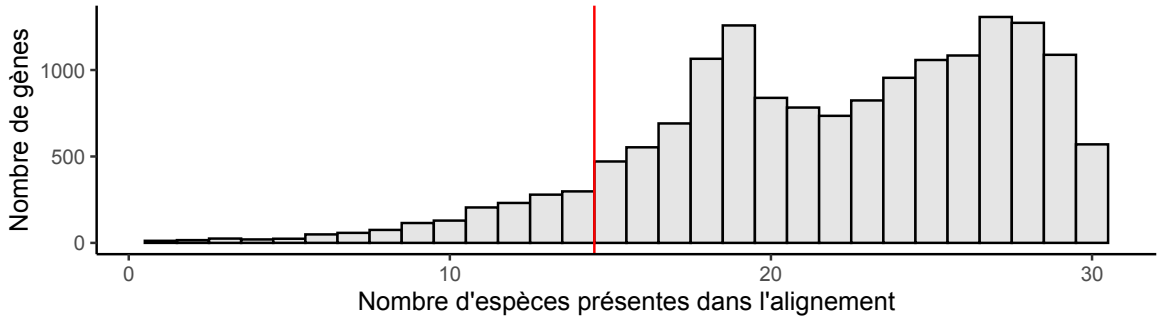

FIGURE 4.4 – Histogramme du nombre de familles de gènes avec un nombre donné d'espèces dans l'alignement. La ligne rouge représente le seuil utilisé pour conserver une famille de gènes.

4.2.6.2 L'analyse des arbres détecte la présence de paralogues

Comme l'annotation des séquences a été réalisée par similarité de séquences et non pas en utilisant des inférences d'arbres des gènes, nous suspections que des paralogues aient été groupés ensemble dans des familles de gènes (cf Partie 2).

Pour tenter d'identifier et d'enlever ces familles de gènes, nous avons détourné l'utilisation de phylter (`https://github.com/damiendevienne/phylter`), un outil en cours de développement par Damien De Vienne (LBBE, Lyon). Dans son utilisation normale, cet outil détecte et enlève des séquences aberrantes dans des jeux de données phylogénomiques. Cette méthode fonctionne par itérations, en enlevant les séquences de certaines espèces pour chaque famille de gène et en optimisant un score de concordance entre les matrices de distances des différentes familles de gènes. Dans notre cas, nous l'avons utilisé pour identifier les familles et les séquences qui sont responsables d'une discordance entre l'arbre obtenu pour une famille donnée et l'arbre moyen obtenu pour l'ensemble des données. Cette procédure pourra donc enlever les familles de gènes qui ont une forte probabilité de contenir des paralogues.

L'utilisation de phylter sur notre jeu de données montre que le nettoyage des alignements par HMMCleaner améliore significativement la concordance générale du jeu de données selon phylter, c'est à dire que les arbres individuels de chaque famille sont davantage similaires entre eux (le score de concordance de phylter passe de 0.52-0.58 à 0.88-0.91). Nous montrons en outre que la stringence du taux d'indels autorisés dans les alignements n'a pas d'influence notable sur la

concordance générale du jeu de données. Cependant, même après ce nettoyage scrupuleux, il reste des séquences identifiées par phylter, dont la position est aberrante dans l'arbre de la famille de gènes. Après un nettoyage par HMMCleaner suivi de la suppression des sites avec plus de 10% d'indels, nous avons identifié 420 familles de gènes que nous avons éliminées. Nous avons examiné à la main une dizaine de cas, qui correspondaient tous à des familles comprenant des mélanges de paralogues. Ceci reste à quantifier de manière générale, mais nous avons par précaution choisi d'éliminer toutes ces familles de gènes identifiées par phylter du jeu de données.

L'utilisation de phylter dans notre jeu de données est un moyen original de détecter les paralogues. Une étude plus poussée des séquences détectées par phylter permettrait de déterminer si d'autres phénomènes sont responsables d'incongruences dans nos phylogénies (composition en bases, chimères, etc). Dans toutes ces étapes de nettoyage, il faut prendre garde à ne pas écarter des gènes intéressants car montrant un signal fort de convergence évolutive, car justement ces gènes risquent de présenter des incongruences phylogénétiques.

4.3 Résultats préliminaires et Discussion

4.3.1 Résultats préliminaires sur l'analyse des séquences

Pour étudier la convergence dans les séquences codantes, nous avons utilisé les deux méthodes de détection qui donnaient les meilleurs performances d'après les résultats présentés dans le second chapitre, PCOC et Tdg09. Bien que plus performant, nous avons écarté diffsel à cause de son temps de calcul trop conséquent pour être utilisé sur des milliers de gènes. Nous avons réutilisé le pipeline implémenté pour les besoins de l'article sur la comparaison des méthodes de détection de la convergence (Partie 3). Cela a permis de gagner du temps sur l'implémentation et de garantir la reproductibilité des analyses.

4.3.1.1 Utilisation des gènes simulés pour définir un seuil de détection

4.3.1.1.1 Définir un seuil de détection spécifique à notre analyse

Lors du développement de PCOC et des comparaisons d'outils présentés au second chapitre (Partie 3), nous avons montré qu'il était impossible de fixer des seuils pour chaque méthode valables sur tout type de jeu de données. Ici, nous avons repris la méthodologie utilisée dans la publication associée à PCOC pour définir un seuil théorique qui correspondait à nos données réelles. Nous avons simulé des sites convergents (vrais positifs) et non-convergents (vrais négatifs) à l'aide du simulateur présent dans PCOC en utilisant la vraie topologie annotée avec les 9 transitions convergentes réelles de notre arbre d'espèces comme cadre pour la simulation afin de se rapprocher au plus près de nos données.

Les sites convergents ont été simulés selon le modèle PCOC, c'est à dire avec un changement convergent de profils d'acides aminés au niveau des transitions (PC) associé à une substitution (OC). Nous avons rendu nos simulations plus réalistes en simulant des gènes avec un nombre variable de transitions environnementales : nous avons généré des gènes avec des changements de profils d'acides aminés dans 3, 5, 7 et 9 transitions. Ainsi, certains gènes simulés sont totalement convergents (9 transitions) tandis que d'autres ne le sont que partiellement (convergences dans 3 sous-arbres sur 9, par exemple). Les simulations non-convergentes sont quant à elles simulées sans changement de profils associés aux transitions.

Nous avons également décidé de complexifier les simulations (convergentes et non convergentes), qui étaient par construction favorables aux méthodes de détection. Nous avons ajouté des petites variations aléatoires dans les longueurs de branches et introduit des changements de profils d'acides aminés dans les branches indépendamment des transitions convergentes.

Ensuite, nous avons utilisé ces sites pour étalonner les deux méthodes de détection utilisées, PCOC et Tdg09. Pour cela, nous avons fait tourner les méthodes sur ces sites et nous avons calculé pour chacun des seuils possibles (entre 0 et 1) trois indicateurs de performances :

— la **sensibilité** (sensitivity), c'est à dire le taux de sites "vrais positifs" effectivement détectés comme convergents par les méthodes parmi l'ensemble des positifs. Cet indicateur nous permet de mesurer la capacité à détecter des sites convergents s'il y en a.

— la **spécificité** (specificity), c'est à dire le taux de sites "vrais négatifs" détecté comme convergent par les méthodes parmi l'ensemble des négatifs. Cet indicateur nous permet de mesurer le taux d'erreur des méthodes lorsqu'il n'y a que des sites non-convergents.

— la **précision** (precision98_02), c'est à dire le taux de sites "vrais positifs" parmi l'ensemble des sites détectés comme convergents. Cet indicateur nous permet d'avoir une mesure théorique de la proportion de sites réellement convergents parmi l'ensemble des sites détectés comme convergents dans notre jeu de données. Cette mesure dépend donc de la proportion de sites convergents et non-convergents étudiés. Comme on ne s'attend pas à ce que la convergence soit majoritaire dans notre jeu de données (cf section 3.2.1.2) nous avons choisi un mélange de 98% de sites "vrais négatifs" et de 2% de sites "vrais positifs".

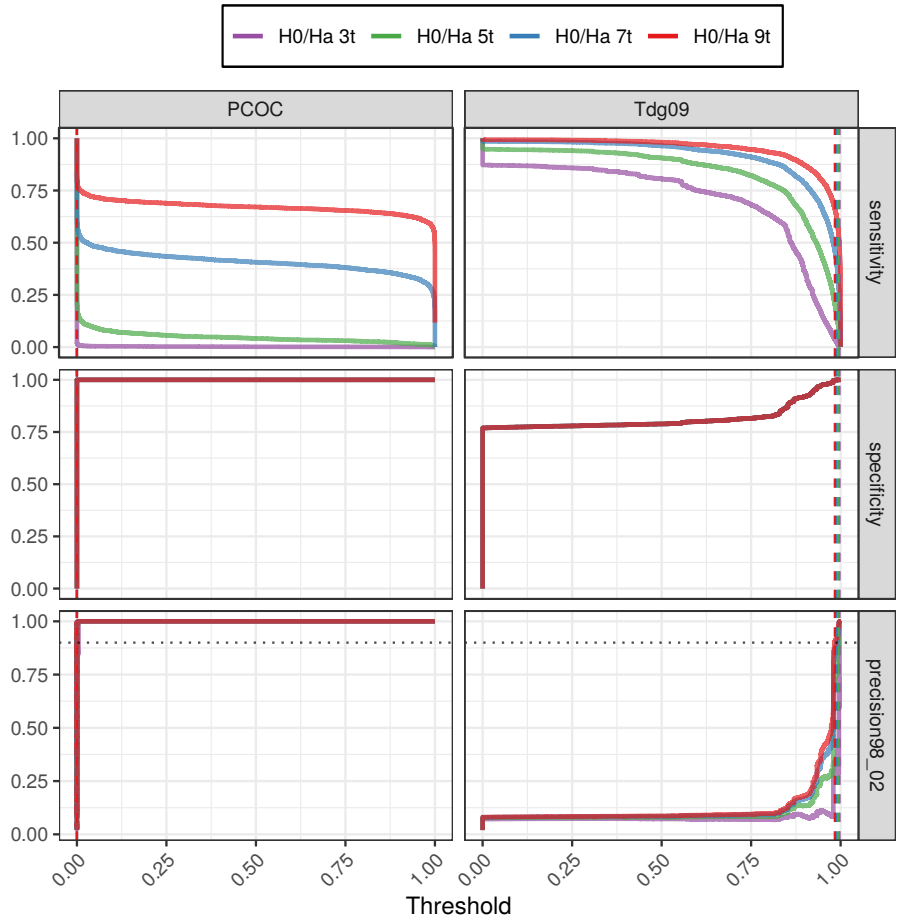

FIGURE 4.5 – Performances des méthodes de détection de la convergence sur des données simulées. Les simulations ont été réalisées avec un nombre effectif de transitions prises en compte différent : 3 (H0/Ha 3t) , 5 (H0/Ha 5t), 7 (H0/Ha 7t) ou 9 (H0/Ha 9t) des 9 transitions réelles. Les barres verticales représentent le seuil à définir pour atteindre une précision de 0.9 (ligne horizontale noire).

La figure 4.5 est construite sur le même modèle que celle du second article du chapitre 2. Elle montre la sensibilité, la spécificité et la précision obtenues en fonction du seuil utilisé pour déclarer un site convergent. La sensibilité et la spécificité diffèrent entre PCOC et Tdg09. On voit

que PCOC est très spécifique, assez sensible pour des gènes simulés avec au moins 7 transitions d'acides aminés sur les 9 transitions convergentes. Tdg09 est quant à lui très sensible quel que soit le nombre de transitions effectivement réalisés, par contre il est moins spécifique que PCOC. Cette différence de comportement des méthodes explique les courbes de précisions. Pour Tdg09, la précision augmente à partir du moment où le seuil est suffisamment stringent pour que la spécificité remonte. Pour PCOC, la spécificité est très bonne à partir d'un seuil peu stringent ce qui permet une très bonne précision à partir de ce seuil.

Basé sur ces simulations, nous avons défini un seuil tel que les méthodes aient une précision théorique de 90%, c'est à dire un taux de faux positif de 10% dans les sites détectés. Nous avons obtenu les seuils de 0.997 pour Tdg09, 5.6e-06 pour PCOC (Figure 4.6) .

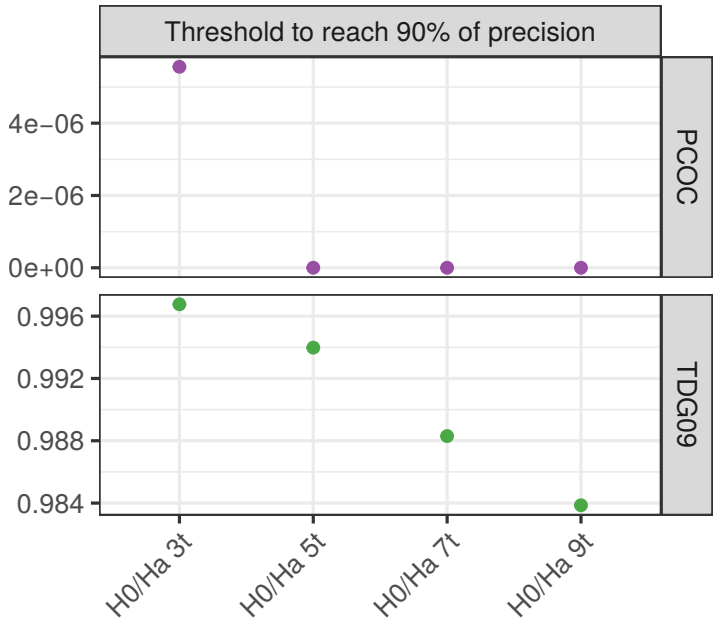

FIGURE 4.6 – Variation du seuil à définir pour atteindre 90% de précision en fonction du type de simulation. (cf légende Figure 4.5)

4.3.1.1.2 Le seuil défini de manière théorique ne semble pas assez stringent pour être appliqué aux données réelles

La figure 4.7 montre la distribution du nombre de sites réels ayant obtenu un score supérieur à un seuil donné, selon les différentes méthodes. Les lignes en pointillés indiquent les seuils définis précédemment. Si l'on applique ces seuils, pour Tdg09, on retiendrait 2 933 sites, soit 2 258 gènes avec au moins un site convergent et pour PCOC, 31 570 sites, soit 9 305 gènes avec au moins un site convergent. Parmi ces 9 305 gènes, le taux de sites convergents se situe entre 0.05% et 65%.

D'après ces seuils, la proportion de sites convergents dans le jeu de données serait très importante, 16% pour Tdg09 et 66% pour PCOC. Ces valeurs sont très élevées et pourraient relever de l'artefact.

D'un point de vue biologique, le jeu de données pourrait contenir un signal de convergence venant d'un processus confondant comme nous l'avons vu dans le second chapitre avec, par exemple, un changement de taille efficace des populations. On pourrait aussi penser au fait que ces espèces se sont adaptées à un milieu de vie assez hostile et cela a pu être le résultat d'un taux accéléré d'évolution. Pour tester cette dernière hypothèse, il faudrait regarder si les espèces convergentes présentent un taux d'évolution importante.

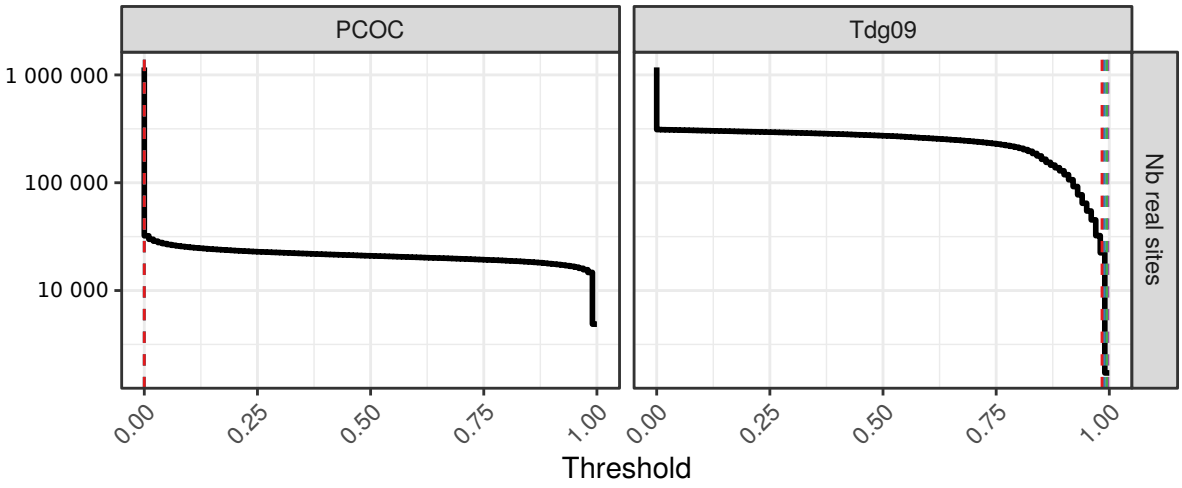

FIGURE 4.7 – Nombre de sites réels détecté en fonction du seuil fixé pour chacune des méthodes de détection de la convergence. Les seuils définis par les simulations sont indiqués par des lignes pointillées verticales.

D'après ces seuils, la proportion de sites convergents dans le jeu de données serait très importante, 16% pour Tdg09 et 66% pour PCOC. Ces valeurs sont très élevées et pourraient relever de l'artefact.

D'un point de vue biologique, le jeu de données pourrait contenir un signal de convergence venant d'un processus confondant comme nous l'avons vu dans le chapitre 2 avec, par exemple, un changement de taille efficace des populations. On pourrait aussi penser au fait que ces espèces se sont adaptées à un milieu de vie assez hostile et cela a pu être le résultat d'un taux accéléré d'évolution. Pour tester cette dernière hypothèse, il faudrait regarder si les espèces convergentes présentent un taux d'évolution importante.

D'un point de vue technique, le jeu de données pourrait encore contenir des paralogues. En effet, si dans un alignement, les séquences des espèces convergentes sont des paralogues, la similarité entre ces séquences serait supérieure à celle avec les autres séquences des espèces non-convergentes et donc cela pourrait mimer un signal de convergence. Les gènes pour lesquels on obtient 65% de sites convergents pourraient correspondre à ce genre de situations et cela malgré le fait que nous avons utilisé phylter. Une autre possibilité est que les simulations des sites non-convergents ne sont pas suffisamment réalistes ce qui empêche de fixer le seuil d'une manière convenable. Elles seraient donc, dans l'état, inutiles pour fixer des seuils dans le cadre d'une application sur des données empiriques.

Enfin, nous voudrions quantifier si la quantité de convergence observée est supérieure à ce qui serait attendu par hasard. Pour cela, nous envisageons de tester si le nombre de gènes convergents obtenus avec l'annotation réelle des environnements de chacune des espèces est supérieur aux nombres de gènes convergents obtenus lorsque que les environnements associés à chacune des espèces seraient attribués par hasard, c'est à dire que l'on déplacerait de manière aléatoire les transitions convergentes dans la phylogénie.

Pour conclure, nous restons convaincus de l'intérêt de cette méthode pour définir les seuils des différentes méthodes, mais il faut encore travailler sur l'implémentation de cette idée et écarter les possibles artefacts techniques pour pouvoir se concentrer sur les explications biologiques.

4.3.1.2 Des méthodes avec des résultats divergents

Bien que nous ne soyons pas entièrement satisfaits de la définition du seuil de nos méthodes, nous avons étudié l'intersection des résultats des méthodes. Nous avons considéré, en première approximation, un gène comme convergent, s'il possède au moins un site identifié comme convergent.

La figure 4.8 montre l'intersection des résultats avec les seuils présentés précédemment ou avec une définition arbitraire et plus stringente du seuil (0.999) pour réduire le nombre de gènes détectés. L'intersection des gènes présents dans les résultats des deux méthodes diminue fortement lorsque le seuil augmente. Cela suggère que les résultats ne sont pas corrélés. Ceci peut s'expliquer par le fait que PCOC et Tdg09 n'utilisent pas les mêmes bases pour détecter la convergence. En effet, bien que chacun cherche à détecter un changement de profils d'acides aminés, Tdg09 définit ses profils pour chacun des sites en utilisant la composition observée en acides aminés alors que PCOC utilise des profils prédéfinis. Ceci pourrait donc entraîner la détection de sites convergents de différentes natures.

Pour le moment, on définit un gène comme convergent à partir de la présence d'un site convergent mais il faudrait plutôt développer une méthode avec un indice permettant d'intégrer les scores de tous les sites d'un gène (cf. discussion second chapitre, section 3.4.2). De cette manière, on pourrait identifier des gènes avec un site ayant un score de convergence très élevé ou des gènes avec plusieurs sites dont les scores sont assez élevés.

