Etude évolutive multi-niveaux des gènes de la multiciliation chez les métazoaires

Audrey Defosset

UNIVERSITÉ DE STRASBOURG

ÉCOLE DOCTORALE DES SCIENCES DE LA VIE ET DE LA SANTE

ICube – UMR 7357

THÈSE présentée par :

Audrey DEFOSSET

soutenue le : **11 septembre 2020**

pour obtenir le grade de : **Docteur de l'université de Strasbourg**

Discipline/ Spécialité : Sciences de la vie et de la Santé – spécialité Bioinformatique

Etude évolutive multi-niveaux des gènes de la multiciliation chez les Métazoaires

THÈSE dirigée par :

Mme LECOMPTE Odile	Professeur, Université de Strasbourg
M POCH Olivier	Directeur de recherche, CNRS

RAPPORTEURS EXTERNES :

M BARBRY Pascal	Directeur de recherche, CNRS
M PONTAROTTI Pierre	Directeur de recherche, CNRS

EXAMINATRICE INTERNE :

Mme FRIEDRICH Anne	Maitre de conférences, Université de Strasbourg

EXAMINATRICE EXTERNE :

Mme LANE Lydie	Co-directrice du groupe CALIPHO, Université de Genève, SIB

Remerciements

Je tenais avant tout à remercier le Dr Pascal Barbry, le Dr Pierre Pontarotti, le Dr Anne Friedrich et le Dr Lydie Lane d'avoir accepté de faire partie de mon jury, et pour l'honneur qu'ils me font d'examiner mes travaux.

Ces travaux n'auraient évidemment pas vu le jour sans l'équipe du CSTB, où la bonne humeur est toujours de mise et où il fait bon vivre. C'est notamment cette ambiance familiale, propice à la recherche, à la découverte et à la (sur-)consommation de café qui m'a permis de vivre pleinement et positivement ces trois années dédiées à ma fameuse 'Thèse couleurs'. Pour cela, je tiens à en remercier l'intégralité des membres, passés et présents. Merci à Julie, Laetitia, Anne N, Claudine et Raymond, les 'adultes' de l'équipe, toujours disponibles pour un conseil, répondre à mes questions ou tout simplement pour discuter autour d'un café. Merci à Luc et Arnaud, pour leur savoir informatique et leur statut de service technique pour tous les bugs que j'ai rencontrés. Merci aux nouveaux 'enfants', Thomas, Nicolas et Romain, et aux anciens, Pierre W. et Julio : ce fut un plaisir de partager ces années avec vous dans le rire et la boisson. Merci Gopal et Kirsley (et Pochi !), compagnons de bureau, de m'avoir supportée pendant ces années, malgré mes accès de folie réguliers. Merci Christelle d'avoir fait de mes derniers mois au labo des moments de fous rires sans égal, bien que notre temps ensemble ait été écourté (Merci Covid19 !). Bienvenue dans la famille ! Merci à tous les autres membres que j'ai pu oublier qui, d'une façon ou d'une autre, ont contribué à mon quotidien tout au long de ce doctorat.

Je n'oublie en revanche pas mes directeurs de thèse, sans qui rien de tout ça n'aurait pu exister. Merci Olivier pour ta bonne humeur, tes conseils avisés, tes idées pour le moins illuminées, tes récits d'aventures passées, et surtout ta passion que tu ne cesses de communiquer. Enfin, merci Odile pour toutes les façons dont tu as pu contribuer à cette thèse : tout d'abord pour m'avoir fait découvrir la génomique comparative, pour avoir accepté de m'encadrer lors de mon premier stage, pour avoir toujours été à l'écoute de mes problèmes, personnels ou professionnels, pour avoir su me remotiver quand j'étais au plus bas (avec ou sans tarte au citron), pour avoir toujours cru en moi, pour m'avoir poussée quand il le fallait et finalement, pour avoir relu et corrigé ce manuscrit qui existe grâce à toi.

J'aimerais ensuite remercier mes amis, pour leur soutien et pour m'avoir procuré des distractions quand j'en avais le plus besoin.