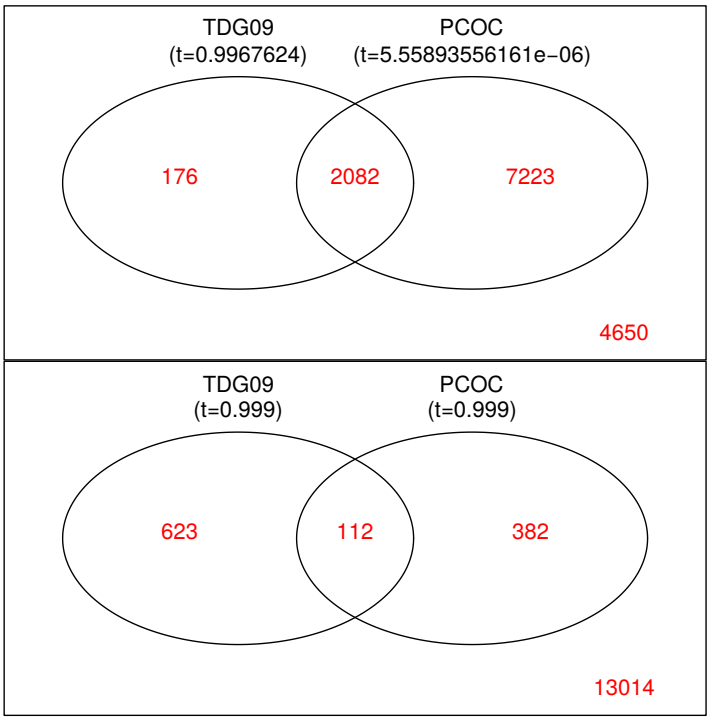

FIGURE 4.8 – Intersection entre les gènes avec au moins un site convergent détecté pour chacune des méthodes en fonction de seuils définis par des simulations théoriques (haut) et des seuils arbitraires (bas).

4.3.1.3 Des analyses d'ontologie de gènes (GO) révèlent de faibles enrichissements

Nous avons ensuite regardé si les 112 gènes détectés à la fois par PCOC et par Tdg09 aux seuils les plus stringents, c'est à dire ceux qui semblent porter le plus fort signal de convergence, semblent liés par leur fonction biologique. Pour cela nous avons procédé à une analyse d'enrichissement en utilisant une ontologie, la *Gene Ontology* (GO) (CONSORTIUM, 2018). Cette ontologie est un regroupement structuré de termes biologiques auxquels sont associés des gènes. Le but de cette analyse est donc de voir si une fonction biologique est sur-représentée dans le groupe de gènes étudiés. L'ontologie est structurée en trois sous-ensembles : les termes associés à des fonctions moléculaires (MF), à des processus cellulaires (BP) ou à des composants cellulaires (CC).

Dix termes GO appartenant à l'ensemble des fonctions moléculaires (MF) ressortent significativement enrichis (Figure 4.9), tels que "cytokine binding" (p-adjust = 0.0047) qui regroupe 7 des 112 gènes étudiés. Aucun terme ne ressort significativement enrichi pour les ensembles processus

cellulaires (BP) et les composants cellulaires (CC).

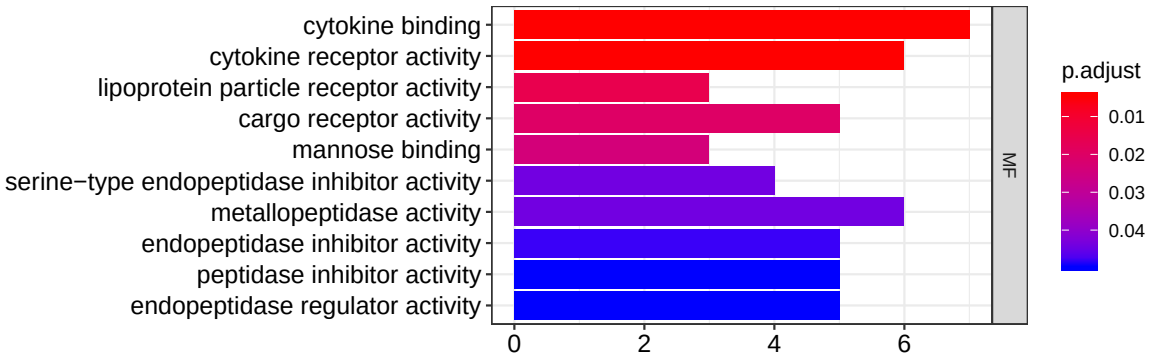

FIGURE 4.9 – Termes GO significativement enrichis en utilisant les 112 gènes détectés comme convergents par PCOC et Tdg09. L'ensemble des gènes du jeu de données a été utilisé comme liste de gènes de référence. La couleur des barres représente la valeur de la p-valeur ajustée pour ces termes et la longueur de la barre le nombre de gènes associés à ce terme présent dans le groupe de gènes étudié. MF : fonctions moléculaires.

A première vue, les termes GO qui ressortent ne nous paraissent pas directement liés à l'adaptation au milieu aride. Avant d'aller plus loin dans l'interprétation biologique, nous aimerions écarter la possibilité que ces termes soient enrichis par hasard. En effet, est-ce que si l'on tire au hasard 112 gènes, on obtiendrait des termes GO enrichis et si oui combien ? Pour tester cela, il faudrait faire des tirages aléatoires de 112 gènes et les utiliser pour faire des analyses d'enrichissement. On pourrait ensuite tester si le nombre de termes significatifs observé avec les vraies données n'appartient pas à la distribution du nombre de termes significatifs obtenus avec des données tirées au hasard. Il se peut également que ces gènes aient un taux d'évolution supérieur aux autres gènes et que des substitutions s'y soient accumulées. Cela augmenterait la probabilité de convergence liée à la dérive. Il faudrait donc également s'intéresser au taux d'évolution de ces gènes pour écarter cette hypothèse.

Les analyses d'enrichissement ne peuvent fournir qu'une information assez grossière. En effet, l'ontologie dépend fortement des analyses sur lesquelles elle s'appuie, ce qui la rend partiale. Certains processus très étudiés sont très complets alors que d'autres processus moins étudiés sont incomplets. L'ontologie n'est donc pas exhaustive mais elle reste une base utile de connaissances.

Cela signifie que si un groupe de gènes est lié par leur fonction biologique, cette fonction ne ressortira pas nécessairement significative dans une analyse d'enrichissement. En effet, il faut qu'au préalable ce groupe de gènes ait été annoté dans une ou plusieurs analyses antérieures. Dans ce cas, il est probable que la fonction biologique liant ce groupe de gènes sorte enrichie dans notre analyse.

Cependant, l'absence d'enrichissement dans les termes GO ne signifie pas absence de pertinence biologique. Et les analyses d'enrichissement sont un moyen facile à mettre en place pour débuter l'interprétation biologique de résultats.

4.3.2 Résultats préliminaires sur l'analyse des niveaux d'expression

4.3.2.1 Un signal de convergence au niveau de l'expression des gènes semble se détacher

4.3.2.1.1 La variance associée aux environnements mésiques et xériques est plus importante qu'attendue par hasard

Aucune composante de l'ACP ne correspondait à la différence d'environnements, mésique versus xérique, dans les données. Cependant, cette variable explique 6.1% de la variance des données d'après une analyse par BCA (une analyse multivariée qui cherche une variance associée

à un processus donné). Si l'on retire au préalable l'effet propre à la famille phylogénétique (en utilisant une WCA, Within-Class Analysis), l'environnement n'explique plus que 3.8% de la variance. Bien que modeste, cette différence est largement significative, comme en attestent des estimations obtenues après tirage au hasard des étiquettes d'environnements (p-valeur < 10^{-3}, Figure 4.10, gauche). L'effet de l'environnement reste significatif après contrôle pour l'effet de la famille (p-valeur=0.004, Figure 4.10 droite) ce qui est en faveur de la présence de convergence dans les données.

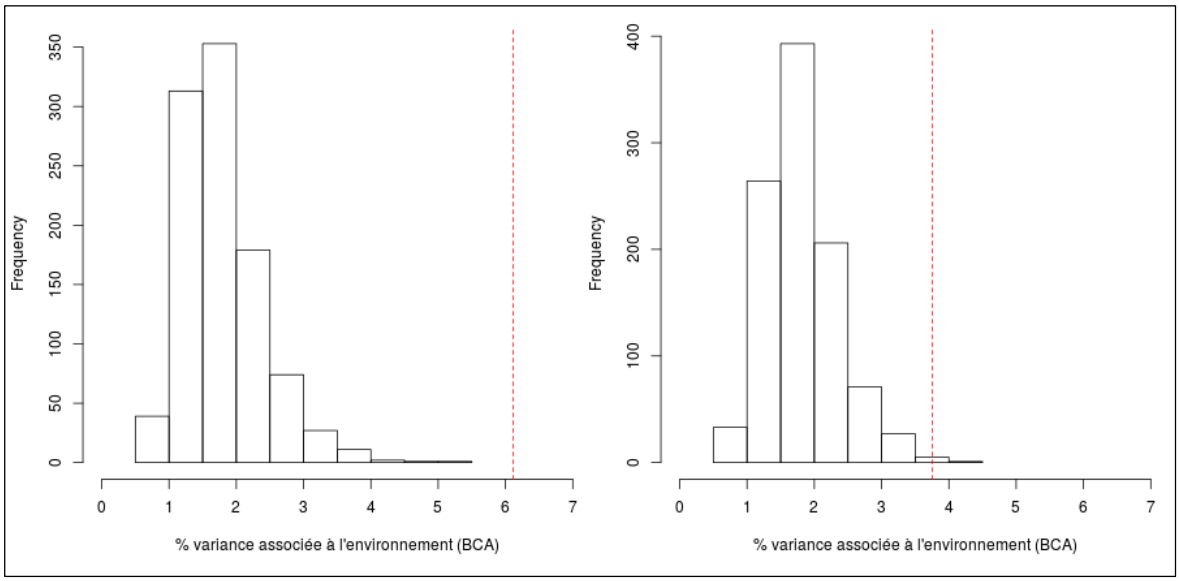

FIGURE 4.10 – Pourcentage de variance associée à la différence entre environnement xérique et mésique, pour les données d'expression, estimé par BCA (Between-Class Analysis, package ade4 (DRAY et collab., 2007)). Les valeurs observées sont indiquées par des traits pointillés en rouge et la distribution des valeurs attendues (obtenues par 1 000 mélanges aléatoires des étiquettes environnementales) représentée en noir. A gauche, valeurs obtenues sans prise en compte de la phylogénie, à droite, avec prise en compte de la phylogénie par une WCA préalable (Within-Class Analysis, package ade4 (DRAY et collab., 2007)).

4.3.2.1.2 Le nombre de gènes DE n'est pas plus important qu'attendu par hasard

En parallèle des analyses sur les séquences codantes, nous avons réalisé des analyses sur les niveaux d'expression des gènes exprimés dans le rein. Nous avons cherché à identifier des gènes différentiellement exprimés entre le milieu aride et mésique en utilisant DESeq2 (LOVE et collab., 2014), un outil habituellement utilisé pour faire ce type d'analyse. Lorsque l'on contraste les individus en fonction de leur environnement (mésique et xérique), on obtient 83 gènes différentiellement exprimés (DE) ($design =\sim environnement$, padj <0.1).

Nous avons cherché à savoir si ce nombre était plus qu'attendu par hasard. Pour cela, nous avons simulé 1 000 scénarios aléatoires où nous avons attribué au hasard les étiquettes "xérique" et "mésique" aux espèces et procédé à une analyse DE (Figure 4.11). Le nombre observé de gènes DE (ligne rose en pointillés) n'est pas significativement supérieur à ce qui est attendu par hasard (P=0.105). Nous ne pouvons donc pas affirmer qu'il y ait davantage de gènes à l'expression convergente dans les niveaux d'expression que ce qui est attendu par hasard d'après notre méthode de simulation.

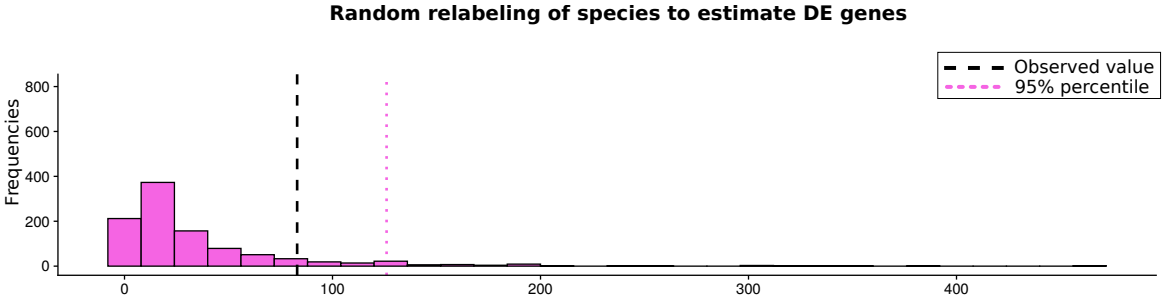

FIGURE 4.11 – Distribution du nombre de gènes DE obtenus par assignation aléatoire des environnements (mésique et xérique) des espèces. 1 000 tirages aléatoires ont été réalisés. La barre verticale noire indique la valeur réelle observée et la barre verticale rose indique la valeur pour le 95ème percentile.

4.3.2.1.3 Le nombre de gènes DE "up" régulés est plus important qu'attendu par hasard

Contrairement aux analyses classiques de gènes DE, nous n'avons pas divisé notre groupe de gènes DE en gènes sur-exprimés et sous-exprimés. En effet, au regard des profils d'expression de ces gènes, nous avons subdivisé ces profils de gènes en cinq sous-classes associées à de possibles processus d'adaptation (Figure 4.12). Les définitions des classes qui sont présentées ci-dessous ont été faites de manière pragmatique afin d'avoir un premier aperçu.

La première classe est celle des gènes "up" (Figure 4.12 a). Elle comprend 26 gènes tels qu'il y a un décalage global vers la hausse de l'expression de ces gènes pour les individus xériques par rapport aux individus mésiques. D'un point de vue biologique, il s'agirait de gènes qui ont globalement convergé vers des niveaux d'expression supérieurs lors de l'adaptation des espèces au milieu aride. Un gène doit remplir ces trois conditions pour être placé dans cette classe : le niveau d'expression moyen de ce gène pour les individus xériques doit être supérieur à celui des mésiques, le niveau d'expression minimum de ce gène pour les individus xériques doit être 1,5 fois supérieur à celui des mésiques et le niveau d'expression maximum de ce gène pour les individus xériques doit être supérieur à celui des mésiques.

La seconde classe est celle des gènes "down" (Figure 4.12). Elle comprend 17 gènes. Elle est définie dans les sens opposés de la catégorie "up" afin d'identifier les gènes ayant un décalage global vers la baisse de l'expression de ces gènes pour les individus xériques par rapport aux individus mésiques. D'un point de vue biologique, il s'agirait de gènes qui ont globalement convergé vers des niveaux d'expressions inférieurs lors de l'adaptation des espèces au milieu aride.

La troisième classe est celle des gènes "constraint" (Figure 4.12). Elle comprend 11 gènes tels que l'ensemble des niveaux d'expression de ces gènes chez les individus xériques représente un sous ensemble de possibilités exploité par les individus mésiques. D'un point de vue biologique, il s'agirait de gènes dont le niveau d'expression est contraint chez les espèces arides, ne représentant qu'une partie de la diversité ancestrale. Un gène doit remplir ces trois conditions pour être placé dans cette classe : la variance de l'expression de ce gène pour les individus mésiques doit être 2 fois supérieure à celle des xériques, le niveau d'expression maximal de ce gène pour les individus xériques doit être 2 fois inférieur à celui des mésiques et le niveau d'expression minimal de ce gène pour les individus mésiques doit être inférieur à 1.5 fois celui des xériques.

La quatrième classe est celle des gènes "diversified" (Figure 4.12). Elle comprend 29 gènes. Elle est définie dans le sens opposé de la classe "constraint". Cela garantit que l'ensemble des niveaux d'expression de ces gènes pour les individus xériques représente une augmentation de la diversité par rapport à celle présente chez les individus mésiques. D'un point de vue biologique, il s'agirait de gènes dont l'expression varie fortement par rapport à la diversité ancestrale mais qui n'est pas partagée par l'ensemble des individus xériques. Il pourrait donc s'agir de convergences non partagées entre l'ensemble des espèces xériques.

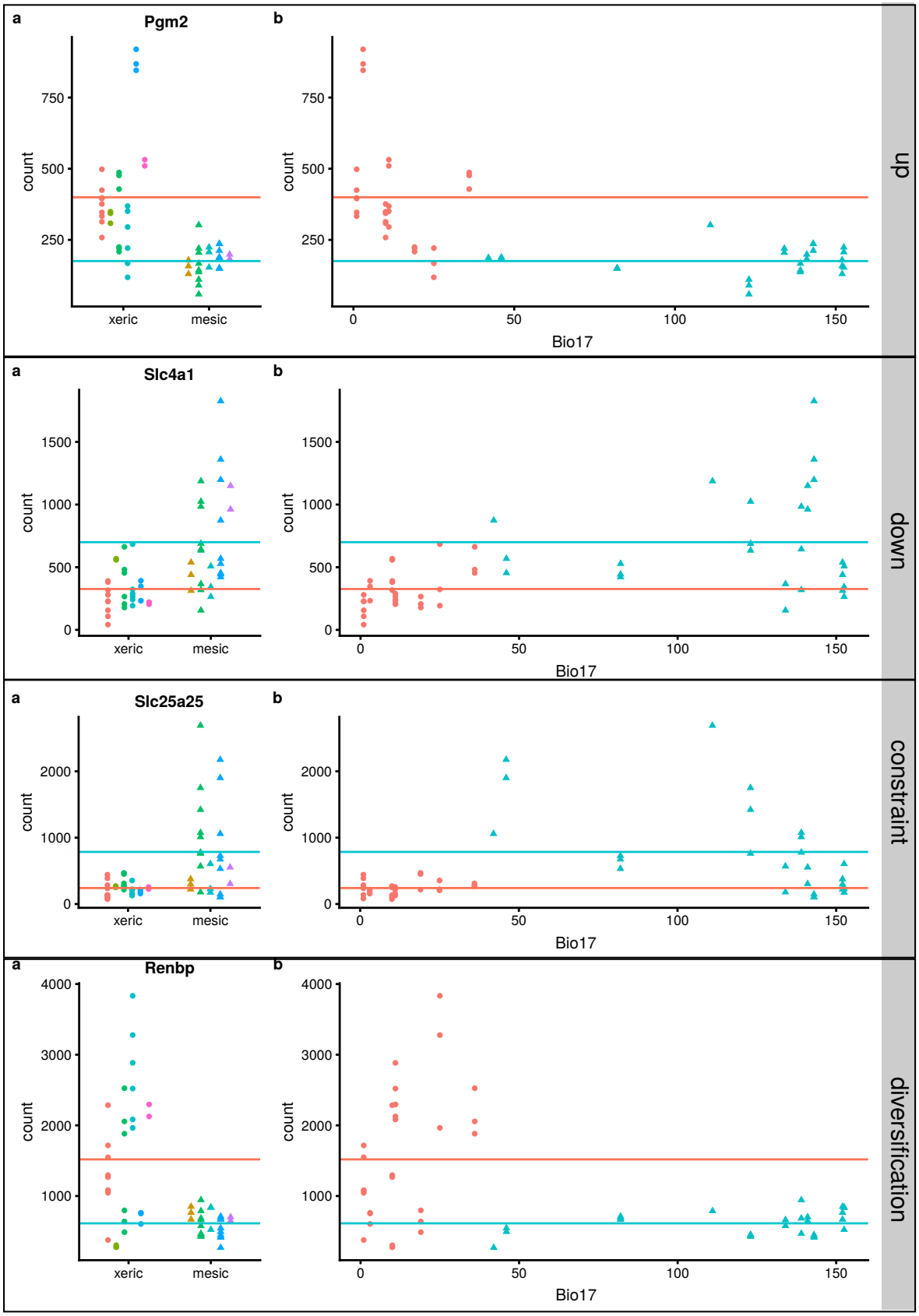

FIGURE 4.12 – Présentation de gènes représentatifs de chacune des catégories, "up", "down","constraint" et "diversified". Pour chacun des gènes, le nombre de comptes normalisés pour chacun des individus est représenté en fonction de sa condition xérique ou mésique (gauche) et en fonction de sa valeur de Bio17 (droite). La couleur des points correspond à la famille à laquelle appartient l'individu associé à ce compte pour le panel de gauche et sa valeur de Bio17 pour celui de droite. Les lignes indiquent la moyenne des comptes pour ce gène pour l'ensemble des individus de chaque condition, en rouge les xériques et en bleu les mésiques.

La dernière correspond aux gènes "unclassified" qui ne rentrent dans aucune de ces catégories et qui n'ont pas été classifiés. Dans notre cas, tous les gènes ont été attribués à l'une des quatre premières classes, bien que la définition des seuils a été réalisée indépendamment des gènes DE obtenus.

Après avoir effectué cette classification pour les données réelles, nous avons testé si le nombre de gènes obtenu dans chacune des classes était supérieur au nombre attendu par hasard. Au cours des 1 000 simulations présentées précédemment, nous avons également procédé à la classification des gènes dans ces quatre classes. Les gènes ne rentrant dans aucune de ces classes sont définis comme "unclassified".

La distribution des nombres de gènes DE obtenus lors de tirages aléatoires des environnements pour chacune des espèces (Figure 4.13) montre que le nombre de gènes "up" est plus important qu'attendu avant correction pour les tests multiples (P=0.022), le nombre de gènes "down" est marginalement plus important (P=0.077) et les nombres observés dans les autres classes ne sont pas différents de ce qui est attendu (P=0.121 pour la classe "constraint", P=0.098 "diversified"). Par contre, le nombre de gènes DE "unclassified" est souvent supérieur à la valeur observée, mais cela n'est pas significatif (1-P=0.122).

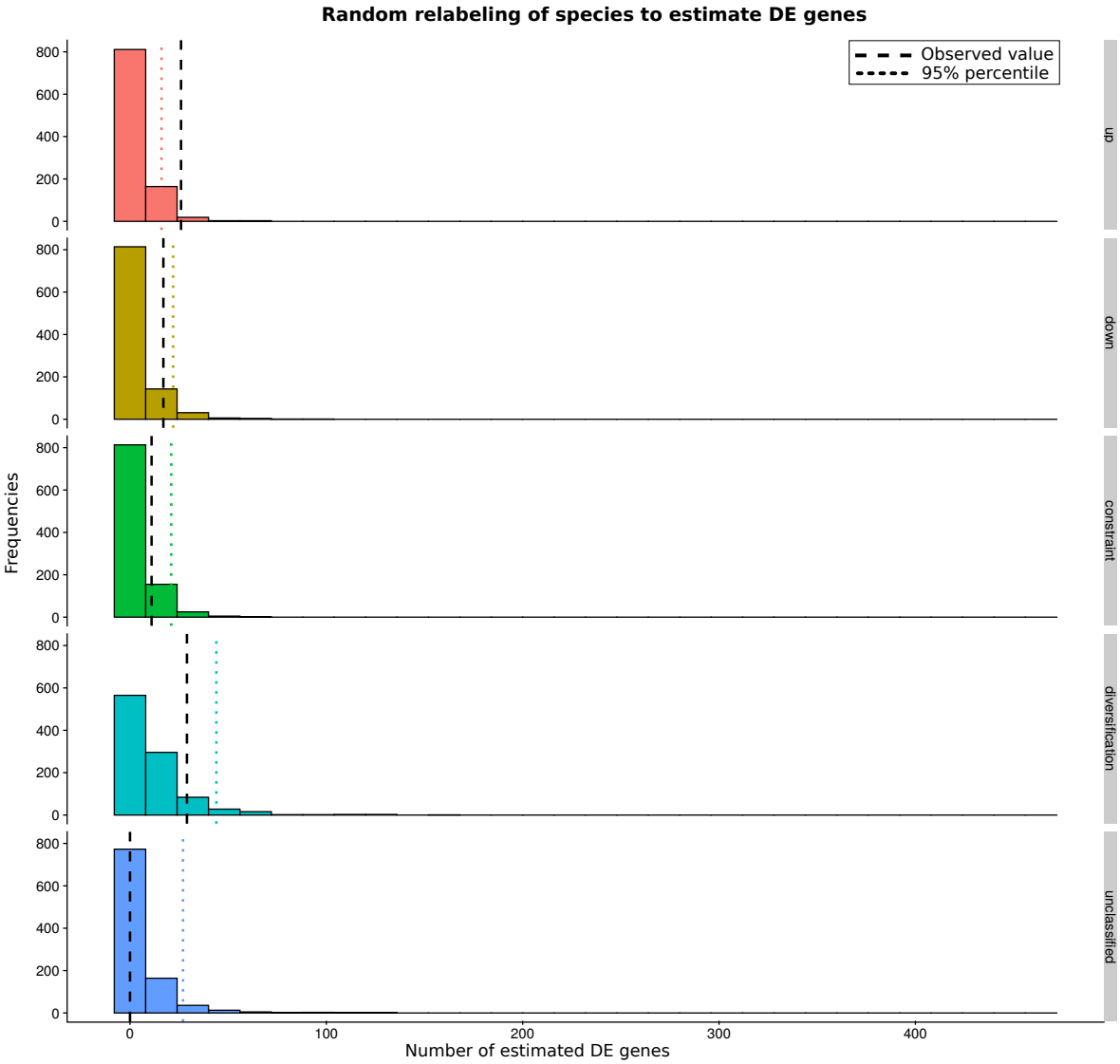

FIGURE 4.13 – Distribution du nombre de gènes DE obtenus par classe par assignation aléatoire des environnements (mésique et xériques) des espèces. 1 000 tirages aléatoires ont été réalisés. Les barres verticales noires indiquent les valeurs réelles observées pour chacune des classes et les barres verticales colorées indiquent les valeurs pour les 95ème percentiles pour chacune des classes.

En conclusion, ces résultats montrent qu'il y a des gènes qui ont une réponse convergente de leur expression en réponse à une adaptation à l'aridité. De plus, le nombre de ces gènes est presque supérieur à ce qui est attendu par hasard. Ensuite, la classification des gènes DE a montré que le nombre de gènes ayant une expression radicalement décalée ("up" et "down") entre les individus mésiques et xériques est plus importante que ce qui est attendu par hasard (effet marginalement significatif). Tout cela suggère qu'il existe un groupe de gènes dont l'expression doit être modifiée positivement ou négativement pour l'ensemble des espèces xériques.

Ces résultats sont encourageants mais nous pensons que nous ne sommes pas totalement équitables lors du test de significativité des valeurs observées. En effet, la méthode de tirage aléatoire ne nous semble pas, intuitivement, optimale : les scénarios aléatoires partagent des transitions présentes dans le scénario réel et peuvent donc contenir du signal de convergence. Cela pourrait augmenter le nombre de gènes DE et donc augmenter artificiellement la p-valeur. Nous aimerions prendre en compte cette distance entre le scénario réel et celui simulé dans notre test, mais nous n'avons pas encore de solution.

Nous réfléchissons aussi à utiliser la distribution des p-valeurs obtenues pour les gènes DE en complément de la distribution du nombre de gènes DE trouvés. En effet, le fait que l'on trouve de manière répétée des gènes "unclassified" dans les simulations suggère que les profils d'expression ne sont pas marqués et que donc les p-valeurs sont à la limite de significativité.

4.3.2.2 Des gènes DE sont liés à des fonctions biologiques pouvant être liées à l'adaptation aux milieux xériques

4.3.2.2.1 Des analyses d'ontologie de gènes (GO) révèlent de faibles enrichissements

De manière indépendante aux simulations pour savoir si le nombre de gènes DE obtenus était plus grand qu'attendu par hasard, nous avons regardé si on pouvait faire un lien entre les fonctions biologiques des gènes DE et l'adaptation des espèces aux milieux arides. Pour cela, nous avons utilisé à nouveau des analyses d'enrichissement en termes GO. La figure 4.14 résume l'ensemble des termes GO significativement enrichis dans chacun des groupes de gènes.

Dans ces termes GO enrichis, nous pouvons remarquer des termes qui pourraient être liés à une adaptation en milieu aride. Par exemple, dans les gènes "down", deux termes GO ressortent liés aux membranes et particulièrement au transport des lipides. On retrouve également des termes liés aux transports des ions dans les gènes "constraint" et dans l'ensemble du groupe de gènes DE, notamment incluant plusieurs termes liés à l'activité du transport transmembranaire (catégories "total" et "constraint"). Les termes liés aux processus apoptotiques (catégories "total" et "constraint") peuvent également avoir un lien avec l'adaptation à la vie aride, comme il a été montré et discuté par (MacManes, 2017), de même que celui concernant le réticulum endoplasmique. En effet, l'hyperosmolarité (qui apparaît lors d'une insuffisance en eau) peut entraîner l'arrêt du cycle cellulaire, l'interruption des mécanismes de réparation de l'ADN, l'inhibition de la transcription et de la traduction et l'inhibition du repliement protéique dans le réticulum endoplasmique. Chez les espèces adaptées à la vie désertique, les réponses face à l'hyperosmolarité seraient moins importantes. En effet, les organismes en milieu xérique subissent des périodes de stress hydrique prolongés. Afin de survivre à de plus longues périodes de pénurie, le seuil de tolérance à l'hyperosmolarité est plus grand, retardant ainsi le déclenchement de survie cellulaire. La catégorie de gènes "up" semble contenir des termes GO non identifiés dans de précédentes études. Ces termes ont pour la plupart un lien avec le cytosquelette, les filaments d'actine et cellules musculaires. Or, il a été montré que les filaments d'actine sont modifiés dans les cellules rénales musculaires lors d'ischémie (diminution de l'apport sanguin artériel, privant les cellules d'apport en oxygène, augmentation de la capacité de constriction) (Genescà et collab., 2006; Kwon et collab., 2002) , induisant de l'apoptose.

Ces analyses d'enrichissement en termes GO permettent d'orienter l'interprétation biologique des gènes DE. Mais cela ne dispense pas d'une analyse fine de la liste de gènes car certains gènes peuvent avoir un lien avec l'adaptation à l'aridité sans qu'il n'y ait d'enrichissement d'un groupe de gènes.

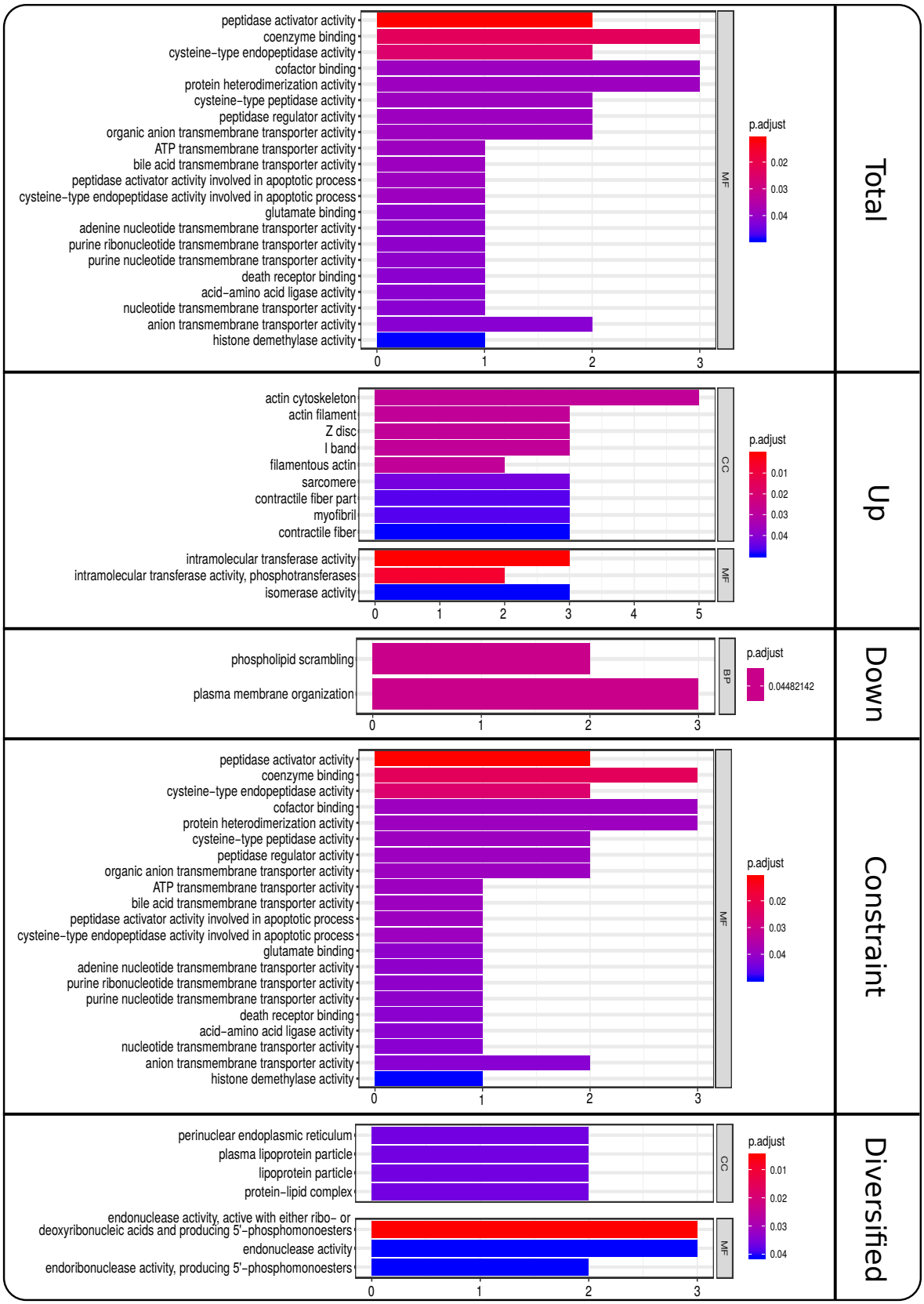

FIGURE 4.14 – Résumé des termes GO significatifs en fonction des groupes de gènes étudiés. L'ensemble des gènes du jeu de données "expressions" a été utilisé comme liste de gènes de référence. La couleur des barres représente la valeur de la p-valeur ajustée pour ces termes et la longueur de la barre le nombre de gènes associés à ce terme présent dans le groupe de gènes étudiés. MF : Fonctions Moléculaires, CC : Composants Cellulaires, BP : Processus Biologiques.

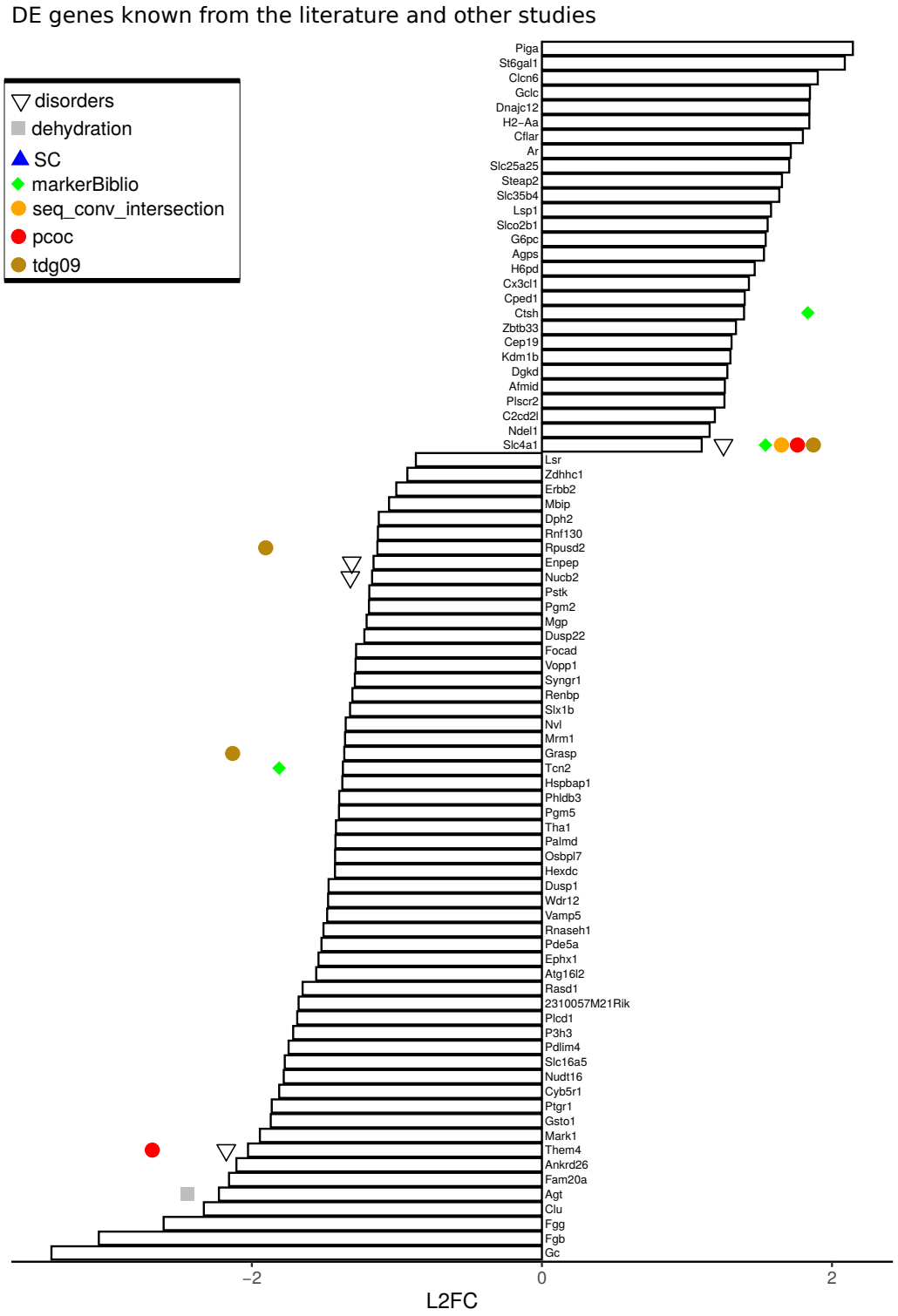

FIGURE 4.15 – Croisement entre les gènes DE et les gènes liés au rein dans la littérature. Les gènes DE sont ordonnés en fonction de leur Log2 Fold Change. Explication des listes croisées : "disorders" contient 186 gènes impliqués dans des maladies rénales d'après (PARK et collab., 2018), 106 gènes extraits de la banque OMIM (désordres génétiques associés au rein chez la souris) et 22 gènes extraits à l'aide du package PDB (phénotype HPO, ClinVar et variants uniprot) ; "dehydration" recense 464 gènes trouvés différentiellement exprimés suite à des expériences de déshydratation (MACMANES, 2017) ; "SC" contient 198 gènes spécifiques des types cellulaires rénaux identifiés par "single cell rna-seq analyses" (CAO et collab., 2018; PARK et collab., 2018) ; "markerBiblio" est une liste de 222 gènes recensés dans la bibliographie rénale ; "seq_conv_intersection" contient les 122 gènes trouvés à la fois par PCOC et Tdg09 ; "pcoc" contient 494 gènes trouvés par PCOC avec un seuil de 0.999 ; et finalement "tdg09" contient les 745 gènes identifiés par Tdg09 avec un seuil de 0.999.

4.3.2.2.2 Des gènes DE semblent biologiquement intéressants

Lorsque l'on croise les listes de gènes DE de notre analyse avec des listes de gènes identifiés dans la littérature comme ayant un lien avec des maladies du rein (PARK et collab., 2018), des variants génétiques ou simplement les gènes spécifiquement exprimés dans les différents types cellulaires rénaux, nous identifions dix gènes connus. Dans la catégorie des gènes "down", *Slc4a1* est impliqué dans les insuffisances rénales chroniques par (PARK et collab., 2018) et est également présent dans la bibliographie du rein (recueil réalisé au sein de l'équipe). *Slc4a1*, aussi appelé *AE1*, est un transporteur membranaire SLC présent au niveau des podocytes glomérulaires contribuant au bon fonctionnement de la barrière de filtration. *Ctsh* (Cathepsin H precursor) est également l'unique gène trouvé dans la catégorie "constraint" dont il y a une référence bibliographique liée au rein. Ce dernier est un marqueur du tubule proximal. La catégorie "up" est celle ayant le plus de gènes connus (*Agt*, *Them4*, *Tcn2*, *Grasp*, *Nucb2*, *Enpep*, *Rpusd2*). Parmi ces gènes, *Agt* pour Angiotensin, est différentiellement exprimé lors d'une sévère déshydratation. L'angiotensine est fondamentale pour l'équilibre rénal. Elle fait partie du système rénine angiotensine qui régule la pression sanguine et la balance des fluides et du sel dans le corps. Aucun gène spécifiquement connu n'appartient à la catégorie "diversified".