Merci au groupe des anciens BSBI, toujours à l'écoute de mes plaintes et de mes monologues interminables, et ce à n'importe quelle heure. Merci à Yanni(s) d'avoir été là pour me conseiller durant toutes ces années au laboratoire, que ça soit sur des questions personnelles ou professionnelles, et d'avoir continué à le faire après son départ. Merci à Camille pour sa compassion infinie pour toutes mes mésaventures, ses photos de chats et ses connaissances douteuses sur les *serial killers*. Merci Sam qui a toujours su me faire rire avec un commentaire ou un mème bien placé. Merci Alexia d'être un moteur de motivation d'activité physique, même si je reste lamentable en sport.

Merci aux rescapés de la TSA, toujours présents après une décennie, qui me rappellent régulièrement d'où je viens et à quel point j'ai pu évoluer ces dernières années : merci aux filles, Bénédicte et Marion, pour leurs discussions toujours très pertinentes, qu'elles soient médicales ou à

propos de robinets, et aux garçons, Adrien, Florian et Yannick, toujours partants pour un verre et pour rire.

Merci Pierre, Christophe et surtout Sophie, de m'avoir accueillie à bras ouverts dans leur monde (à coup d'envolées lyriques et de surligneur fluo) qu'il me tarde de rejoindre à plein temps.

Je remercie ma famille, en particulier ma mère, qui me soutient toujours (malgré elle) : promis, un jour j'arrêterai les études.

Il me reste une dernière personne à remercier, qui a été à mes côtés du premier au dernier jour, qui a su me soutenir autant qu'elle a pu me distraire, qui m'a permis d'agrandir ma famille à quatre pattes avec l'addition d'un petit loup, et qui, je l'espère, sera à mes côtés dans toutes mes prochaines aventures. Merci Ophélia, pour tout.

Sommaire

Liste des figures

Liste des tableaux

Abréviations

DD : *Deuterosome dependent*

GWAS : *Genome-Wide Association Study*

IFT *: Intraflagellar Transport*

LECA : *Last Eukaryotic Common Ancestor*

MCCs : *Multiciliated cells*

MCD : *Mother centriole dependent*

MSA : *Multiple Sequence Alignment*

NCID : *Notch Intracellular Domain*

PCD : *Primary ciliary dyskinesia*

PCP : *Planar Cell Polarity*

PDGF : *Platelet-derived Growth Factor*

RGMC : *Reduced generation of multiple motile cilia*

SNP : *Single nucleotide polymorphism*

SNV : *Single nucleotide variant*

WES : *Whole exome sequencing*

WGS : *Whole genome sequencing*

Avant-propos

L'objectif de ces travaux de thèse est l'étude de la multiciliation de façon intégrative à travers plusieurs approches de génomique. Dans cette optique, nous avons développé une nouvelle approche de génomique comparative qui, appliquée à la multiciliation et combinée à des approches fonctionnelles, nous a permis d'identifier des candidats potentiels à la multiciliation.

L'introduction de ce manuscrit se compose de trois chapitres dont le but est de présenter à la fois les problématiques ayant donné lieu à ces travaux de thèse mais également les concepts sur lesquels ils s'appuient.

Les deux premiers chapitres posent le contexte biologique dans lequel s'inscrivent ces travaux, en traitant successivement le cil et la multiciliation. Nos connaissances actuelles en termes de structure, de fonction et d'évolution y seront résumées en détaillant plus particulièrement la multiciliation, qui a été au centre de cette thèse et pour laquelle de nombreuses interrogations persistent à ce jour. Nous soulignerons également l'intérêt à la fois biologique et médical de l'étude de ces deux processus par leur implication dans les maladies rares connues sous le nom de ciliopathies dont les causes génétiques sont encore fréquemment non identifiées.

Le troisième chapitre est dédié à la génomique à l'ère des disciplines à haut débit et présentera de manière succincte les différents domaines qui la composent. Nous verrons dans cette partie comment chacune des approches décrites a pu contribuer à l'extension de nos connaissances concernant le cil et la multiciliation. La dernière section de ce chapitre s'intéressera aux différentes manières dont il est possible de combiner des approches de génomique par une démarche intégrative, qui a par la suite guidé notre étude sur la multiciliation.

La partie **Contributions** se décline également en trois chapitres de résultats, présentant les différentes étapes de notre étude ayant conduit à une analyse intégrative de la multiciliation.

Le premier chapitre décrit la réalisation d'une étude évolutive des gènes connus de la multiciliation par l'application de diverses méthodes de génomique comparative. Nous verrons la manière dont nous avons exploité les relations génotype-phénotype pour identifier de nouveaux candidats potentiels de la multiciliation, et comment ces résultats ont permis de mettre en évidence des cas de conservation atypique qui nous ont mené au développement d'une nouvelle approche de génomique comparative.

Le second chapitre présente BLUR, une ressource de génomique comparative multi-niveaux que nous avons développée pour permettre les comparaisons de protéomes et les études par profilage phylogénétique sur plusieurs niveaux de granularité. Nous décrirons également le site web que nous avons développé pour permettre l'utilisation interactive du programme, ainsi que l'ensemble des outils que nous avons implémentés pour réaliser des analyses approfondies des résultats.

Le troisième chapitre porte sur l'analyse intégrative de la multiciliation, réunissant un pan évolutif représenté par l'application de BLUR, et un pan fonctionnel qui comprend la combinaison de

résultats d'études de transcriptomique dédiées à la multiciliation. De cette manière, nous avons pu mettre en évidence des nouveaux candidats potentiels à la multiciliation.

Enfin, un chapitre de conclusions soulignera la place et l'intérêt de la génomique comparative dans l'ère actuelle de la biologie fonctionnelle à haut débit, et son potentiel dans l'étude des processus complexes. Le dernier chapitre de ce manuscrit est consacré aux matériels et méthodes employés au cours de ces travaux de thèse.

INTRODUCTION

Chapitre 1 : Le cil

Observés pour la première fois par Antoni van Leeuwenhoek au 17ème siècle, les cils suscitent l'intérêt de beaucoup de chercheurs, en particulier depuis qu'il est devenu clair qu'ils étaient impliqués à la fois dans les processus de développement et dans de nombreuses pathologies humaines. Malgré l'attention portée à cette organelle, beaucoup de choses à son sujet sont encore mal connues et elle fait l'objet d'un grand nombre d'études.

Nous verrons dans ce chapitre la structure et les fonctions des différents types de cil existant chez les eucaryotes, ainsi qu'un aperçu des diverses pathologies pouvant toucher le cil.

1. Structure du cil eucaryote

On distingue le cil eucaryote du flagelle procaryote qui, malgré une dénomination commune (on pense notamment au flagelle du spermatozoïde) sont deux entités entièrement différentes. Bien qu'il s'agisse de deux structures présentant des fonctions motrices que l'on retrouve à l'extérieur de la cellule, ils ne sont en revanche ni homologues, ni similaires en terme d'architecture. En effet, le flagelle procaryote est un complexe moléculaire ancré dans la membrane et présentant un filament de flagelline (Khan and Scholey, 2018), tandis que le cil est une extension de la membrane plasmique que l'on retrouve à la surface de la plupart des cellules eucaryotes, composée d'un cytosquelette de microtubules, et ancrée à la membrane cellulaire *via* une structure centrosomale appelée corps basal (*Figure 1-1*). Bien que continu avec la cellule et la membrane cellulaire, le cil est un compartiment à part entière, possédant son propre protéome régulé par un système de transport spécifique. Il existe chez les eucaryotes deux types majeurs de cils, le cil primaire et le cil motile, dont les fonctions différentes se traduisent par des structures présentant quelques variations, mais dont les composants fondamentaux restent similaires (Hoyer-Fender, 2013).

1.1. Composants du cil

Le corps basal. Dérivé du centriole, le corps basal est formé de 9 triplets de microtubules organisés de manière circulaire. Il agit comme un socle sur lequel peut se construire le cil, et sert de point de départ au développement de l'axonème. La fixation du corps basal à la membrane se fait grâce à des fibres de transitions résultat de la spécialisation des appendices distaux du centriole père. On retrouve également des pieds basaux, dérivés des appendices subdistaux, qui permettent la fixation des microtubules cytoplasmiques à la base du cil. Enfin, il existe au niveau de l'extrémité proximale du corps basal une structure striée appelée racine ciliaire, qui semble également avoir un rôle d'ancrage du cil (*Figure 1-1*).