L'analyse d'enrichissement en termes GO, ainsi que le croisement des gènes DE avec les gènes connus dans la littérature suggèrent fortement que les gènes DE ont des fonctions biologiques liées à l'adaptation à l'aridité. On pourra procéder à une interprétation biologique plus poussée de ces résultats.

4.3.2.3 Le signal phylogénétique doit être pris en compte mais cela est encore difficile

Dans les analyses ci-dessus, nous n'avons pas pris l'effet phylogénétique en compte bien que les analyses multivariées ont montré la présence d'un fort effet phylogénétique dans les données, car DEseq2 ne le permet pas. En première approximation, nous avons, en parallèle de l'analyse présentée précédemment, considéré les familles de chacun des échantillons comme covariables lors de l'analyse d'expressions différentielles entre les environnements mésique et xérique afin de retirer l'effet de la famille. Mais nous ne pensons pas que ce soit la solution idéale car les deux effets "famille" et "écologie" sont ici confondus. En effet, certaines familles comprennent uniquement des espèces xériques ou uniquement des espèces mésiques.

Nous avons testé une autre alternative en nous plaçant au niveau de la famille. Pour cela nous avons mené des analyses spécifiques à l'échelle des trois familles les mieux échantillonnées (*muridae*, *cricetidae* et *heteromyidae*). Nous avons trouvé beaucoup plus de gènes DE (respectivement 1 274, 293 et 1 624) que lorsque l'on considérait l'ensemble des espèces et ceci semble cohérent avec de précédentes analyses (MARRA et collab., 2014). Cependant nous ne pouvons pas tester si ces valeurs sont plus qu'attendues par hasard. En effet, le nombre de combinaisons lors du mélange aléatoire des environnements est très faible (5 pour les *Muridae*, 3 pour les *Heteromyidae* et 15 pour les *Cricetidae*). Faire 1 000 simulations reviendrait à échantillonner de nombreuses fois les mêmes combinaisons et donc, nécessairement, la vraie combinaison réelle. On ne peut pas non plus échantillonner des espèces d'une autre famille car il faudrait prendre en compte la structure phylogénétique sous-jacente. 13 gènes DE sont en communs dans les trois analyses intra-familles.

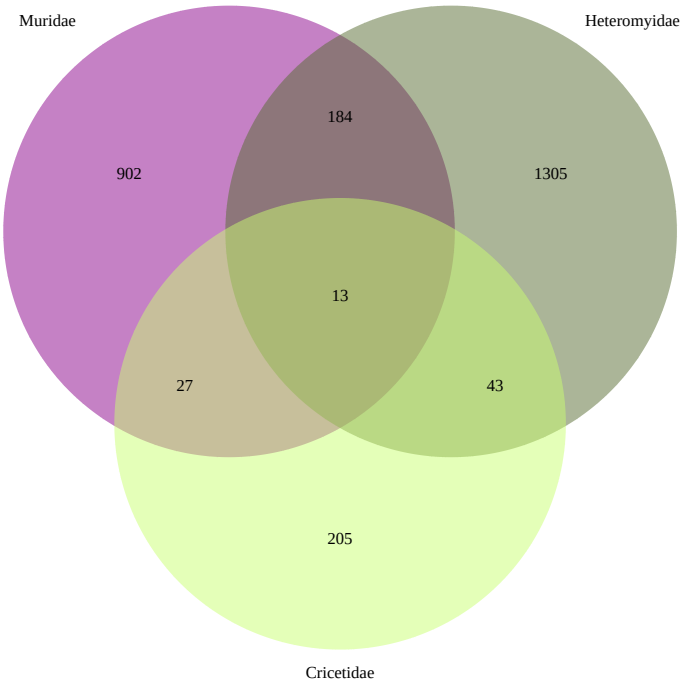

FIGURE 4.16 – Diagramme de Venn de l'intersection entre les listes de gènes DE pour les analyses d'expressions différentielles intra-familles taxonomiques.

Ces analyses à l'échelle intra-famille sont très intéressantes car elles permettraient de tester si l'on trouve plus de convergences à des distances phylogénétiques plus petites. Cependant, ces analyses intra-famille ne sont pas directement comparables entre elles et ni avec l'analyse globale qui considère l'ensemble des espèces car DEseq2 est très sensible au variation du nombre d'individus global et du nombre d'individus dans chacune des conditions analysées (ici, mésique contre xérique). En effet, la p-valeur pour un gène présentant un changement d'expression comparable entre les conditions mésique et xérique dans un jeu de données à 22 espèces sera bien inférieure que dans un jeu de données de 3 espèces.

Pour conclure, on devrait prendre en compte l'effet de la phylogénie dans l'analyse de gènes DE mais DESeq2 ne le permet pas et nous n'avons pas encore trouvé de manière satisfaisante pour le faire.

4.3.3 Comparaison des résultats provenant des analyses des niveaux d'expression et des séquences

4.3.3.1 *Slc4a1* est le seul gène présent à la fois dans les résultats des analyses des niveaux d'expression, de PCOC et de Tdg09

Slc4a1 est le seul gène qui possède des sites détectés comme convergents par Tdg09 et PCOC et qui a également été détecté comme DE. Trois sites ont été détectés convergents dans ce gène, le 47, le 488 et le 590 (Figure 4.17). Pour évaluer les effets de ces changements, il faudrait comparer ces sites aux sites fonctionnels de ce transporteur à anion et aux mutations humaines.

On peut voir que les sites détectés par PCOC et Tdg09 ne sont pas les mêmes. Le site 47 est détecté par Tdg09 alors que le site 590 par PCOC. Le site 488 est détecté par PCOC mais possède un bon score pour Tdg09. Lorsque l'on regarde de plus près la composition des sites, il semble que Tdg09 détecte des changements de profils d'acides aminés qui sont composés d'un petit nombre d'acides aminés tandis que PCOC détecte des changements de profils d'acides aminés qui sont composés d'un plus grand nombre d'acides aminés.

Par exemple pour le site 47, Tdg09 capte un changement de l'acide aminé majoritaire entre les espèces convergentes et non-convergentes de la lysine (K) vers l'arginine (R), deux acides aminés chargés positivement. Il existe deux possibilités pour expliquer que PCOC ne détecte pas ce site, soit le nombre de transitions effectivement convergentes est trop faible par rapport au nombre total de transitions ; soit les profils d'acides aminés prédéfinis dans PCOC ne sont pas capables de détecter ce genre de changement.

Pour le site 590, on voit qu'il a de nouveaux acides aminés dans les espèces convergentes (une lysine (K), une histidine (H), une alanine (A), une glycine (G), une thréonine (T)) qui ne sont pas présents dans les espèces non-convergentes majoritairement composées de sérines (S), asparagines (N) et d'arginines (R). PCOC détecte ce site car d'une part, les compositions en acides aminés sont très différentes (partie ProfileChange du modèle PCOC) et d'autre part il y a de nombreux changements au niveau des transitions (partie OneChange du modèle PCOC). Tdg09 ne détecte pas ce site car il ne peut pas définir *de novo* le profil d'acides aminés de ce site car par construction il écarte tous les acides aminés présents une seule fois.

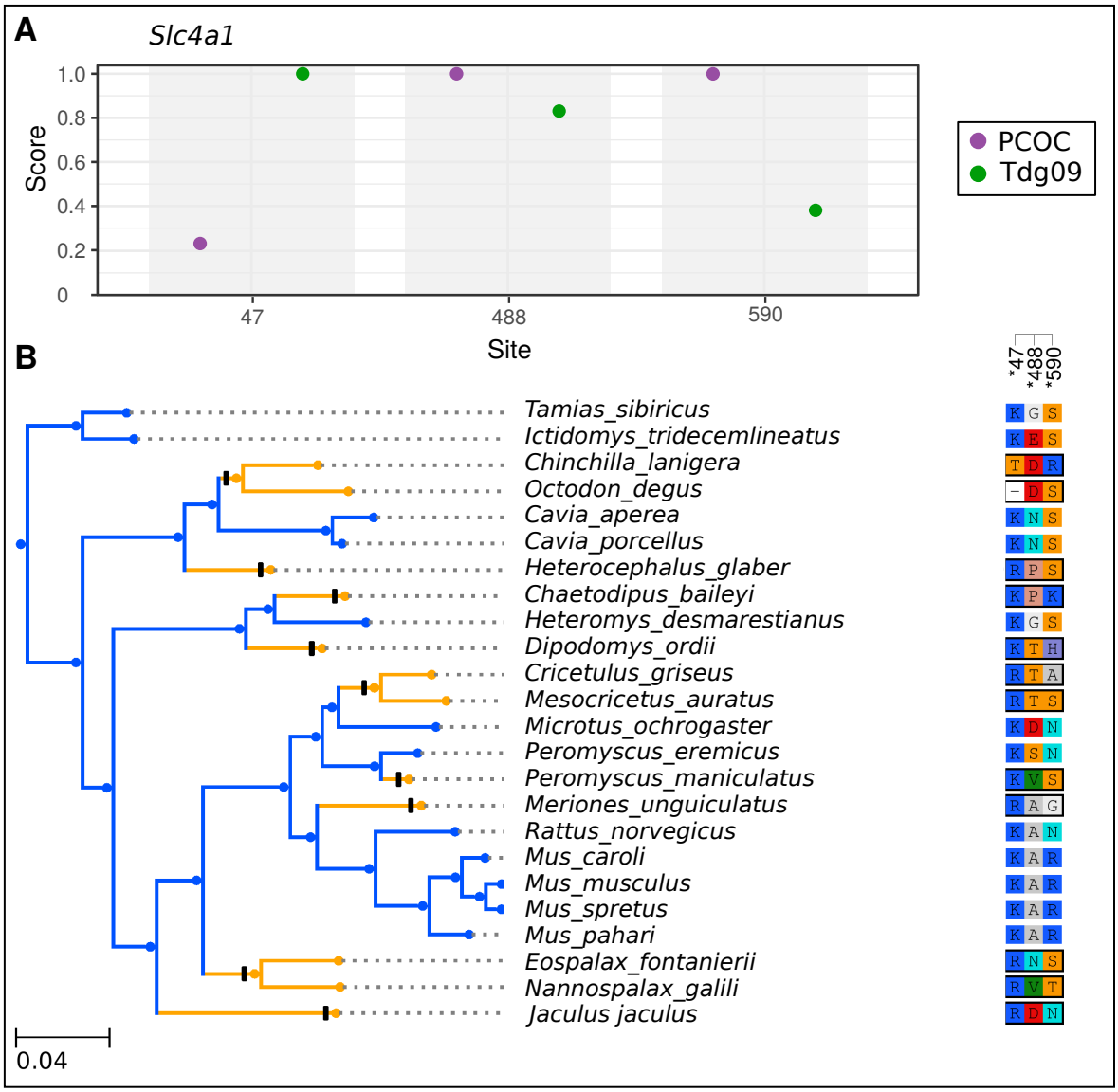

FIGURE 4.17 – Sites du gène *Slc4a1* détectés comme convergents par PCOC et Tdg09 avec des seuils de 0.999. (A) Score de chacune des méthodes pour les 3 sites détectés comme convergents. (B) Arbre des espèces pour lesquelles une séquence est disponible pour ce gène. Les espèces convergentes sont annotées en orange, les espèces non-convergentes en bleu et les transitions convergentes en noire. L'alignement protéique de ces 3 sites est disposé à droite de l'arbre.

L'analyse de ces deux sites valide bien que PCOC et Tdg09 ne détectent pas les mêmes sites comme nous avons discuté précédemment et donc explique bien la faible intersection entre les gènes trouvés par PCOC et par Tdg09.

4.3.3.2 La faible intersection entre les résultats des deux analyses (niveau d'expression et séquences) peut être un signe d'adaptations convergentes de natures différentes

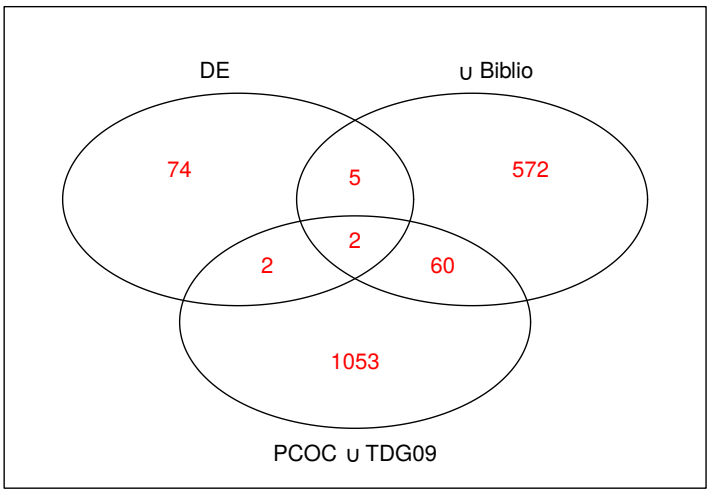

FIGURE 4.18 – Diagramme de Venn montrant l'intersection des résultats des analyses de séquences, des analyses des niveaux d'expression et de la bibliographie.

Il y a quatre gènes qui ont été détectés par PCOC et/ou Tdg09 et également par l'analyse de niveau d'expression. Deux d'entre eux sont connus dans la littérature pour avoir un lien avec le rein, *Slc4a1* et *Them4* et les deux autres sont *Grasp* et *Rpusd2*.

Les modifications des gènes différentiellement exprimés se situent au niveau des régions non codantes (Régions cis et trans) permettant ainsi une modification locale, pour un organe en particulier. En revanche, une modification au niveau de la séquence touche l'ensemble des cellules de l'organisme et donc tous les organes. On ne s'attend donc pas à avoir une intersection forte entre les gènes DE et les gènes avec de la convergence dans les séquences.

Deux hypothèses pourraient expliquer la présence de gènes dans les deux types de résultats. La première, dans le cas d'une convergence adaptative où, par exemple, une substitution avantageuse modifie la fonction de la protéine dont la surexpression est également avantageuse. La seconde possibilité est un relâchement de pression convergente, c'est à dire dans le cas d'une convergence non adaptative. Par exemple, la fonction d'un gène peut ne plus être importante, ce qui peut entraîner une diminution de son expression et une accumulation de substitutions dans sa séquence codante car le gène n'est plus sous sélection et donc n'a plus d'impact sur la fitness.

Par contre, si l'adaptation aux milieux arides des rongeurs a été favorisée par la modification convergente de processus biologiques, on pourrait s'attendre à retrouver de manière préférentielle les gènes issus des deux types d'analyses dans des réseaux de gènes communs.

4.3.4 Les gènes identifiés par les deux analyses pourraient être impliqués dans les mêmes processus biologiques

En parallèle des analyses d'ontologie, nous avons également regardé si les gènes détectés comme convergents par les deux approches appartenaient aux mêmes réseaux de gènes. Pour cela, nous avons utilisé String (SNEL, 2000) pour identifier les interactions connues entre les 112 gènes trouvés à la fois par PCOC et Tdg09 et les 83 gènes DE.

Parmi les 192 gènes présents dans la base de données de String, 115 gènes ont une interaction avec au moins un autre gène, c'est à dire que les deux gènes ont été associés à une même fonction (Figure 4.19, seuls les gènes avec une interaction sont présents). Au total, 179 associations ont été trouvées, ce qui est plus qu'attendu par hasard (p-valeur=1.32e-11). Bien que le test d'enrichissement soit significatif, il faut prendre ce résultat avec précaution car String, contrairement aux analyses d'enrichissement GO, ne prend pas en compte une liste de gènes d'arrière-plan. C'est à dire la liste de gènes contenus dans le jeu de données qui pourrait être initialement enrichie et donc fausser les résultats. String utilise l'ensemble des gènes présents dans le génome de la souris. Or, les gènes contenus dans nos jeux de données sont seulement les gènes exprimés dans le rein. On peut alors s'attendre à ce qu'ils soient en interaction. Il se peut donc que l'enrichissement que l'on observe est artefactuelle.

Cependant, l'analyse du réseau formé par l'ensemble des interactions met en évidence certains processus biologiques qui impliquent des gènes avec un signal de convergence. 8 gènes identifiés comme convergents sont impliqués dans les processus de "divisions cellulaires, condensation de la chromatine, mitose" dont la majorité proviennent de l'analyse de séquences (Ensemble de gènes entouré à gauche dans la Figure 4.19). Plusieurs gènes sont impliqués dans la "réparation de l'ADN" (*Parp9*, *Slx1b*, *Brca1*). On retrouve également des transporteurs à anions dont 8 présentant des interactions avec d'autres gènes convergents (*Lrp2*, *G6pc*, *Apoa4*, *Abcb11*, *Abca7*, *Slco2b1*, *Slc16a5*, *C2cd2l*) et 6 sans intéractions (*Slc35b4*, *Slc38a8*, *Slc4a1*, *Slc25a25*, *Plscr2*, *Atp10a*). Le processus "complement and coagulation cascades" est également bien représenté par 8 gènes (*Itgb2*, *Serpina5*, *Fgg*, i, *Fgb*, *A2m*, *Clu*, Mbl2). Parmi ces 8 gènes, les gènes Fibrinogen (*fgg*, *fgb* et *fga*) sont également impliqués dans la régulation de l'apoptose et les voies de signalisation MAPK et sont considérés comme biomarqueurs lors de transplantation rénale. Plusieurs gènes connus dans le rein semblent également interagir entre eux ou avoir de nombreuses connexions, comme par exemple *Agt*, *Lrp2*, *Enpep* et *Ctsh*.

Enfin, les gènes identifiés comme convergents par les analyses basées sur les séquences codantes et sur leur niveau d'expressions se retrouvent de manière homogène dans le réseau, c'est à dire que l'on ne retrouve pas de manière préférentielle les gènes issus de l'analyse de séquences dans une partie du réseau et les gènes issus de l'analyses DE dans une autre. L'ensemble des gènes montrant un signal de convergence intervient donc dans des processus biologiques communs. Il faudrait tester de manière rigoureuse si ces associations ne sont pas simplement dûes au hasard. Quoiqu'il en soit, ce résultat est encourageant.

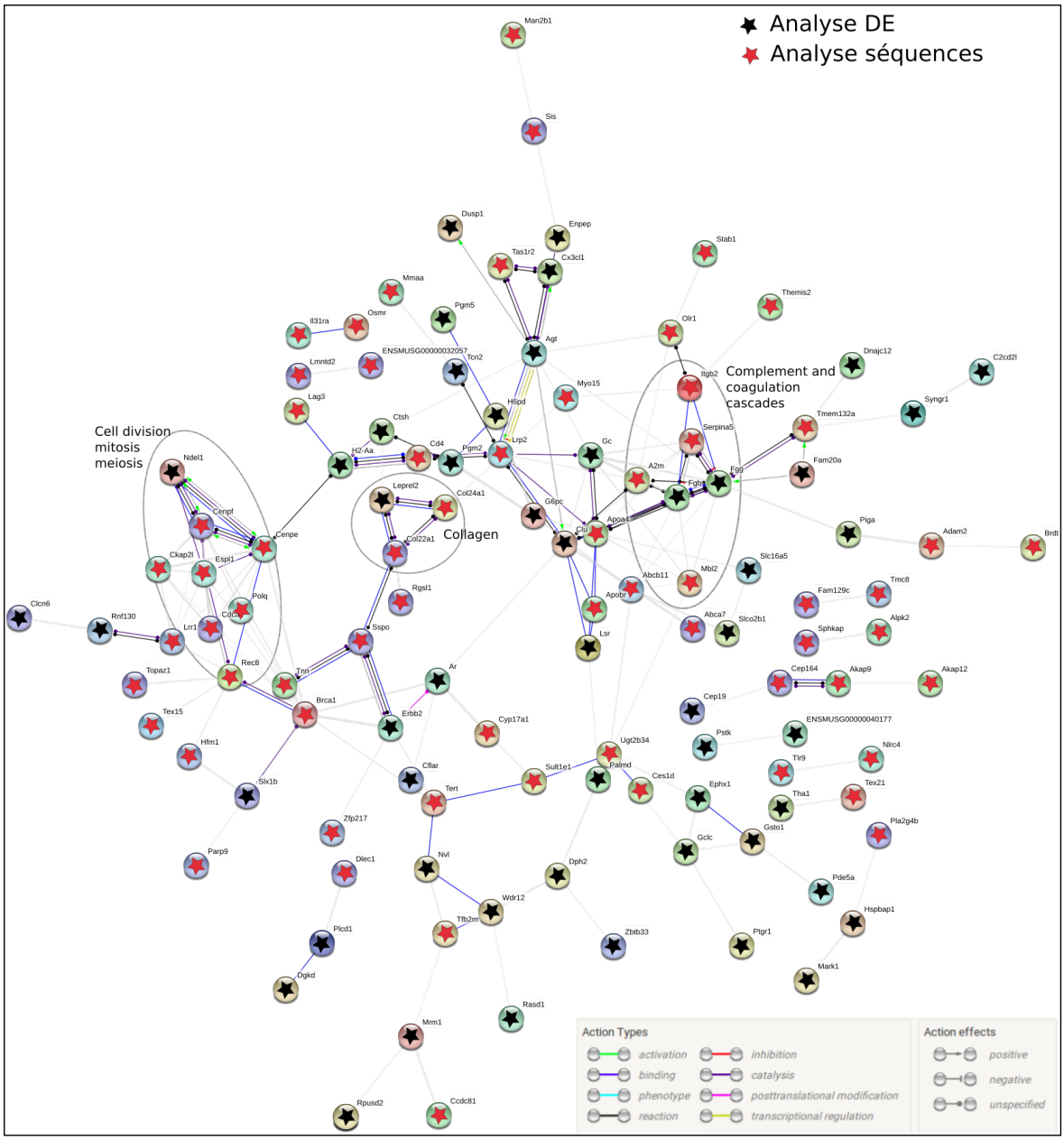

FIGURE 4.19 – Utilisation du logiciel String (SNEL, 2000) pour déterminer les interactions protéiques dans la liste de gènes contenant les 112 gènes trouvés par les analyses de séquences et les 83 gènes DE. Seuls sont représentés les gènes ayant une association avec une autre protéine. Les gènes avec une étoile noire sont des gènes DE alors que les gènes avec une étoile rouge proviennent du groupe de l'intersection des séquences PCOC et Tdg09.