L'axonème. Formé par 9 doublets périphériques de microtubule, l'axonème représente la partie axiale et saillante du cil. Chaque doublet est constitué d'un protofilament A de tubuline complet, adjacent à un protofilament B incomplet et l'extrémité positive des filaments est située dans la partie distale du cil, par laquelle ce dernier croit. On le retrouve dans deux configurations majeures, selon le type de cil considéré. Dans le cas du cil motile, les doublets sont accompagnés d'une paire centrale de microtubules (structure dite « 9+2 ») et de bras de dynéine, nécessaires au mouvement. Le cil non motile quant à lui est dépourvu de microtubules centraux et de dynéine, il adopte une structure dite « 9+0 ». Il existe par ailleurs des cas exceptionnels dont nous parlerons dans une partie ultérieure.

La zone de transition. La zone de transition est la région ciliaire à la base de l'axonème faisant suite au corps basal et aux fibres de transition. Elle est caractérisée par la présence de projections en forme de Y qui ont pour rôle de connecter chaque doublet de l'axonème à la membrane ciliaire et au collier ciliaire, un ensemble de structures circulaires membranaires. Cette zone permet la compartimentation du cil et agit comme une barrière sélective contrôlant les entrées et les sorties de protéines dans le cil, et par conséquent, sa composition moléculaire. Elle interagit également avec la machinerie de transport intra-flagellaire pour permettre le transport contrôlé de molécules dans le cil (Reiter et al., 2012).

La membrane ciliaire. Continue avec la membrane plasmique, elle est toutefois distincte de cette dernière de par sa composition lipidique et protéique, et joue un rôle important dans la signalisation ciliaire. Selon la localisation et la fonction du cil considéré, la membrane présentera différents types de récepteurs transmembranaires, les plus connus étant les récepteurs impliqués dans la voie de signalisation *Hedgehog* (Garcia et al., 2018).