4.4 Conclusions et perspectives

Le but de cette analyse était, d'une part, de se confronter à la réalité de données biologiques et, d'autre part, d'aborder la problématique de la quantité et du type de convergence présente lors de l'acquisition d'un phénotype convergent tel que l'adaptation à la vie aride chez les rongeurs.

1/ Tout d'abord, cette analyse sur la convergence nous a permis de montrer que la qualité du jeu de données est primordiale.

En effet, la définition des groupes d'orthologues est très importante car la présence de paralogues au sein du jeu de données peut induire un faux signal de convergence. Par exemple, nous avons des gènes écartés par phylter qui étaient détectés comme convergents par Tdg09 et PCOC

dans des résultats précédents (non-présentés ici).

De plus, il faut faire attention lors du nettoyage des données à conserver le signal de convergence. La convergence peut être interprétée par certains nettoyeurs d'alignements comme un problème d'orthologie entre les sites. Par exemple, dans une version précédente du pipeline de construction du jeu de données, nous utilisions Trimal (CAPELLA-GUTIERREZ et collab., 2009) plutôt que HMMcleaner pour le nettoyage de notre jeu de données. Lors de l'analyse de ce précédent jeu de données, nous ne détections aucun site convergent dans nos données. Nous avons compris la raison après avoir changé le nettoyeur de séquences dans notre pipeline. En effet, Trimal attribue un score d'homologie pour chacun des sites et supprime tous les sites avec un score inférieur à un seuil. Or, nous avions défini ce score à 0.8 ce qui écartait tous les sites qui portaient un signal de convergence (Figure 4.20).

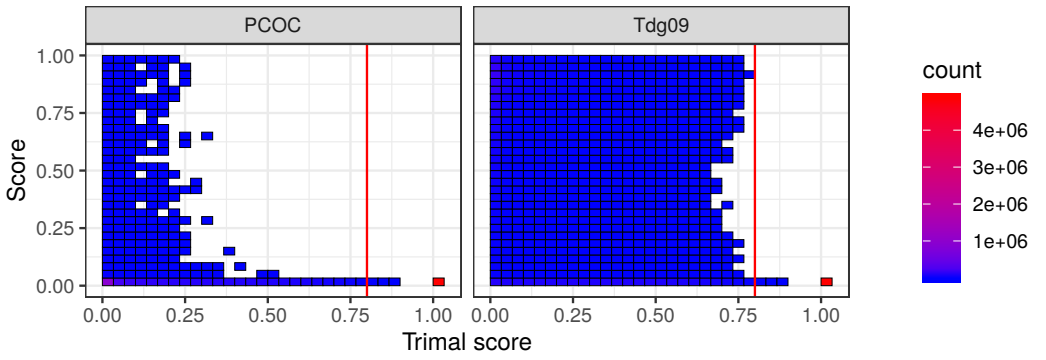

FIGURE 4.20 – Nombre de sites du jeu de données réel en fonction de leur score attribué par Trimal et leur score obtenu avec PCOC et Tdg09. La barre rouge indique le seuil à partir duquel un site avec un score inférieur était écarté par Trimal dans une précédente analyse.

2/ Nous nous sommes également rendu compte que l'interprétation des résultats de l'analyse de séquences est encore compliquée et ce pour trois raisons. Le grand nombre de gènes détectés convergents par PCOC et Tdg09 est intriguant et nécessite de nouvelles analyses. Est-ce qu'il s'agit d'un problème de définition du seuil ? Est-ce que c'est un problème de qualité du jeu de données ? Est-ce qu'il y a un processus biologique qui introduit de manière adaptative ou neutre de la convergence ?

D'autre part, il est difficile d'interpréter biologiquement des changements de profils d'acides aminés détectés par PCOC et Tdg09. En effet, les deux méthodes utilisent des profils d'acides aminés dont la définition est indépendante des propriétés physico chimique des acides aminés. J'ai implémenté la possibilité d'utiliser des profils d'acides aminés définis par l'utilisateur dans PCOC et il serait intéressant de tester cette nouvelle version avec des profils basés sur les connaissances physico-chimiques des acides aminés. Cela permettrait d'interpréter plus facilement les sites ressortant comme convergents.

Enfin, l'hétérogénéité du jeu de données "séquences" peut nous être défavorable. En effet, nous avons choisi de conserver toutes les familles de gènes avec au moins 15 espèces. Cela a permis de travailler sur un jeu de données possédant 14 554 gènes plutôt que 570 gènes si l'on avait considéré uniquement les familles de gènes avec les 30 espèces. Cependant, cela peut introduire du bruit dans les résultats. En effet, les scores des méthodes de détection ne sont pas faits pour tenir compte du nombre d'espèces ainsi que du nombre de transitions présentes. Un score de 0.7 pour un site présent dans un alignement de 30 espèces avec 9 transitions est différent d'un score de 0.7 pour un alignement avec 15 espèces et 4 transitions. Il faudrait étudier l'influence des ces deux paramètres sur les scores.

Finalement, lorsque ces problèmes auront été résolus, nous pourrons envisager d'utiliser diffsel sur les gènes identifiés par PCOC et Tdg09. Cela permettra de confirmer le signal de convergence contenu dans ces gènes et d'identifier les sites les plus prometteurs pour les tester fonctionnelle-

ment.

3/ Les résultats des analyses d'expression permettent en partie une interprétation biologique en lien avec la convergence mais il nous reste à prendre en compte le signal phylogénétique.

Tout d'abord, l'analyse de la variance contenue dans le jeu de données montre qu'il y a un signal de convergence entre les espèces xériques supérieur à ce qui est attendu par hasard et cela même si l'on retire l'effet phylogénétique.

Ensuite, les analyses d'expression ont mis en évidence des gènes différentiellement exprimés dans notre jeu de données complet à l'échelle des rongeurs, mais également sur des temps phylogénétiques moindre, à l'échelle des familles. Dans chacune des analyses, de nombreux gènes DE identifiés ont un lien connu avec le rein, que ce soit des marqueurs spécifiques (exemple : *Ctsh*), des gènes impliqués dans des maladies rénales (exemple : *Slc4a1*, *Them4*), ou bien des gènes mis en évidence par des expériences de déshydratation (*Agt*). Cependant, on ne trouve aucune Aquaporine, protéine fondamentale pour le bon fonctionnement du rein. De manière intéressante, plusieurs gènes ressortent lors des différents jeux et design utilisés pour l'expression (*Slc4a1*, *Pde5a*, *Them4*, *Dnajc12*).

Même si ces résultats sont encourageants, dans le cas du jeu de données complet, le signal phylogénétique est très important et dans le cas des analyses intra-familles, le nombre d'espèces et d'échantillons est probablement trop faible pour réaliser des tirages aléatoires significatifs. Pour essayer de s'affranchir de ce signal phylogénétique, on pourrait réaliser une analyse où l'on étudie uniquement des paires d'espèces convergentes et non-convergentes très apparentées. Ce design en paire a été utilisé par (YOUNG et collab., 2019) pour étudier l'apparition convergente de la monogamie dans cinq couples d'espèces allant de rongeurs jusqu'aux poissons en passant par les oiseaux et les grenouilles (Figure 4.21). Seulement cela nous obligerait à réduire de manière considérable notre jeu de données de 22 à 10 espèces.

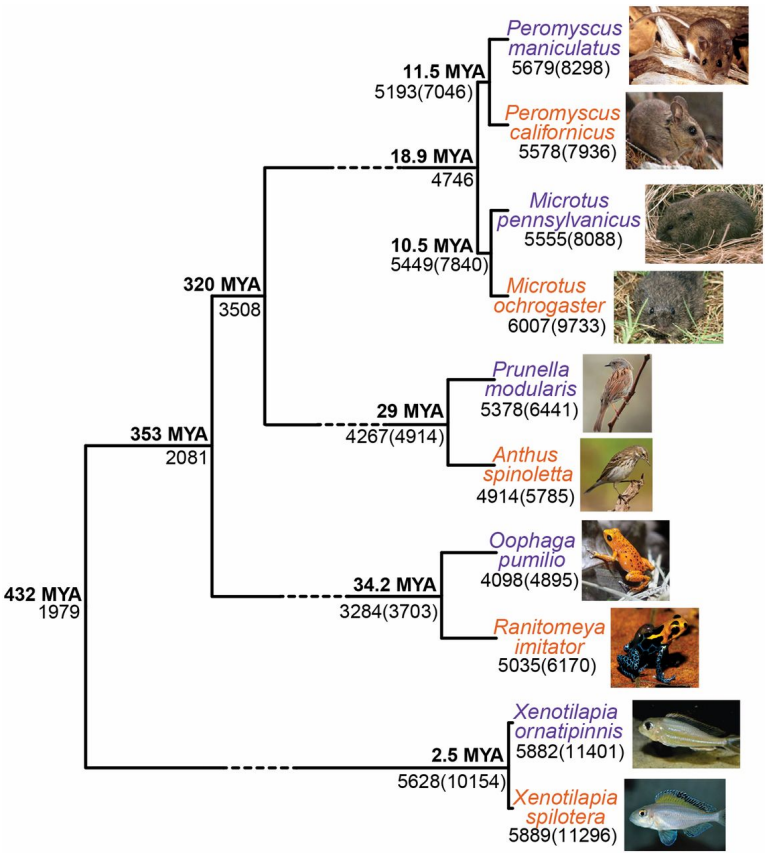

FIGURE 4.21 – Design en paire utilisé dans l'étude de la monogamie chez les vertébrés (YOUNG et collab., 2019).

De plus, un nouveau jeu de données est en construction. En effet, l'équipe a réalisé un échantillonnage afin de compléter certaines familles de rongeurs. Cet échantillonnage contient 17 espèces (dont trois souches de *Mus musculus* différentes), dont 14 nouvelles pouvant être ajoutées à l'analyse d'expression. Les échantillons de rein ont d'ores et déjà été séquencés et assemblés, mais pas encore analysés.

4/ Nous souhaitons également compléter cette analyse par une étude des taux d'évolution des gènes.

En effet, l'étude du taux d'évolution des gènes nous permettra de définir s'il s'agit d'une adaptation convergente comme chez les mammifères marins (CHIKINA et collab., 2016) ou bien un relâchement de la pression de sélection comme c'est le cas chez les animaux souterrains (PARTHA et collab., 2017) avec la perte des yeux par exemple. Nous envisageons d'utiliser RERconverge (KOWALCZYK et collab., 2019) qui est un nouvel outil pour étudier les différents taux d'évolution dans un contexte de convergence et également PAML (YANG, 1997) l'outil classique pour faire ce genre d'étude. Ce genre d'analyses a été utilisé dans le cadre d'études d'évolution convergente telles que l'étude de la convergence de l'espérance de vie chez le poisson clown (SAHM et collab., 2019) ou bien l'acquisition indépendante de la monogamie chez les vertébrés (YOUNG et collab., 2019).

Pour conclure, nous avons réussi à construire un jeu de données permettant d'analyser l'adaptation convergente à la vie en milieu aride chez les rongeurs. Les résultats préliminaires sur le jeu de données "expressions" soutiennent que l'on observe plus de convergence que ce qui est attendu par hasard et propose des gènes pouvant être liés à cette adaptation. Cependant, il soulève des interrogations et des difficultés techniques qui n'ont jamais été abordées. Les résultats préliminaires sur le jeu de données "séquences" reste plus difficile à interpréter mais semble contenir un signal de convergence. L'ensemble de ces analyses soutient des résultats qui vont dans le même sens et qui nous encouragent à approfondir nos analyses.

4.5 Matériels supplémentaires

TABLEAU 4.1 – Résumé détaillé des statistiques d'assemblages pour les 22 espèces de notre jeu de données pour lesquelles on a utilisé des données RNA-Seq.

Espèce	Référence	Nombre de reads utilisés (en millions)	N50	Taille moyenne des transcripts	Taille médiane des transcripts	% de gènes BUSCO complets retrouvés	Nombre de MGI
Abrothrix longipilis hirtus	Valdez et al. 2015	402.6M	3663	1342.95	432	86.5%	12770
Abrothrix olivaceus	Giorello et al. 2014	363M	3179	1152.21	403	87.9%	13133
Cavia porcellus	Fushan et al. 2015	101.2M	2594	1158.16	460	71.6%	11464
Chaetodipus baileyi	Marra et al. 2014	168M	2743	1137.68	428	84.0%	12573
Chinchilla lanigera	Broad Institute data	77 M	2515	980.42	377	82.2%	12672
Dipodomys spectabilis	Marra et al. 2014	32.6 M	2140	1052.60	477	69.1 %	10960
Eospalax fontanierii baileyi	Shao et al. 2015	43.6 M	2514	1295.99	630	70.7%	11882
Fukomys damarensis	Fang et al. 2014	139.6 M	1951	857.56	366	78.8%	13349
Fukomys mechowii	Fritz Lipmann Institute	100.4M	1560	912.53	481	76.5%	12775
Fukomys micklemi	Fritz Lipmann Institute	41.8 M	1482	897.71	488	75.5%	12900
Heterocephalus glaber	Bens et al. 2018	84.1M	1575	816.76	391	65.2%	11413
Heteromys desmarestianus	Marra et al. 2014	60.8M	2592	1182.89	482.5	78.4%	12272
Meriones unguiculatus	Fushan et al. 2015	120.2M	2690	1162.39	443	73.4%	11349
Mesocricetus auratus	Fushan et al. 2015	117.6M	2799	1111.10	394	78.3 %	12255
Mus caroli	EMBL-EBI	107.2M	1300	672.31	321	81.1%	12678
Mus musculus	Fushan et al. 2015	82.6M	2473	1135.07	468.5	67.7%	10838
Mus pahari	EMBL-EBI	63M	821	601.24	327	75.5%	12373
Myodes glareolus	Lund University	82.4M	2545	1058.98	419	78.5%	12462
Peromyscus eremicus	MacManes 2017	145.8M	1726	863.29	403	83.5%	13496
Peromyscus leucopus	Fushan et al. 2015	72.4M	2017	1004.19	462	62.5%	10915
Tamias sibiricus	Fushan et al. 2015	80M	2083	1006.13	446	66.7%	11159
Rattus norvegicus	Fushan et al. 2015	98.8M	2568	1206.07	513	71.2%	11294

5 Conclusion et perspectives générales

Sommaire

5.1 Conclusion et perspectives générales

L'objectif général de ma thèse était d'étudier la convergence au niveau génomique et plus particulièrement la quantité de convergence génomique pouvant être liée à l'acquisition d'un phénotype convergent via l'étude d'un jeu de données réel. Mais pour cela, j'ai dû au préalable mettre en place des outils bioinformatiques permettant de réaliser ce genre d'analyses. Au cours de ma thèse, j'ai tout d'abord réalisé des travaux basés sur des approches théoriques utilisant des données simulées puis des travaux basés sur une approche empirique utilisant des données réelles.

Dans un premier temps, j'ai implémenté CAARS, un outil permettant de construire des jeux de données de bonne qualité qui peuvent ensuite être utilisés dans des analyses de génomique comparative telles que les analyses de convergence. En plus d'apporter une solution intégrée pour la construction de jeux de données, CAARS permet l'annotation des paralogues dans les familles de gènes et propose un module facultatif d'affinage des assemblages.

J'ai ensuite créé un nouvel outil, PCOC, pour détecter la convergence présente dans les séquences codantes. Cet outil est basé sur la mise en œuvre de deux composantes qui nous paraissaient plus adaptées pour détecter la convergence que les méthodes qui étaient disponibles au début de ma thèse. Dans PCOC, un site est détecté convergent s'il présente, d'une part, un changement de profils d'acides aminés au niveau des transitions convergentes et, d'autre part, des substitutions au niveau de ces transitions convergentes. Cette dernière composante est l'une des particularités de PCOC qui lui permet d'identifier uniquement les sites présentant un signal de convergence particulièrement corrélé aux phénotypes convergents.

Nous avons également mis en évidence que des processus autres que la convergence adaptative pouvaient mener à la convergence génomique. C'est le cas des variations de tailles efficaces des populations des espèces convergentes. Or, ces processus peuvent avoir lieu de manière concomitante mais indépendante de l'adaptation convergente de ces espèces. Ces autres processus confondants doivent donc être pris en compte lors de l'interprétation des résultats car nous n'en connaissons pas les proportions.

Enfin, nous avons mis en place une analyse de la convergence génomique dans un jeu de données réel : l'étude de l'adaptation convergente à l'aridité chez les rongeurs. Les résultats sont encore préliminaires et ne nous permettent pas encore d'adresser la question de la quantité de convergence génomique associée au phénotype convergent étudié. Les premières interprétations biologiques des listes de gènes qui ont été identifiés comme convergents par les différentes méthodes (analyses d'expressions différentielles, Tdg09, PCOC) montrent un potentiel lien entre la fonction de ces gènes et l'adaptation à l'aridité. Cependant, la manière de définir ces listes n'est pas encore satisfaisante notamment à cause de la fixation des seuils pour Tdg09 et PCOC et la prise en compte du signal phylogénétique pour les analyses d'expressions différentielles. De plus, nous n'avons pas encore étudié les facteurs confondants potentiellement présents dans notre jeu de données pouvant induire de la convergence génomique.

Les résultats obtenus dans le cas de l'approche empirique ont mis en évidence des difficultés sur l'applicabilité des modèles développés sur données simulées dans un contexte réel. De plus, certaines observations suggèrent le besoin de développer de nouvelles fonctionnalités dans les modèles de détection de la convergence génomique pour aborder des questions biologiques que je développerai ensuite :

— Comment prendre en compte les facteurs confondants potentiels lors de la détection de convergence génomique associée au phénotype convergent étudié ?
— Comment prendre en compte de manière plus précise le phénotype convergent ?
— Comment quantifier la convergence qui n'est pas partagée par l'ensemble des espèces convergentes ?
— Quelle serait la meilleure manière de mener une analyse sur des nouvelles données réelles

pour identifier la convergence génomique liée à la convergence phénotypique ?

5.1.1 Comment prendre en compte les facteurs confondants potentiels lors de la détection de convergence génomique associée au phénotype convergent étudié ?

Dans notre étude de la convergence sur un jeu de données réel, nous suspectons la présence de facteurs confondants liés à des processus biologiques pouvant mimer de la convergence comme nous l'avons étudié dans le second chapitre. Cependant, nous n'avons pas encore les moyens de prendre en compte ces facteurs dans les outils actuels (PCOC, Tdg09, diffsel).

Il serait intéressant de pouvoir prendre en compte des covariables pour chacune des espèces telles que leur taux de GC. Une manière de faire celà serait de s'inspirer de la méthode multinomiale, qui est basée sur un test de χ^2, pour développer un modèle linéaire généralisé. Ce modèle devrait avoir des performances au moins équivalentes à la méthode multinomiale qui avait des performances légèrement inférieures aux méthodes actuelles. Cependant, il sera tout aussi rapide que la méthode multinomiale et donc beaucoup plus rapide que les méthodes actuelles. De plus, il sera facilement extensible. Par exemple, on pourra faire en sorte qu'il prenne en compte la structure phylogénétique. On pourra également l'étendre pour prendre en compte des facteurs confondants comme le taux de GC. Poursuivre le développement de cette méthode pourrait être une voie intéressante à suivre en parallèle des développements actuels.