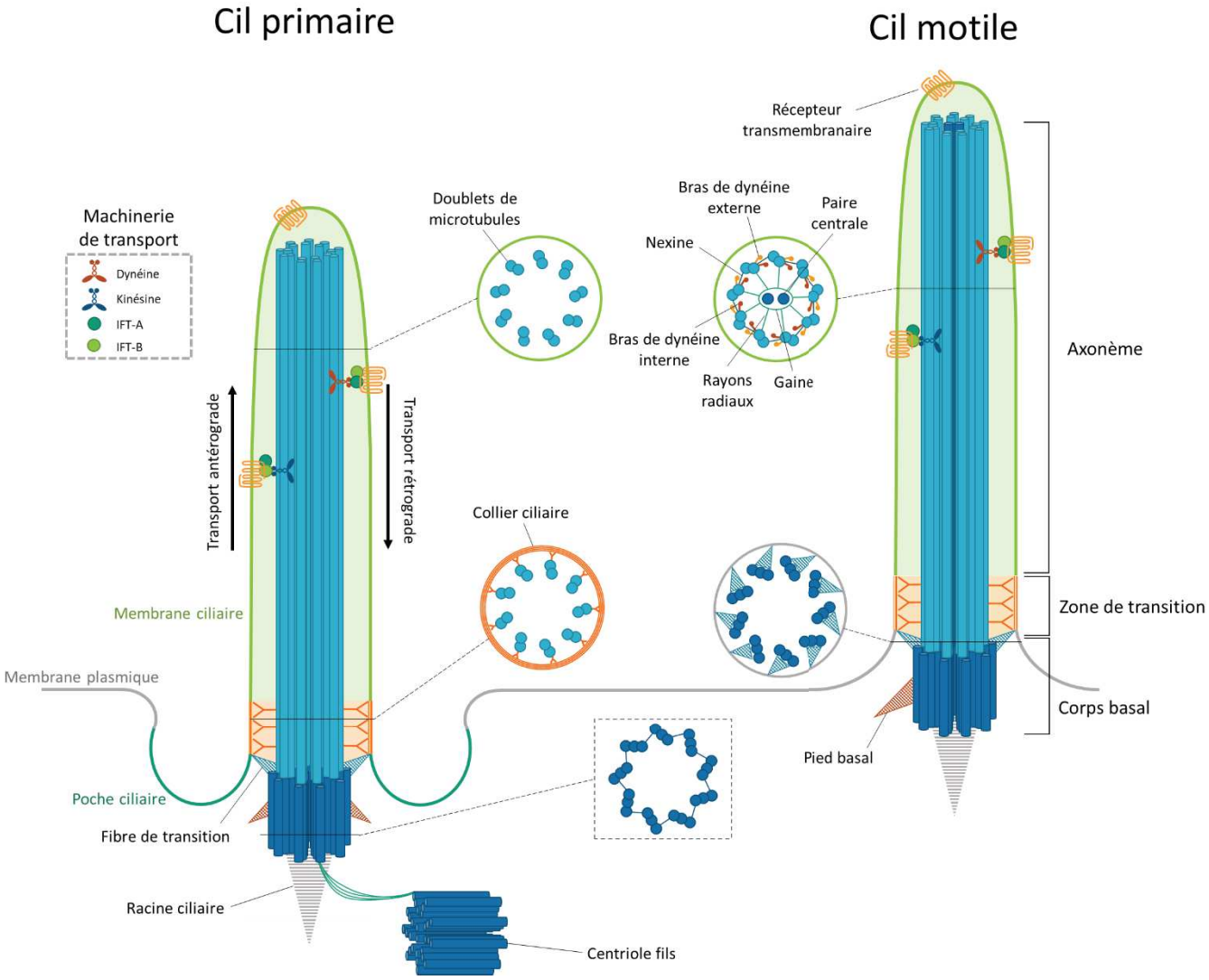

Figure 1-1: Structures des cils primaire et motile chez les eucaryotes avec coupes transversales.

Le transport intra-flagellaire. L'élongation de l'axonème et le maintien de la structure ciliaire sont assurés par une machinerie de transport spécifique, le cil n'étant pas capable de synthétiser ses propres protéines et la zone de transition empêchant les échanges avec la cellule. Il existe notamment le complexe IFT (*Intraflagellar Transport*), que l'on peut diviser en deux sous-complexes : IFT-A pour le transport rétrograde et IFT-B pour le transport antérograde. Ainsi, le complexe IFT-B, en s'associant à la kinésine, est requis pour l'assemblage et le maintien du cil, tandis que le complexe IFT-A qui s'associe à la dynéine a une fonction dans le renouvellement des protéines (Ishikawa and Marshall, 2017). On retrouve également un second complexe associé aux protéines IFT, le BBSome, dont le rôle est celui d'adaptateur de cargo reconnaissant les protéines destinées à la membrane ciliaire.

1.2. Cil primaire

Le cil primaire est une structure dynamique qui s'assemble et se désassemble selon les étapes du cycle cellulaire. On le retrouve à la surface de la plupart des cellules différenciées des vertébrés lorsque celles-ci sont dans un état quiescent, il disparaît ensuite lorsque la cellule entre dans un nouveau cycle de prolifération. Il se développe à partir du centriole père après que celui-ci ait migré vers la membrane, accompagné du centriole fils. Il arrive dans certains cas que le cil ne soit pas entièrement externalisé et que les centrioles soient logés dans une invagination de la membrane, donnant lieu à une poche ciliaire.

1.3. Cil motile

Comme nous l'avons vu précédemment, l'axonème du cil motile est caractérisé par une structure de type « 9+2 » et la présence de protéines spécialisées permettant le mouvement. Les doublets périphériques sont liés entre eux par des ponts de nexine et chaque doublet possède des bras de dynéine externe et interne, servant de moteurs moléculaires. La paire centrale de microtubules est entourée d'une gaine projetant des rayons vers chaque doublet périphérique. L'ensemble de ces composants a pour but de faciliter le glissement des doublets les uns par rapport aux autres lors de la courbure du cil.

Une seconde différence notable entre le cil motile et le cil primaire est la structure du corps basal, qui reflète à la fois le mode de génération du cil et sa fonction. Contrairement au cil primaire, transitoire et issu du centriole père, le cil motile ne possède qu'une seule structure centriolaire à sa base, générée *de novo* et ancrée par un seul pied basal, orienté dans le sens du battement du cil.

Au sein des cils motiles, il existe également le cas particulier du flagelle des spermatozoïdes chez les mammifères, dont la structure présente quelques spécificités (*Figure 1-2*). On retrouve ainsi des fibres denses externes de kératine associées à chaque doublet de microtubules sur une partie du flagelle, dont le rôle est de supporter la tension générée lors du mouvement. Pour permettre un fonctionnement optimal et une autonomie du spermatozoïde, le flagelle est doté d'une région comprenant une gaine de mitochondries ayant pour but de fournir toute l'énergie nécessaire au déplacement du gamète (Lindemann and Lesich, 2016).

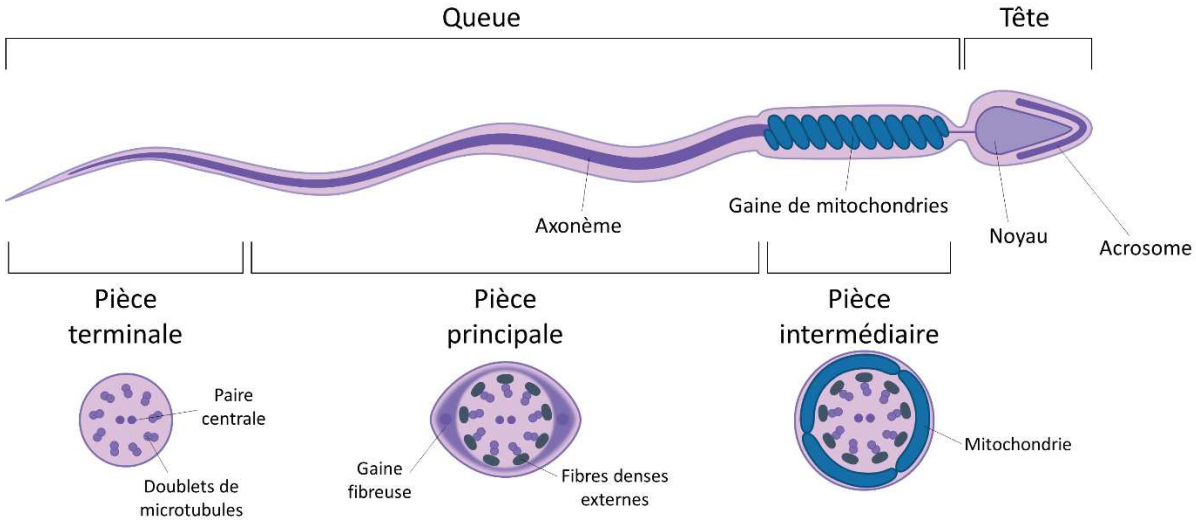

Figure 1-2: Structure du spermatozoïde de mammifère avec coupes transversales des différentes zones de la queue.

1.4. Cas particuliers

Bien que la majorité des cils que l'on peut rencontrer soit classable dans les deux catégories vues précédemment, il existe des exceptions, dont deux notables chez l'Homme. Le cil nodal, localisé au niveau du nœud de Hensen lors du développement embryonnaire, présente une configuration de type « 9+0 » bien qu'il soit motile. Il est dépourvu de la paire centrale de microtubules mais possède des bras de dynéine qui lui permettent de générer un flux orienté à l'origine de l'asymétrie gauche/droite du corps. A l'inverse, une sous-population de cellules de la cochlée dispose de cils non motiles de structure « 9+2 » appelés kinocils (Falk et al., 2015).

2. Fonctions du cil

Nous avons vu dans la partie précédente que les cils peuvent, de manière générale, être classés dans deux catégories selon leurs structures : il en est de même pour leurs fonctions. Habituellement, on séparera les fonctions de type sensoriel, assurées par les cils primaires, et de type moteur, assurées par les cils motiles.