5.1.2 Comment prendre en compte de manière plus précise le phénotype convergent ?

Dans les travaux théoriques, le fait qu'une espèce soit convergente est un paramètre fixé. Cependant, nous nous sommes rendu compte que la caractérisation de l'état convergent ou non-convergent des espèces d'un jeu de données empirique peut s'avérer plus compliquée pour deux raisons. La première est biologique. Il n'existe pas un milieu aride et un milieu mésique mais plutôt un continuum d'environnement de très aride à très humide auquel des espèces vont s'adapter. De plus, l'adaptation peut être polyfactorielle et il existe différentes stratégies pour y parvenir. Dans le cas de l'aridité, on a vu que des espèces pouvaient s'adapter en adoptant une vie nocturne et d'autres en fabriquant de l'eau métabolique à partir de lipides. La seconde raison est que l'on n'a pas accès à une mesure directe de l'adaptation, mais plutôt une approximation via une variable environnementale. Par exemple, dans notre jeu de données, nous avons utilisé la pluviométrie du trimestre le plus sec pour caractériser l'environnement des espèces. Celà peut être problématique pour l'annotation de l'environnement d'une espèce qui a accès à un point d'eau tel qu'un cours d'eau. En effet, nous nous attendrions à trouver une adaptation à un manque d'eau chez cette espèce alors que ce n'est pas le cas.

Concrètement, cette amélioration est difficilement envisageable dans PCOC car les modèles d'évolution sont appelés via un utilitaire intermédiaire de Bio++ et sont donc difficilement modifiables. Cependant, dans diffsel, qui est aussi développé dans le laboratoire, il est plus facilement envisageable d'implémenter cette modification. En effet, diffsel est un outil modulaire qui pourrait reprendre le modèle utilisé dans un autre programme Coevol (Lartillot et Poujol, 2010) qui permet d'étudier l'évolution de variables quantitatives le long d'une phylogénie. Diffsel est également basé sur la détection d'un changement de profils d'acides aminés entre les espèces non-convergentes et les espèces convergentes mais dans un cadre bayésien. On pourrait, par exemple, réutiliser des modules de Coevol dans diffsel afin de permettre la détection d'un changement de profils d'acides aminés le long d'un continuum. Par exemple, les espèces porteuses des phénotypes les plus extrêmes permettront de définir des profils d'acides aminés extrêmes et les espèces porteuses d'un phénotype intermédiaire se verront attribuer un profil d'acides aminés intermédiaire calculé via une transformation linéaire.

Dans un modèle linéaire généralisé, il est également possible d'implémenter cette amélioration en considérant le phénotype convergent comme continu plutôt que binaire. Un autre problème que

nous avons rencontré est le placement des transitions convergentes dans l'arbre d'espèces. En effet, il n'est pas toujours facile d'orienter les changements phénotypiques dans la phylogénie. On aimerait pouvoir identifier une espèce qui pourrait représenter une réversion. En effet, on ne s'attend pas à ce que cette espèce suive le comportement des espèces convergentes ou non-convergentes. On s'attend donc à ce qu'elle ait un régime particulier. Il serait peut-être plus prudent pour le moment de les enlever du jeu de données. Nous pourrions imaginer des modèles ne prenant en compte que l'état aux feuilles et qui pourraient inférer les états ancestraux. Cela pourrait permettre de prendre en compte les potentielles réversions. Ces modèles ont déjà été implémentés dans la littérature (Chapitre 7, (HARMON, 2019)) et ont été utilisés par la méthode msd qui intègre justement cette caractéristique. Cependant, ce serait très chronophage car le modèle devrait calculer toutes les possibilités pour chacun des noeuds ancestraux, ce qui augmenterait de manière considérable le temps de calcul global.

5.1.3 Comment quantifier la convergence qui n'est pas partagée par l'ensemble des espèces convergentes ?

Les résultats préliminaires obtenus sur le jeu de données rongeurs ne semblent pas contenir des gènes globalement convergents tels que la prestine avec plusieurs substitutions retrouvées dans l'ensemble des espèces convergentes et qui sont fonctionnellement liés à un phénotype convergent. Cela peut s'expliquer par le fait que le phénotype que l'on étudie concerne un plus grand nombre de transitions. En effet, en admettant que nos espèces soient correctement étiquetées et qu'elles soient également adaptées ; plus on augmente le nombre d'espèces, moins on s'attend à trouver le même changement convergent à un site partagé par toutes les espèces. Sinon, cela serait vraiment exceptionnel. Il s'agirait du gène incontournable permettant l'adaptation à l'aridité. En revanche, on s'attend à trouver des sites partagés par des sous-ensembles, d'une part, à cause des degrés divers d'adaptation et, d'autre part, à cause des erreurs d'étiquettes. On voudrait donc être capable de détecter de la convergence présente dans un sous-ensemble d'espèces.

Cette fonctionnalité permettrait d'étudier si des sous-ensembles d'espèces convergentes partagent une plus grande quantité de convergence que d'autres sous-ensembles d'espèces convergentes. En effet, il est possible que les espèces avec les phénotypes les plus extrêmes (dans notre cas, les espèces vivant dans les milieux les plus arides) partagent une plus grande quantité de sites convergents que l'ensemble des espèces convergentes. On peut également se demander si la quantité de convergence est liée aux distances phylogénétiques entre les espèces convergentes, c'est à dire qu'il est plus facile d'obtenir de la convergence entre les espèces phylogénétiquement proches. Ou bien encore, il est possible que des sous-ensembles d'espèces partagent plus de convergences indépendamment de la phylogénie ou du phénotype. Par exemple, il est possible qu'il existe deux chemins pour s'adapter à un environnement et les espèces convergentes ont pris par hasard l'un des chemins sans lien avec leur proximité phylogénétique ou de leur phénotype. Cette fonctionnalité permettra donc de savoir s'il y a plusieurs manières de s'adapter à un environnement et, si oui, s'il y a des chemins préférentiels en lien avec la phylogénie, l'intensité du phénotype ou bien une autre variable.

Notre jeu de données sur l'adaptation à la vie en milieu aride chez les rongeurs convient très bien pour répondre à ces questions. En effet, il comprend neuf transitions convergentes, contre seulement deux ou trois pour les études précédentes. Cependant, pour tester cette hypothèse, il faut être capable d'intégrer dans les outils un moyen de détecter des sites convergents dans un sous ensemble des transitions convergentes et d'avoir un moyen d'extraire cette information. En effet, nous ne pouvons pas tester toutes les combinaisons possibles, car la combinatoire est trop grande.

Nous envisageons d'inclure cette capacité dans PCOC en ajoutant une fonctionnalité qui permet de tester la contribution de chacune des transitions au modèle convergent global. Cela nous permettra ensuite de pondérer les transitions en fonction de leur contribution au modèle global et ainsi de définir un site convergent en fonction des transitions robustes.

5.1.4 Quelle serait la meilleure manière de mener une analyse sur des nouvelles données réelles pour identifier la convergence génomique liée à la convergence phénotypique ?

Finalement, ces travaux vont être probablement suivis par d'autres analyses étudiant de nouveaux phénotypes convergents. Les résultats de ces différentes analyses vont permettre à terme de proposer une réponse globale sur la quantité de convergence attendue dans la nature. Nous avons montré au travers des discussions des différents chapitres que des développements méthodologiques étaient encore nécessaires pour améliorer les méthodes actuelles, notamment lors de la définition du seuil de détection de chacune des méthodes de détection de la convergence. Dès qu'ils seront achevés et validés sur ce jeu de données test, ils pourront être mis en place sur de nouveaux jeux de données. Au vu des mes travaux, j'aimerais apporter différents conseils pour ces analyses futures sur le choix du phénotype, sur la construction du jeu de données et sur les analyses à réaliser.

Tout d'abord, il faut étudier un phénotype convergent pour lequel il existe des hypothèses sur les adaptations fonctionnelles attendues et donc potentiellement sur les gènes ciblés. Par exemple, l'adaptation à la vie en haute altitude permet de faire face à une quantité d'oxygène limitée. Il ne faut pas tout de suite s'attaquer à des phénotypes trop complexes tels que l'acquisition des ailes où la convergence peut ne pas avoir de bases phénotypiques profondes. On pourra peut-être l'envisager dans un second temps. De plus, il faut essayer d'inclure une espèce modèle dans l'échantillonnage des espèces étudiées pour profiter à la fois de l'annotation du génome qui est normalement de bonne qualité sur le plan de l'annotation génomique (définition des transcripts) et sur le plan fonctionnel (fonction des gènes). De plus, il sera plus facile de tester fonctionnellement les gènes candidats potentiellement identifiés sur un organisme modèle pour lequel des outils de biologie moléculaire seront disponibles.

Ensuite, il faut envisager de construire un jeu de données avec un nombre important d'espèces. En effet, plus il y a d'espèces, moins il y a de chances d'observer des processus corrélés au phénotype convergent par hasard. Dans la même idée, il faut essayer d'avoir un maximum de transitions convergentes possibles.

Lors de la construction du jeu de données, il faut se questionner sur l'influence de la présence de convergence dans chacune des étapes, c'est à dire est-ce que le fait que le jeu de données puisse contenir de la convergence peut perturber le fonctionnement d'un outil. Et inversement, il faut également envisager l'impact de chacune des étapes sur notre capacité à détecter la convergence. Est-ce que l'outil peut supprimer le signal de convergence ?

Enfin, lors de la détection de la convergence génomique, il faut dans un premier temps définir les seuils de détection des sites convergents adaptés au jeu de données étudié. La méthode reste à être mise au point à l'aide de notre jeu de données test mais elle reposera sûrement sur des simulations. Ensuite, je pense qu'il faut utiliser dans un premier temps les méthodes les plus rapides et sensibles telles que PCOC et Tdg09 pour identifier des gènes candidats. Puis, lorsque nous aurons implémenté la prise en compte de cofacteurs dans un modèle linéaire généralisé, nous pourrons dans un second temps l'utiliser en parallèle de PCOC et Tdg09 pour étudier et quantifier les proportions de convergence associées à chacun des facteurs confondants.

Cette approche pourra nous apporter des réponses sur la quantité de convergence présente dans la nature. Par contre, pour identifier les sites fonctionnellement liés au phénotype convergent, il faudra également utiliser diffsel qui est plus lent mais plus spécifique sur les gènes précédemment identifiés. Cela permettra d'identifier avec plus de certitude les sites pouvant être liés fonctionnellement au phénotype convergent et ainsi restreindre la liste aux gènes et aux sites les plus prometteurs. Les sites passant ce dernier test pourront ensuite être testés en laboratoire. Ce genre de test est très long et très coûteux, on ne peut donc tester qu'un petit nombre de sites.

Pour finir, nous avons pris le chemin de l'étude de la convergence à la suite des travaux de

(PARKER et collab., 2013). Ces travaux, bien que discutables et discutés, sont les précurseurs de l'étude de la convergence à l'échelle génomique. Ils ont engendré le développement de nouvelles méthodes et d'études théoriques sur les bases biologiques de la convergence génomique dont mes travaux de thèse font partie. Mais vu l'attrait de ce champ de la biologie, je pense que le sujet sera étudié encore pour longtemps.

A Annexes

Sommaire

A.1 Collaborations

A.1.1 Assemblages de transcriptomes et analyse d'expression différentielle

Au cours de ma thèse, j'ai également collaboré avec Thibault Lorin pour l'un de ses projets de thèse. Ce travail a débouché sur un article publié en janvier 2019 dans PIGMENT CELL & MELANOMA Research (SALIS et collab. (2019), https://doi.org/10.1111/pcmr.12766).

Au cours de ce projet, Thibault Lorin et ses collaborateurs s'intéressaient à la caractérisation des pigments responsables de la coloration en orange et blanc chez le poisson clown (*Amphiprion ocellaris*).

Dans ce projet, j'ai aidé Thibault Lorin à réaliser les assemblages des transcriptomes de peau blanche et orange. Je l'ai également conseillé avec Marie Sémon pour la réalisation de l'analyse de l'expression différentielle des gènes entre ces deux tissus.

A.1.2 Assemblages ciblés de gènes et analyses d'épissages alternatifs

J'ai également collaboré avec Marie Fablet. Ce travail a débouché sur un article publié en mars 2019 dans G3 : Genes, Genomes, Genetics (FABLET et collab. (2019), https://doi.org/10.1534/g3.118.200789).

Dans cet article, Marie Fablet et ses collaborateurs s'intéressent à l'effet de l'insertion d'un élément transposable, (*Tirant*), dans deux gènes chez une espèce de drosophile (*Drosophila simulans*).

Cette analyse avait pour but d'étudier l'effet de cette insertion sur le niveau d'expression de ces gènes en comparant différentes populations de *Drosophila simulans* où l'élément transposable était présent ou absent.

Ma contribution a été d'analyser le profil d'épissage alternatif de ces gènes en présence ou en l'absence de l'élément transposable. J'ai pour cela utilisé Apytram qui est l'un des constituants de CAARS (cf section 2.1.3).

Liste complète des références

ALHAJERI, B. H. et S. J. STEPPAN. 2018, «Community structure in ecological assemblages of desert rodents», *Biological Journal of the Linnean Society*, vol. 124, nº 3, doi :10.1093/biolinnean/bly068, p. 308–318. URL https://doi.org/10.1093/biolinnean/bly068. 96

ARENDT, J. et D. REZNICK. 2008, «Convergence and parallelism reconsidered : what have we learned about the genetics of adaptation ?», *Trends in Ecology & Evolution*, vol. 23, nº 1, doi : 10.1016/j.tree.2007.09.011, p. 26–32. URL https://doi.org/10.1016/j.tree.2007.09.011. 36

ASHIHARA, H. 2004, «Distribution and biosynthesis of caffeine in plants», *Frontiers in Bioscience*, vol. 9, nº 1-3, doi :10.2741/1367, p. 1864. URL https://doi.org/10.2741/F1367. 3, 4

BANKIR, L. et C. DE ROUFFIGNAC. 1985, «Urinary concentrating ability : insights from comparative anatomy», *American Journal of Physiology-Regulatory, Integrative and Comparative Physiology*, vol. 249, nº 6, doi :10.1152/ajpregu.1985.249.6.r643, p. R643–R666. URL https://doi.org/10.1152/ajpregu.1985.249.6.r643. 96

BLOOM, J. D. 2016, «Identification of positive selection in genes is greatly improved by using experimentally informed site-specific models», doi :10.1101/037689. URL https://doi.org/10.1101/037689. 78

BOLGER, A. M., M. LOHSE et B. USADEL. 2014, «Trimmomatic : a flexible trimmer for illumina sequence data», *Bioinformatics*, vol. 30, nº 15, doi :10.1093/bioinformatics/btu170, p. 2114–2120. URL https://doi.org/10.1093/bioinformatics/btu170. 98

BRAY, N. L., H. PIMENTEL, P. MELSTED et L. PACHTER. 2016, «Near-optimal probabilistic RNA-seq quantification», *Nature Biotechnology*, vol. 34, nº 5, doi :10.1038/nbt.3519, p. 525–527. URL https://doi.org/10.1038/nbt.3519. 99

CAMACHO, C., G. COULOURIS, V. AVAGYAN, N. MA, J. PAPADOPOULOS, K. BEALER et T. L. MADDEN. 2009, «BLAST+ : architecture and applications», *BMC Bioinformatics*, vol. 10, nº 1, doi :10.1186/1471-2105-10-421, p. 421. URL https://doi.org/10.1186/1471-2105-10-421. 99

CAO, J., D. A. CUSANOVICH, V. RAMANI, D. AGHAMIRZAIE, H. A. PLINER, A. J. HILL, R. M. DAZA, J. L. MCFALINE-FIGUEROA, J. S. PACKER, L. CHRISTIANSEN, F. J. STEEMERS, A. C. ADEY, C. TRAPNELL et J. SHENDURE. 2018, «Joint profiling of chromatin accessibility and gene expression in thousands of single cells», *Science*, vol. 361, nº 6409, doi :10.1126/science.aau0730, p. 1380–1385. URL https://doi.org/10.1126/science.aau0730. 96, 118

CAPELLA-GUTIERREZ, S., J. M. SILLA-MARTINEZ et T. GABALDON. 2009, «trimAl : a tool for automated alignment trimming in large-scale phylogenetic analyses», *Bioinformatics*, vol. 25, nº 15, doi :10.1093/bioinformatics/btp348, p. 1972–1973. URL https://doi.org/10.1093/bioinformatics/btp348. 125

CHABROL, O., M. ROYER-CARENZI, P. PONTAROTTI et G. DIDIER. 2018, «Detecting the molecular basis of phenotypic convergence», *Methods in Ecology and Evolution*, vol. 9, nº 11, doi : 10.1111/2041-210x.13071, p. 2170–2180. URL https://doi.org/10.1111/2041-210x.13071. 78

CHAN, Y. F., M. E. MARKS, F. C. JONES, G. VILLARREAL, M. D. SHAPIRO, S. D. BRADY, A. M. SOUTHWICK, D. M. ABSHER, J. GRIMWOOD, J. SCHMUTZ, R. M. MYERS, D. PETROV, B. JONSSON, D. SCHLUTER, M. A. BELL et D. M. KINGSLEY. 2009, «Adaptive evolution of pelvic reduction in sticklebacks by recurrent deletion of a pitx1 enhancer», *Science*, vol. 327, nº 5963, doi : 10.1126/science.1182213, p. 302–305. URL https://doi.org/10.1126/science.1182213. 35

CHIKINA, M., J. D. ROBINSON et N. L. CLARK. 2016, «Hundreds of genes experienced convergent shifts in selective pressure in marine mammals», *Molecular Biology and Evolution*, vol. 33, nº 9, doi :10.1093/molbev/msw112, p. 2182–2192. URL https://doi.org/10.1093/molbev/msw112. 127

CLARKE, K., Y. YANG, R. MARSH, L. XIE et Z. K. K. 2013, «Comparative analysis of de novo transcriptome assembly», *Science China Life Sciences*, vol. 56, nº 2, doi :10.1007/s11427-013-4444-x, p. 156–162. URL https://doi.org/10.1007/s11427-013-4444-x. 43

CONSORTIUM, G. O. 2018, «The gene ontology resource : 20 years and still going strong», *Nucleic acids research*, vol. 47, nº D1, doi :10.1093/nar/gky1055, p. D330–D338. URL https://doi.org/10.1093/nar/gky1055. 110

COUVILLON, M. J., H. A. TOUFAILIA, T. M. BUTTERFIELD, F. SCHRELL, F. L. RATNIEKS et R. SCHÜRCH. 2015, «Caffeinated forage tricks honeybees into increasing foraging and recruitment behaviors», *Current Biology*, vol. 25, nº 21, doi : 10.1016/j.cub.2015.08.052, p. 2815–2818. URL https://doi.org/10.1016/j.cub.2015.08.052. 20

DOUZERY, E. J. P., C. SCORNAVACCA, J. ROMIGUIER, K. BELKHIR, N. GALTIER, F. DELSUC et V. RANWEZ. 2014, «OrthoMaM v8 : A database of orthologous exons and coding sequences for comparative genomics in mammals», *Molecular Biology and Evolution*, vol. 31, nº 7, doi :10.1093/molbev/msu132, p. 1923–1928. URL https://doi.org/10.1093/molbev/msu132. 41

DRAY, S., A. B. DUFOUR et D. CHESSEL. 2007, «The ade4 package-ii : Two-table and k-table methods», *R news*, vol. 7, nº 2, p. 47–52. 104, 112

DU, K., L. YANG et S. HE. 2015, «Phylogenomic analyses reveal a molecular signature linked to subterranean adaptation in rodents», *BMC Evolutionary Biology*, vol. 15, nº 1, doi :10.1186/s12862-015-0564-1. URL https://doi.org/10.1186/s12862-015-0564-1. 96

DURET, L. et N. GALTIER. 2009, «Biased gene conversion and the evolution of mammalian genomic landscapes», *Annual Review of Genomics and Human Genetics*, vol. 10, nº 1, doi :10.1146/annurev-genom-082908-150001, p. 285–311. URL https://doi.org/10.1146/annurev-genom-082908-150001. 35

EDWARDS, S. V., L. LIU et D. K. PEARL. 2007, «High-resolution species trees without concatenation», *Proceedings of the National Academy of Sciences*, vol. 104, nº 14, doi :10.1073/pnas.0607004104, p. 5936–5941. URL https://doi.org/10.1073/pnas.0607004104. 91

EWELS, P., M. MAGNUSSON, S. LUNDIN et M. KÄLLER. 2016, «MultiQC : summarize analysis results for multiple tools and samples in a single report», *Bioinformatics*, vol. 32, nº 19, doi : 10.1093/bioinformatics/btw354, p. 3047–3048. URL https://doi.org/10.1093/bioinformatics/btw354. 98