2.1. Détection et transmission de signaux

Longtemps considéré comme un vestige cellulaire, l'importance du cil primaire dans la détection et la transmission de signaux, particulièrement au cours du développement, a pu être mise en évidence à travers de nombreuses études et lui vaut aujourd'hui le nom « d'antenne de la cellule » (Singla and Reiter, 2006). Grâce au nombre important de récepteurs, canaux transmembranaires et autres protéines localisés dans sa membrane, le cil est capable d'interpréter des signaux provenant de son environnement, qu'ils soient de type optique, chimique, osmotique ou mécanique. En réponse à ces stimuli, le cil initie diverses cascades de signalisation impliquées à la fois dans le développement embryonnaire mais également dans des fonctions cellulaires primordiales telles que la régulation de la prolifération, la différenciation, la mise en place de la polarité, l'apoptose ou encore l'homéostasie tissulaire (Pala et al., 2017).

La fonction pour laquelle le cil primaire est peut-être le mieux connu est son implication dans les voies de signalisation Hedgehog, Wnt, PDGF (*Platelet-derived Growth Factor*) et Notch,

régulateurs clés du développement animal et du maintien de l'intégrité des différents organes. Dans certains tissus, les cils développent des fonctions spécifiques, notamment au niveau de la rétine et des reins. Les photorécepteurs de la rétine sont formés à partir de cils primaires fortement modifiés, et contiennent de nombreux pigments ainsi que des protéines impliquées dans la transduction de la lumière (Wheway et al., 2014). Au niveau des reins, le rôle du cil primaire est mécanosenseur ; la détection d'un flux de fluide va engendrer un afflux de calcium dans la cellule (Praetorius and Spring, 2003).

2.2. Mouvements et flux de liquides

Contrairement aux cils primaires que l'on retrouve à la surface de la majorité des cellules, les cils motiles ne sont présents que sur certains types cellulaires spécifiques. Chez les mammifères, les spermatozoïdes disposent d'un cil motile unique spécialisé, plus communément appelé flagelle, dont le but est de permettre le déplacement des spermatozoïdes le long du tractus génital féminin. Plus généralement, les cils motiles sont trouvés en grand nombre à la surface de cellules épithéliales multiciliées qui, chez les mammifères, tapissent le système respiratoire, les ventricules du cerveau, et le système reproducteur. Ils ont ainsi pour but de créer un flux directionnel permettant la circulation d'un liquide ou le déplacement de particules (Satir and Christensen, 2007). Nous étudierons plus en détail les cellules multiciliées dans le chapitre suivant qui leur est dédié.

3. Les ciliopathies

De par leur implication dans de nombreux processus et leur localisation dans la majorité des cellules différenciées de l'organisme, les cils sont associés à de nombreuses maladies aux phénotypes complexes et variés. Tout comme pour la structure et la fonction du cil, il est possible de distinguer deux catégories de ciliopathies, chacune liée à un type ciliaire. Un type particulier de ciliopathies touchant spécifiquement les cellules multiciliées appelées génération réduite des cils motiles multiples (*RGMC* : *Reduced generation of multiple motile cilia*) sera traité dans le chapitre suivant.

3.1. Ciliopathies sensorielles

Les cils primaires étant presque ubiquitaires, il n'est pas surprenant que les dysfonctionnements les impliquant soient à l'origine de pathologies multi-systémiques présentant des phénotypes variés. Les troubles les plus souvent observés dans le cadre de ciliopathies sensorielles sont des atteintes rénales, en particulier la formation de kystes, des atteintes de la rétine, des troubles du développement mental, des anomalies du squelette comme la polydactylie, ou encore l'obésité. Il existe un important chevauchement phénotypique entre les différentes ciliopathies connues, rendant ainsi le diagnostic clinique parfois complexe, bien que l'association spécifique de certains signes puisse orienter vers un syndrome particulier (*Figure 1-3*). Parmi les ciliopathies à atteintes multi-systémiques, on retrouve notamment les syndromes de Joubert, de Bardet-Biedl et de Meckel (Bachmann-Gagescu, 2014; Reiter and Leroux, 2017). D'autres pathologies ne touchent qu'un seul système, comme la néphronophtise et la polykystose rénale atteignant les reins ou l'amaurose congénitale de Leber qui touche les yeux. L'ensemble de ces pathologies appartient à la classe des maladies rares dont la prévalence en Europe est inférieure à 1 personne atteinte sur 2000.