FABLET, M., A. JACQUET, R. REBOLLO, A. HAUDRY, C. REY, J. SALCES-ORTIZ, P. BAJAD, N. BURLET, M. F. JANTSCH, M. P. G. GUERREIRO et C. VIEIRA. 2019, «Dynamic interactions between the genome and an endogenous retrovirus : Tirant in drosophilasimulans wild-type strains», *G3 : Genes, Genomes, Genetics*, doi :10.1534/g3.118.200789, p. g3.200 789.2018. URL https://doi.org/10.1534/g3.118.200789. 137

FABRE, P.-H., L. HAUTIER, D. DIMITROV et E. J. P. DOUZERY. 2012, «A glimpse on the pattern of rodent diversification : a phylogenetic approach», *BMC Evolutionary Biology*, vol. 12, nº 1, doi : 10.1186/1471-2148-12-88, p. 88. URL https://doi.org/10.1186/1471-2148-12-88. 96, 100, 101

FOOTE, A. D., Y. LIU, G. W. C. THOMAS, T. VINAŘ, J. ALFÖLDI, J. DENG, S. DUGAN, C. E. VAN ELK, M. E. HUNTER, V. JOSHI, Z. KHAN, C. KOVAR, S. L. LEE, K. LINDBLAD-TOH, A. MANCIA, R. NIELSEN, X. QIN, J. QU, B. J. RANEY, N. VIJAY, J. B. W. WOLF, M. W. HAHN, D. M. MUZNY, K. C. WORLEY, M. T. P. GILBERT et R. A. GIBBS. 2015, «Convergent evolution of the genomes of marine mammals», *Nature Genetics*, vol. 47, nº 3, doi :10.1038/ng.3198, p. 272–275. URL https://doi.org/10.1038/ng.3198. 6, 61, 95

FRANCO, A. D., R. POUJOL, D. BAURAIN et H. PHILIPPE. 2019, «Evaluating the usefulness of alignment filtering methods to reduce the impact of errors on evolutionary inferences», *BMC Evolutionary Biology*, vol. 19, nº 1, doi :10.1186/s12862-019-1350-2. URL https://doi.org/10.1186/s12862-019-1350-2. 99

FRIEDMAN, M., K. SHIMADA, L. D. MARTIN, M. J. EVERHART, J. LISTON, A. MALTESE et M. TRIEBOLD. 2010, «100-million-year dynasty of giant planktivorous bony fishes in the mesozoic seas», *Science*, vol. 327, nº 5968, doi :10.1126/science.1184743, p. 990–993. URL https://doi.org/10.1126/science.1184743. 18

GENESCÀ, M., A. SOLA et G. HOTTER. 2006, «Actin cytoskeleton derangement induces apoptosis in renal ischemia/reperfusion», *Apoptosis*, vol. 11, nº 4, doi :10.1007/s10495-006-4937-1, p. 563–571. URL https://doi.org/10.1007/s10495-006-4937-1. 116

GIORELLO, F. M., M. FEIJOO, G. D'ELÍA, D. E. NAYA, L. VALDEZ, J. C. OPAZO et E. P. LESSA. 2018, «An association between differential expression and genetic divergence in the patagonian olive mouse (abrothrix olivacea)», *Molecular Ecology*, vol. 27, nº 16, doi :10.1111/mec.14778, p. 3274–3286. URL https://doi.org/10.1111/mec.14778. 96

GIORELLO, F. M., M. FEIJOO, G. D'ELÍA, L. VALDEZ, J. C. OPAZO, V. VARAS, D. E. NAYA et E. P. LESSA. 2014, «Characterization of the kidney transcriptome of the south american olive mouse abrothrix olivacea», *BMC Genomics*, vol. 15, nº 1, doi :10.1186/1471-2164-15-446, p. 446. URL https://doi.org/10.1186/1471-2164-15-446. 96

GOLDBOGEN, J., D. CADE, J. CALAMBOKIDIS, A. FRIEDLAENDER, J. POTVIN, P. SEGRE et A. WERTH. 2017, «How baleen whales feed : The biomechanics of engulfment and filtration», *Annual Review of Marine Science*, vol. 9, nº 1, doi :10.1146/annurev-marine-122414-033905, p. 367–386. URL https://doi.org/10.1146/annurev-marine-122414-033905. 18

GOMPEL, N. et B. PRUD'HOMME. 2009, «The causes of repeated genetic evolution», *Developmental Biology*, vol. 332, nº 1, doi : 10.1016/j.ydbio.2009.04.040, p. 36–47. URL https://doi.org/10.1016/j.ydbio.2009.04.040. 36

GOULD, S. J. 1990, *Wonderful life : the Burgess Shale and the nature of history*, WW Norton & Company. 14

GRABHERR, M. G., B. J. HAAS, M. YASSOUR, J. Z. LEVIN, D. A. THOMPSON, I. AMIT, X. ADICONIS, L. FAN, R. RAYCHOWDHURY, Q. ZENG, Z. CHEN, E. MAUCELI, N. HACOHEN, A. GNIRKE, N. RHIND, F. DI PALMA, B. W. BIRREN, C. NUSBAUM, K. LINDBLAD-TOH, N. FRIEDMAN et A. REGEV. 2011, «Full-length transcriptome assembly from RNA-seq data without a reference genome», *Nature Biotechnology*, vol. 29, nº 7, doi :10.1038/nbt.1883, p. 644–652. URL https://doi.org/10.1038/nbt.1883. 43, 98

GUÉGUEN, L., S. GAILLARD, B. BOUSSAU, M. GOUY, M. GROUSSIN, N. C. ROCHETTE, T. BIGOT, D. FOURNIER, F. POUYET, V. CAHAIS, A. BERNARD, C. SCORNAVACCA, B. NABHOLZ, A. HAUDRY, L. DACHARY, N. GALTIER, K. BELKHIR et J. Y. DUTHEIL. 2013, «Bio++ : Efficient extensible libraries and tools for computational molecular evolution», *Molecular Biology and Evolution*, vol. 30, nº 8, doi :10.1093/molbev/mst097, p. 1745–1750. URL https://doi.org/10.1093/molbev/mst097. 66

HAAS, B. J., A. PAPANICOLAOU, M. YASSOUR, M. GRABHERR, P. D. BLOOD, J. BOWDEN, M. B. COUGER, D. ECCLES, B. LI, M. LIEBER, M. D. MACMANES, M. OTT, J. ORVIS, N. POCHET, F. STROZZI, N. WEEKS, R. WESTERMAN, T. WILLIAM, C. N. DEWEY, R. HENSCHEL, R. D. LEDUC, N. FRIEDMAN et A. REGEV. 2013, «De novo transcript sequence reconstruction from RNA-seq using the trinity platform for reference generation and analysis», *Nature Protocols*, vol. 8, nº 8, doi :10.1038/nprot.2013.084, p. 1494–1512. URL https://doi.org/10.1038/nprot.2013.084. 98

HARMON, L. 2019, «Phylogenetic comparative methods : Learning from trees», doi :10.32942/osf.io/e3xnr. URL https://doi.org/10.32942/osf.io/e3xnr. 132

HELED, J. et A. J. DRUMMOND. 2009, «Bayesian inference of species trees from multilocus data», *Molecular Biology and Evolution*, vol. 27, nº 3, doi :10.1093/molbev/msp274, p. 570–580. URL https://doi.org/10.1093/molbev/msp274. 91

HERRERO, J., M. MUFFATO, K. BEAL, S. FITZGERALD, L. GORDON, M. PIGNATELLI, A. J. VILELLA, S. M. J. SEARLE, R. AMODE, S. BRENT, W. SPOONER, E. KULESHA, A. YATES et P. FLICEK. 2016, «Ensembl comparative genomics resources», *Database*, vol. 2016, doi :10.1093/database/bav096, p. bav096. URL https://doi.org/10.1093/database/bav096. 41

HEWAVITHARANAGE, P., S. KARUNARATNE et N. KUMAR. 1999, «Effect of caffeine on shot-hole borer beetle (xyleborusfornicatus) of tea (camellia sinensis)», *Phytochemistry*, vol. 51, nº 1, doi : 10.1016/s0031-9422(98)00610-4, p. 35–41. URL https://doi.org/10.1016/s0031-9422(98)00610-4. 20

HOEKSTRA, H. E., R. J. HIRSCHMANN, R. A. BUNDEY, P. A. INSEL et J. P. CROSSLAND. 2006, «A single amino acid mutation contributes to adaptive beach mouse color pattern», *Science*, vol. 313, nº 5783, doi :10.1126/science.1126121, p. 101–104. URL https://doi.org/10.1126/science.1126121. 29

HOLLINGSWORTH, R. G., J. W. ARMSTRONG et E. CAMPBELL. 2002, «Caffeine as a repellent for slugs and snails», *Nature*, vol. 417, nº 6892, doi :10.1038/417915a, p. 915–916. URL https://doi.org/10.1038/417915a. 20

HUANG, R., A. J. O'DONNELL, J. J. BARBOLINE et T. J. BARKMAN. 2016, «Convergent evolution of caffeine in plants by co-option of exapted ancestral enzymes», *Proceedings of the National Academy of Sciences*, vol. 113, nº 38, doi :10.1073/pnas.1602575113, p. 10 613–10 618. URL https://doi.org/10.1073/pnas.1602575113. 23, 24, 25, 26

HUERTA-CEPAS, J., D. SZKLARCZYK, K. FORSLUND, H. COOK, D. HELLER, M. C. WALTER, T. RATTEI, D. R. MENDE, S. SUNAGAWA, M. KUHN, L. J. JENSEN, C. VON MERING et P. BORK. 2015, «eggNOG 4.5 : a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences», *Nucleic Acids Research*, vol. 44, nº D1, doi :10.1093/nar/gkv1248, p. D286–D293. URL https://doi.org/10.1093/nar/gkv1248. 41, 99

HÖLZER, M. et M. MARZ. 2019, «De novo transcriptome assembly : A comprehensive cross-species comparison of short-read RNA-seq assemblers», *GigaScience*, vol. 8, nº 5, doi :10.1093/gigascience/giz039. URL https://doi.org/10.1093/gigascience/giz039. 43

IRWIN, D. M., T. D. KOCHER et A. C. WILSON. 1991, «Evolution of the cytochromeb gene of mammals», *Journal of Molecular Evolution*, vol. 32, nº 2, doi :10.1007/bf02515385, p. 128–144. URL https://doi.org/10.1007/bf02515385. 12

ISHIKAWA, M., I. YUYAMA, H. SHIMIZU, M. NOZAWA, K. IKEO et T. GOJOBORI. 2016, «Different endosymbiotic interactions in two hydra species reflect the evolutionary history of endosymbiosis», *Genome Biology and Evolution*, vol. 8, nº 7, doi :10.1093/gbe/evw142, p. 2155–2163. URL https://doi.org/10.1093/gbe/evw142. 44

JOHNSON, K. P., K. K. WALDEN et H. M. ROBERTSON. 2013, «Next-generation phylogenomics using a target restricted assembly method», *Molecular Phylogenetics and Evolution*, vol. 66, nº 1, doi :10.1016/j.ympev.2012.09.007, p. 417–422. URL https://doi.org/10.1016/j.ympev.2012.09.007. 45

JOHNSON, R. J., P. STENVINKEL, T. JENSEN, M. A. LANASPA, C. RONCAL, Z. SONG, L. BANKIR et L. G. SÁNCHEZ-LOZADA. 2016, «Metabolic and kidney diseases in the setting of climate change, water shortage, and survival factors», *Journal of the American Society of Nephrology*, vol. 27, nº 8, doi :10.1681/asn.2015121314, p. 2247–2256. URL https://doi.org/10.1681/asn.2015121314. 96

KATOH, K. et D. M. STANDLEY. 2013, «MAFFT multiple sequence alignment software version 7 : Improvements in performance and usability», *Molecular Biology and Evolution*, vol. 30, nº 4, doi :10.1093/molbev/mst010, p. 772–780. URL https://doi.org/10.1093/molbev/mst010. 99

KELLEY, N. P. et N. D. PYENSON. 2015, «Evolutionary innovation and ecology in marine tetrapods from the triassic to the anthropocene», *Science*, vol. 348, nº 6232, doi :10.1126/science.aaa3716, p. aaa3716–aaa3716. URL https://doi.org/10.1126/science.aaa3716. 17

KIM, C.-S. et D.-M. SHIN. 2016, «Improper hydration induces global gene expression changes associated with renal development in infant mice», *Genes & Nutrition*, vol. 11, nº 1, doi :10.1186/s12263-016-0544-0. URL https://doi.org/10.1186/s12263-016-0544-0. 96

KORDONOWY, L., K. D. LOMBARDO, H. L. GREEN, M. D. DAWSON, E. A. BOLTON, S. LACOURSE et M. D. MACMANES. 2017, «Physiological and biochemical changes associated with acute experimental dehydration in the desert adapted mouse, peromyscus eremicus», *Physiological Reports*, vol. 5, nº 6, doi :10.14814/phy2.13218, p. e13 218. URL https://doi.org/10.14814/phy2.13218. 96

KOWALCZYK, A., W. K. MEYER, R. PARTHA, W. MAO, N. L. CLARK et M. CHIKINA. 2019, «RERconverge : an r package for associating evolutionary rates with convergent traits», *Bioinformatics*, doi :10.1093/bioinformatics/btz468. URL https://doi.org/10.1093/bioinformatics/btz468. 127

KOZLOV, A. M., D. DARRIBA, T. FLOURI, B. MOREL et A. STAMATAKIS. 2019, «RAxML-NG : a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference», *Bioinformatics*, doi :10.1093/bioinformatics/btz305. URL https://doi.org/10.1093/bioinformatics/btz305. 99

KUMAR, S., G. STECHER, M. SULESKI et S. B. HEDGES. 2017, «TimeTree : A resource for timelines, timetrees, and divergence times», *Molecular Biology and Evolution*, vol. 34, nº 7, doi :10.1093/molbev/msx116, p. 1812–1819. URL https://doi.org/10.1093/molbev/msx116. 6

KWON, O., C. L. PHILLIPS et B. A. MOLITORIS. 2002, «Ischemia induces alterations in actin filaments in renal vascular smooth muscle cells», *American Journal of Physiology-Renal Physiology*, vol. 282, nº 6, doi :10.1152/ajprenal.00294.2001, p. F1012–F1019. URL https://doi.org/10.1152/ajprenal.00294.2001. 116

LAND, M. F. et D.-E. NILSSON. 2012, *Animal eyes*, Oxford University Press. 18, 19

LARTER, M., A. DUNBAR-WALLIS, A. E. BERARDI et S. D. SMITH. 2018, «Convergent evolution at the pathway level : Predictable regulatory changes during flower color transitions», *Molecular Biology and Evolution*, vol. 35, nº 9, doi :10.1093/molbev/msy117, p. 2159–2169. URL https://doi.org/10.1093/molbev/msy117. 22

LARTILLOT, N. et H. PHILIPPE. 2004, «A bayesian mixture model for across-site heterogeneities in the amino-acid replacement process», *Molecular Biology and Evolution*, vol. 21, nº 6, doi :10.1093/molbev/msh112, p. 1095–1109. URL https://doi.org/10.1093/molbev/msh112. 65

LARTILLOT, N. et R. POUJOL. 2010, «A phylogenetic model for investigating correlated evolution of substitution rates and continuous phenotypic characters», *Molecular Biology and Evolution*, vol. 28, nº 1, doi :10.1093/molbev/msq244, p. 729–744. URL https://doi.org/10.1093/molbev/msq244. 131

LE, S. Q., N. LARTILLOT et O. GASCUEL. 2008, «Phylogenetic mixture models for proteins», *Philosophical Transactions of the Royal Society B : Biological Sciences*, vol. 363, nº 1512, doi :10.1098/rstb.2008.0180, p. 3965–3976. URL https://doi.org/10.1098/rstb.2008.0180. 65

LI, E. et Y. ZHANG. 2014, «DNA methylation in mammals», *Cold Spring Harbor Perspectives in Biology*, vol. 6, nº 5, doi :10.1101/cshperspect.a019133, p. a019 133–a019 133. URL https://doi.org/10.1101/cshperspect.a019133. 35

LI, Y., Z. LIU, P. SHI et J. ZHANG. 2010, «The hearing gene prestin unites echolocating bats and whales», *Current Biology*, vol. 20, nº 2, doi :10.1016/j.cub.2009.11.042, p. R55–R56. URL https://doi.org/10.1016/j.cub.2009.11.042. 30, 31

LIU, Z., F.-Y. QI, X. ZHOU, H.-Q. REN et P. SHI. 2014, «Parallel sites implicate functional convergence of the hearing gene prestin among echolocating mammals», *Molecular Biology and Evolution*, vol. 31, nº 9, doi :10.1093/molbev/msu194, p. 2415–2424. URL https://doi.org/10.1093/molbev/msu194. 30, 31

LOSOS, J. 2017, *Improbable Destinies : How Predictable is Evolution ?*, Penguin UK. 5

LOVE, M. I., W. HUBER et S. ANDERS. 2014, «Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2», *Genome Biology*, vol. 15, nº 12, doi :10.1186/s13059-014-0550-8. URL https://doi.org/10.1186/s13059-014-0550-8. 103, 112

MACMANES, M. D. 2017, «Severe acute dehydration in a desert rodent elicits a transcriptional response that effectively prevents kidney injury», *American Journal of Physiology-Renal Physiology*, vol. 313, nº 2, doi :10.1152/ajprenal.00067.2017, p. F262–F272. URL https://doi.org/10.1152/ajprenal.00067.2017. 96, 116, 118

MacManes, M. D. et M. B. Eisen. 2014, «Characterization of the transcriptome, nucleotide sequence polymorphism, and natural selection in the desert adapted mousePeromyscus eremicus», *PeerJ*, vol. 2, doi :10.7717/peerj.642, p. e642. URL https://doi.org/10.7717/peerj.642. 96

Manceau, M., V. S. Domingues, C. R. Linnen, E. B. Rosenblum et H. E. Hoekstra. 2010, «Convergence in pigmentation at multiple levels : mutations, genes and function», *Philosophical Transactions of the Royal Society B : Biological Sciences*, vol. 365, nº 1552, doi :10.1098/rstb.2010.0104, p. 2439–2450. URL https://doi.org/10.1098/rstb.2010.0104. 27, 28, 29, 96

Maor, R., T. Dayan, H. Ferguson-Gow et K. E. Jones. 2017, «Temporal niche expansion in mammals from a nocturnal ancestor after dinosaur extinction», *Nature Ecology & Evolution*, vol. 1, nº 12, doi :10.1038/s41559-017-0366-5, p. 1889–1895. URL https://doi.org/10.1038/s41559-017-0366-5. 8

Marra, N. J., S. H. Eo, M. C. Hale, P. M. Waser et J. A. DeWoody. 2012, «A priori and a posteriori approaches for finding genes of evolutionary interest in non-model species : Osmoregulatory genes in the kidney transcriptome of the desert rodent dipodomys spectabilis (banner-tailed kangaroo rat)», *Comparative Biochemistry and Physiology Part D : Genomics and Proteomics*, vol. 7, nº 4, doi :10.1016/j.cbd.2012.07.001, p. 328–339. URL https://doi.org/10.1016/j.cbd.2012.07.001. 96, 103

Marra, N. J., A. Romero et J. A. DeWoody. 2014, «Natural selection and the genetic basis of osmoregulation in heteromyid rodents as revealed by RNA-seq», *Molecular Ecology*, vol. 23, nº 11, doi :10.1111/mec.12764, p. 2699–2711. URL https://doi.org/10.1111/mec.12764. 44, 96, 103, 119

Mathavan, S., Y. Premalatha et M. Christopher. 1985, «Effects of caffeine and theophylline on the fecundity of four lepidopteran species.», *Experimental biology*, vol. 44, nº 2, p. 133–138. 20

Morris, S. C. 1999, *The crucible of creation : the Burgess Shale and the rise of animals*, Peterson's. 14

Motta, P. J., M. Maslanka, R. E. Hueter, R. L. Davis, R. de la Parra, S. L. Mulvany, M. L. Habegger, J. A. Strother, K. R. Mara, J. M. Gardiner, J. P. Tyminski et L. D. Zeigler. 2010, «Feeding anatomy, filter-feeding rate, and diet of whale sharks rhincodon typus during surface ram filter feeding off the yucatan peninsula, mexico», *Zoology*, vol. 113, nº 4, doi :10.1016/j.zool.2009.12.001, p. 199–212. URL https://doi.org/10.1016/j.zool.2009.12.001. 18

Mundy, N. I. 2009, «Conservation and convergence of colour genetics : MC1r mutations in brown cavefish», *PLoS Genetics*, vol. 5, nº 2, doi :10.1371/journal.pgen.1000388, p. e1000388. URL https://doi.org/10.1371/journal.pgen.1000388. 28

Nathanson, J. 1984, «Caffeine and related methylxanthines : possible naturally occurring pesticides», *Science*, vol. 226, nº 4671, doi :10.1126/science.6207592, p. 184–187. URL https://doi.org/10.1126/science.6207592. 20

Nowak, R. M. et E. P. Walker. 1999, *Walker's Mammals of the World*, vol. 1, JHU press. 96

O'Leary, M. A., J. I. Bloch, J. J. Flynn, T. J. Gaudin, A. Giallombardo, N. P. Giannini, S. L. Goldberg, B. P. Kraatz, Z.-X. Luo, J. Meng, X. Ni, M. J. Novacek, F. A. Perini, Z. Randall, G. W. Rougier, E. J. Sargis, M. T. Silcox, N. B. Simmons, M. Spaulding, P. M. Velazco, M. Weksler, J. R. Wible et A. L. Cirranello. 2013, «Response to comment on "the placental mammal ancestor and the post-k-pg radiation of placentals"», *Science*, vol. 341, nº 6146, doi :10.1126/science.1238162, p. 613–613. URL https://doi.org/10.1126/science.1238162. 8

Park, J., R. Shrestha, C. Qiu, A. Kondo, S. Huang, M. Werth, M. Li, J. Barasch et K. Suszták. 2018, «Single-cell transcriptomics of the mouse kidney reveals potential cellular targets of kidney disease», *Science*, vol. 360, nº 6390, doi :10.1126/science.aar2131, p. 758–763. URL https://doi.org/10.1126/science.aar2131. 96, 118, 119

Parker, J., G. Tsagkogeorga, J. A. Cotton, Y. Liu, P. Provero, E. Stupka et S. J. Rossiter. 2013, «Genome-wide signatures of convergent evolution in echolocating mammals», *Nature*, vol. 502, nº 7470, doi :10.1038/nature12511, p. 228–231. URL https://doi.org/10.1038/nature12511. 61, 95, 134

Partha, R., B. K. Chauhan, Z. Ferreira, J. D. Robinson, K. Lathrop, K. K. Nischal, M. Chikina et N. L. Clark. 2017, «Subterranean mammals show convergent regression in ocular genes and enhancers, along with adaptation to tunneling», *eLife*, vol. 6, doi :10.7554/elife.25884. URL https://doi.org/10.7554/elife.25884. 127

Parto, S. et N. Lartillot. 2018, «Correction : Molecular adaptation in rubisco : Discriminating between convergent evolution and positive selection using mechanistic and classical codon models», *PLOS ONE*, vol. 13, nº 4, doi :10.1371/journal.pone.0196267, p. e0196267. URL https://doi.org/10.1371/journal.pone.0196267. 78

Pereira, R. J., F. S. Barreto, N. T. Pierce, M. Carneiro et R. S. Burton. 2016, «Transcriptome-wide patterns of divergence during allopatric evolution», *Molecular Ecology*, vol. 25, nº 7, doi :10.1111/mec.13579, p. 1478–1493. URL https://doi.org/10.1111/mec.13579. 44

Pessia, E., A. Popa, S. Mousset, C. Rezvoy, L. Duret et G. A. B. Marais. 2012, «Evidence for widespread GC-biased gene conversion in eukaryotes», *Genome Biology and Evolution*, vol. 4, nº 7, doi :10.1093/gbe/evs052, p. 675–682. URL https://doi.org/10.1093/gbe/evs052. 35, 91

Powell, S., K. Forslund, D. Szklarczyk, K. Trachana, A. Roth, J. Huerta-Cepas, T. Gabaldón, T. Rattei, C. Creevey, M. Kuhn, L. J. Jensen, C. von Mering et P. Bork. 2013, «eggNOG v4.0 : nested orthology inference across 3686 organisms», *Nucleic Acids Research*, vol. 42, nº D1, doi :10.1093/nar/gkt1253, p. D231–D239. URL https://doi.org/10.1093/nar/gkt1253. 99

Pradervand, S., A. Z. Mercier, G. Centeno, O. Bonny et D. Firsov. 2010, «A comprehensive analysis of gene expression profiles in distal parts of the mouse renal tubule», *Pflügers Archiv - European Journal of Physiology*, vol. 460, nº 6, doi :10.1007/s00424-010-0863-8, p. 925–952. URL https://doi.org/10.1007/s00424-010-0863-8. 96

Revell, L. J. 2011, «phytools : an r package for phylogenetic comparative biology (and other things)», *Methods in Ecology and Evolution*, vol. 3, nº 2, doi :10.1111/j.2041-210x.2011.00169.x, p. 217–223. URL https://doi.org/10.1111/j.2041-210x.2011.00169.x. 100

Romero, A. 2012, «When whales became mammals : The scientific journey of cetaceans from fish to mammals in the history of science», dans *New Approaches to the Study of Marine Mammals*, InTech, doi :10.5772/50811. URL https://doi.org/10.5772/50811. 6

Romiguier, J., V. Ranwez, E. J. P. Douzery et N. Galtier. 2010, «Contrasting GC-content dynamics across 33 mammalian genomes : Relationship with life-history traits and chromosome sizes», *Genome Research*, vol. 20, nº 8, doi :10.1101/gr.104372.109, p. 1001–1009. URL https://doi.org/10.1101/gr.104372.109. 36

Rosenblum, E. B., H. E. Hoekstra et M. W. Nachman. 2004, «Adaptive reptile color variation and the evolution of the mcir gene», *Evolution*, vol. 58, nº 8, p. 1794–1808. 28

ROSENBLUM, E. B., C. E. PARENT et E. E. BRANDT. 2014, «The molecular basis of phenotypic convergence», *Annual Review of Ecology, Evolution, and Systematics*, vol. 45, nº 1, doi :10.1146/annurev-ecolsys-120213-091851, p. 203–226. URL https://doi.org/10.1146/annurev-ecolsys-120213-091851. 36

ROSENBLUM, E. B., H. ROMPLER, T. SCHONEBERG et H. E. HOEKSTRA. 2009, «Molecular and functional basis of phenotypic convergence in white lizards at white sands», *Proceedings of the National Academy of Sciences*, vol. 107, nº 5, doi :10.1073/pnas.0911042107, p. 2113–2117. URL https://doi.org/10.1073/pnas.0911042107. 27, 28, 29

RÖMPLER, H., N. ROHLAND, C. LALUEZA-FOX, E. WILLERSLEV, T. KUZNETSOVA, G. RABEDER, J. BERTRANPETIT, T. SCHÖNEBERG et M. HOFREITER. 2006, «Nuclear gene indicates coat-color polymorphism in mammoths», *Science*, vol. 313, nº 5783, doi :10.1126/science.1128994, p. 62–62. URL https://doi.org/10.1126/science.1128994. 27, 28, 29

SAHM, A., P. ALMAIDA-PAGÁN, M. BENS, M. MUTALIPASSI, A. LUCAS-SÁNCHEZ, J. DE COSTA RUIZ, M. GÖRLACH et A. CELLERINO. 2019, «Analysis of the coding sequences of clownfish reveals molecular convergence in the evolution of lifespan», *BMC Evolutionary Biology*, vol. 19, nº 1, doi :10.1186/s12862-019-1409-0. URL https://doi.org/10.1186/s12862-019-1409-0. 127

SALIS, P., T. LORIN, V. LEWIS, C. REY, A. MARCIONETTI, M.-L. ESCANDE, N. ROUX, L. BESSEAU, N. SALAMIN, M. SÉMON, D. PARICHY, J.-N. VOLFF et V. LAUDET. 2019, «Developmental and comparative transcriptomic identification of iridophore contribution to white barring in clownfish», *Pigment Cell & Melanoma Research*, vol. 32, nº 3, doi :10.1111/pcmr.12766, p. 391–402. URL https://doi.org/10.1111/pcmr.12766. 137

SANDER, C. et R. SCHNEIDER. 1991, «Database of homology-derived protein structures and the structural meaning of sequence alignment», *Proteins : Structure, Function, and Genetics*, vol. 9, nº 1, doi :10.1002/prot.340090107, p. 56–68. URL https://doi.org/10.1002/prot.340090107. 65

SCHENK, J. J., K. C. ROWE et S. J. STEPPAN. 2013, «Ecological opportunity and incumbency in the diversification of repeated continental colonizations by muroid rodents», *Systematic Biology*, vol. 62, nº 6, doi :10.1093/sysbio/syt050, p. 837–864. URL https://doi.org/10.1093/sysbio/syt050. 96

SIMÃO, F. A., R. M. WATERHOUSE, P. IOANNIDIS, E. V. KRIVENTSEVA et E. M. ZDOBNOV. 2015, «BUSCO : assessing genome assembly and annotation completeness with single-copy orthologs», *Bioinformatics*, vol. 31, nº 19, doi :10.1093/bioinformatics/btv351, p. 3210–3212. URL https://doi.org/10.1093/bioinformatics/btv351. 98

SMITH, S. D., C. ANÉ et D. A. BAUM. 2008, «THE ROLE OF POLLINATOR SHIFTS IN THE FLORAL DIVERSIFICATION OFIOCHROMA(SOLANACEAE)», *Evolution*, vol. 62, nº 4, doi :10.1111/j.1558-5646.2008.00327.x, p. 793–806. URL https://doi.org/10.1111/j.1558-5646.2008.00327.x. 22

SMITH, S. D. et M. D. RAUSHER. 2011, «Gene loss and parallel evolution contribute to species difference in flower color», *Molecular Biology and Evolution*, vol. 28, nº 10, doi :10.1093/molbev/msr109, p. 2799–2810. URL https://doi.org/10.1093/molbev/msr109. 23

SNEL, B. 2000, «STRING : a web-server to retrieve and display the repeatedly occurring neighbourhood of a gene», *Nucleic Acids Research*, vol. 28, nº 18, doi :10.1093/nar/28.18.3442, p. 3442–3444. URL https://doi.org/10.1093/nar/28.18.3442. 122, 124

SPRINGER, M. 2004, «Molecules consolidate the placental mammal tree», *Trends in Ecology & Evolution*, vol. 19, nº 8, doi :10.1016/j.tree.2004.05.006, p. 430–438. URL https://doi.org/10.1016/j.tree.2004.05.006. 7, 8, 10

STEINER, C. C., H. ROMPLER, L. M. BOETTGER, T. SCHONEBERG et H. E. HOEKSTRA. 2008, «The genetic basis of phenotypic convergence in beach mice : Similar pigment patterns but different genes», *Molecular Biology and Evolution*, vol. 26, nº 1, doi :10.1093/molbev/msn218, p. 35–45. URL https://doi.org/10.1093/molbev/msn218. 27, 29, 36

STODDEN, V., P. GUO et Z. MA. 2013, «Toward reproducible computational research : An empirical analysis of data and code policy adoption by journals», *PLoS ONE*, vol. 8, nº 6, doi :10.1371/journal.pone.0067111, p. e67 111. URL https://doi.org/10.1371/journal.pone.0067111. 43

SUZUKI, T. et G. R. WALLER. 1987, «Allelopathy due to purine alkaloids in tea seeds during germination», *Plant and Soil*, vol. 98, nº 1, doi :10.1007/bf02381733, p. 131–136. URL https://doi.org/10.1007/bf02381733. 20

TAMURI, A. U., M. DOS REIS, A. J. HAY et R. A. GOLDSTEIN. 2009, «Identifying changes in selective constraints : Host shifts in influenza», *PLoS Computational Biology*, vol. 5, nº 11, doi :10.1371/journal.pcbi.1000564, p. e1000 564. URL https://doi.org/10.1371/journal.pcbi.1000564. 78

THOMAS, G. W. et M. W. HAHN. 2015, «Determining the null model for detecting adaptive convergence from genomic data : A case study using echolocating mammals», *Molecular Biology and Evolution*, vol. 32, nº 5, doi :10.1093/molbev/msv013, p. 1232–1236. URL https://doi.org/10.1093/molbev/msv013. 61

THOMAS, G. W., M. W. HAHN et Y. HAHN. 2017, «The effects of increasing the number of taxa on inferences of molecular convergence», *Genome Biology and Evolution*, doi :10.1093/gbe/evw306, p. evw306. URL https://doi.org/10.1093/gbe/evw306. 61

THOMPSON, A. W. et G. ORTÍ. 2016, «Annual killifish transcriptomics and candidate genes for metazoan diapause», *Molecular Biology and Evolution*, vol. 33, nº 9, doi :10.1093/molbev/msw110, p. 2391–2395. URL https://doi.org/10.1093/molbev/msw110. 44

THOMPSON, D. A., S. ROY, M. CHAN, M. P. STYCZYNSKY, J. PFIFFNER, C. FRENCH, A. SOCHA, A. THIELKE, S. NAPOLITANO, P. MULLER, M. KELLIS, J. H. KONIECZKA, I. WAPINSKI et A. REGEV. 2013, «Evolutionary principles of modular gene regulation in yeasts», *eLife*, vol. 2, doi :10.7554/elife.00603. URL https://doi.org/10.7554/elife.00603. 45

VOSHALL, A. et E. N. MORIYAMA. 2018, «Next-generation transcriptome assembly : Strategies and performance analaysis», dans *Bioinformatics in the Era of Post Genomics and Big Data*, InTech, doi :10.5772/intechopen.73497. URL https://doi.org/10.5772/intechopen.73497. 43

WALLER, G. R. 1989, «Biochemical frontiers of allelopathy», *Biologia Plantarum*, vol. 31, nº 6, doi :10.1007/bf02876217, p. 418–447. URL https://doi.org/10.1007/bf02876217. 20

WARREN, W. C., L. W. HILLIER, J. A. M. GRAVES, E. BIRNEY, C. P. PONTING, F. GRÜTZNER, K. BELOV, W. MILLER, L. CLARKE, A. T. CHINWALLA et collab.. 2008, «Genome analysis of the platypus reveals unique signatures of evolution», *Nature*, vol. 453, nº 7192, doi :10.1038/nature06936, p. 175–183. URL https://doi.org/10.1038/nature06936. 9

WHITTALL, J. B., C. VOELCKEL, D. J. KLIEBENSTAIN et S. A. HODGES. 2006, «Convergence, constraint and the role of gene expression during adaptive radiation : floral anthocyanins in aquilegia», *Molecular Ecology*, vol. 15, nº 14, doi :10.1111/j.1365-294x.2006.03114.x, p. 4645–4657. URL https://doi.org/10.1111/j.1365-294x.2006.03114.x. 23

WRIGHT, G. A., D. D. BAKER, M. J. PALMER, D. STABLER, J. A. MUSTARD, E. F. POWER, A. M. BORLAND et P. C. STEVENSON. 2013, «Caffeine in floral nectar enhances a pollinator's memory of reward», *Science*, vol. 339, nº 6124, doi :10.1126/science.1228806, p. 1202–1204. URL https://doi.org/10.1126/science.1228806. 15, 20

WU, G. A., J. TEROL, V. IBANEZ, A. LOPEZ-GARCIA, E. PEREZ-ROMAN, C. BORREDA, C. DOMINGO, F. R. TADEO, J. CARBONELL-CABALLERO, R. ALONSO, F. CURK, D. DU, P. OLLITRAULT, M. L. ROOSE, J. DOPAZO, F. G. GMITTER, D. S. ROKHSAR et M. TALON. 2018, «Genomics of the origin and evolution of citrus», *Nature*, vol. 554, nº 7692, doi :10.1038/nature25447, p. 311–316. URL https://doi.org/10.1038/nature25447. 3

YANG, Z. 1997, «PAML : a program package for phylogenetic analysis by maximum likelihood», *Bioinformatics*, vol. 13, nº 5, doi :10.1093/bioinformatics/13.5.555, p. 555–556. URL https://doi.org/10.1093/bioinformatics/13.5.555. 127

YOUNG, R. L., M. H. FERKIN, N. F. OCKENDON-POWELL, V. N. ORR, S. M. PHELPS, Á. POGÁNY, C. L. RICHARDS-ZAWACKI, K. SUMMERS, T. SZÉKELY, B. C. TRAINOR, A. O. URRUTIA, G. ZACHAR, L. A. O'CONNELL et H. A. HOFMANN. 2019, «Conserved transcriptomic profiles underpin monogamy across vertebrates», *Proceedings of the National Academy of Sciences*, vol. 116, nº 4, doi :10.1073/pnas.1813775116, p. 1331–1336. URL https://doi.org/10.1073/pnas.1813775116. 126, 127

ZHENG, J., W. SHEN, D. Z. Z. HE, K. B. LONG, L. D. MADISON et P. DALLOS. 2000, «Prestin is the motor protein of cochlear outer hair cells», *Nature*, vol. 405, nº 6783, doi :10.1038/35012009, p. 149–155. URL https://doi.org/10.1038/35012009. 31

ZHU, X., Y. GUAN, A. V. SIGNORE, C. NATARAJAN, S. G. DUBAY, Y. CHENG, N. HAN, G. SONG, Y. QU, H. MORIYAMA, F. G. HOFFMANN, A. FAGO, F. LEI et J. F. STORZ. 2018, «Divergent and parallel routes of biochemical adaptation in high-altitude passerine birds from the qinghai-tibet plateau», *Proceedings of the National Academy of Sciences*, vol. 115, nº 8, doi :10.1073/pnas.1720487115, p. 1865–1870. URL https://doi.org/10.1073/pnas.1720487115. 36

ZIMMER, C. et D. J. EMLEN. 2015, *Evolution : Making sense of life*, Macmillan Higher Education. 33

ZOU, Z. et J. ZHANG. 2015, «No genome-wide protein sequence convergence for echolocation», *Molecular Biology and Evolution*, vol. 32, nº 5, doi :10.1093/molbev/msv014, p. 1237–1241. URL https://doi.org/10.1093/molbev/msv014. 61

RÉSUMÉ DE LA THÈSE

La convergence phénotypique, c'est-à-dire l'acquisition indépendante de caractères similaires par des espèces différentes, est omniprésente dans la nature et a été souvent étudiée. Mais ce processus évolutif n'est pas bien compris. Par exemple, de nombreux chercheurs cherchent à comprendre s'il existe des bases génétiques convergentes sous-jacentes à ces convergences phénotypiques.

Quelques substitutions convergentes corrélées à un phénotype convergent ont été décrites dans la littérature, mais il existe peu d'études à l'échelle génomique. Ceci peut s'expliquer par deux problèmes méthodologiques : 1/ D'une part, la difficulté de créer des jeux de données multi-espèces pour des analyses comparatives. 2/ D'autre part, le manque de méthodes dédiées à la détection de la convergence à l'échelle génomique.

Au cours de ma thèse, j'ai proposé des solutions à ces deux défis. Dans un premier temps, j'ai créé un programme (CAARS) permettant d'automatiser l'assemblage de jeux de données composés de familles d'orthologues à partir de données RNA-Seq. Puis, j'ai créé un outil (PCOC) pour étudier les substitutions convergentes au sein de séquences codantes, basé sur l'identification de changements de profils d'acides aminés. Ces outils ont été développés dans un souci de reproductibilité et de facilité d'utilisation. J'ai ensuite étudié la capacité de différentes méthodes, dont PCOC, à détecter des substitutions convergentes en présence de facteurs confondants. Enfin, j'ai appliqué ces méthodes à un cas biologique où j'ai cherché à caractériser les bases génomiques de l'adaptation aux milieux arides chez les rongeurs.

PHD THESIS SUMMARY

Phenotypic convergence, the independent acquisition of similar characters by different species, is widespread in nature and has been extensively studied. But this evolutionary process is not well understood. For example, many researchers seek to understand whether there are convergent genetic bases underlying these phenotypic convergences.

Some convergent substitutions correlated with a convergent phenotype have been described in the literature, but there are few studies at the genome scale. This can be explained by two methodological problems : 1 / On the one hand, the difficulty of creating multi-species datasets for comparative analyses. 2 / On the other hand, the lack of dedicated methods to detect convergence at the genomic scale.

During my thesis, I proposed solutions to these two challenges. As a first step, I created a program (CAARS) to automate the assembly of datasets composed of orthologous families from RNA-Seq data. Then I created a tool (PCOC) to study convergent substitutions within coding sequences, based on the identification of amino acid profile changes rather than strict amino acid changes. These tools have been developed for the sake of reproducibility and ease of use. I then studied the ability of different methods, including PCOC, to detect convergent substitutions in the presence of confounding factors. Finally, I applied these methods to a biological case where I sought to characterize the genomic bases of adaptation to arid environments in rodents.