D'un point de vue génétique, plus de 750 gènes sont à ce jour associés au cil et certaines estimations porteraient à plus de 1000 le nombre de gènes impliqués dans le ciliome (van Dam et al., 2019). Parmi ces gènes, 251 sont associés à une ciliopathie dans la base de données Orphanet [http://www.orpha.net/], certains étant mis en cause dans plusieurs syndromes ; on retrouve notamment 6 gènes impliqués à la fois dans les syndromes de Joubert et de Meckel. Les recoupements à la fois phénotypique et génétique que l'on peut observer font des ciliopathies des pathologies complexes parfois difficiles à diagnostiquer.

Figure 1-3: Expression phénotypique de quelques ciliopathies sensorielles. Il existe un fort recoupement entre les syndromes à atteintes multi-systémiques, tandis que d'autres pathologies n'atteignent qu'un seul système. L'intensité des couleurs indique la fréquence du phénotype. ALMS : Syndrome d'Alström ; BBS : Syndrome de Bardet-Biedl ; LCA : Amaurose congénitale de Leber ; JBTS : Syndrome de Joubert ; MKS : Syndrome de Meckel ; JATD : Syndrome de Jeune ; NHPH : Nephronophtise ; OFD : Syndrome Oro-facio-digial I ; MORM : Syndrome MORM ; EVC : Syndrome d'Ellis-Van Creveld.

3.2. Dyskinésies ciliaires primitives

Les dyskinésies ciliaires primitives (PCD : *Primary ciliary dyskinesia*) représentent les ciliopathies touchant les cils motiles. Il s'agit de maladies rares atteignant jusqu'à 1 personne sur 10 000 en Europe à l'origine de phénotypes majoritairement respiratoires, tels que des détresses respiratoires chez les nouveau-nés, une dilatation des bronches, ou des infections respiratoires chroniques. On y retrouve fréquemment associés des troubles auditifs, une infertilité et dans 50% des cas un *situs inversus*. Ce dernier symptôme est dû au dysfonctionnement du cil nodal au cours de développement, perturbant ainsi la mise en place de l'asymétrie gauche/droite ; lorsque ce symptôme est présent, on parle alors de syndrome de Kartagener. Certaines formes sévères de PCD sont également associées à une hydrocéphalie causée par un défaut de drainage de liquide cérébrospinal dans les ventricules du cerveau. A ce jour, des mutations dans plus de 40 gènes ont été associés à une PCD, affectant majoritairement les acteurs du mouvement (bras de dynéine, rayons radiaux…), mais pour 25% des patients la cause génétique n'est pas identifiée (Lucas et al., 2020).

3.3. Dyskinésies ciliaires secondaires

Il existe une classe particulière de défauts du cil dans le système respiratoire pouvant apparaitre de manière secondaire à une autre pathologie ou à différents facteurs ; on parle alors de dyskinésie ciliaire secondaire ou acquise. Les symptômes généralement associés à ces pneumopathies incluent des cils plus longs ou plus courts, parfois absents, ainsi que des anomalies de l'évacuation du mucus. Typiquement, ces troubles peuvent être induits par des éléments externes, tels que la consommation de cigarettes ou par la pollution environnementale, mais également par des facteurs physiopathologiques comme les infections récurrentes, l'asthme, la bronchectasie ou encore les bronchopneumopathies chroniques obstructives (Tilley et al., 2015).

4. Perspective évolutive du cil

D'un point de vue évolutif, le cil est une structure retrouvée dans tous les clades majeurs des eucaryotes, suggérant ainsi sa présence dans le dernier ancêtre commun des eucaryotes (*LECA : Last Eukaryotic Common Ancestor*). Des études ont montré que le cil retrouvé chez LECA possédait vraisemblablement une architecture de type « 9+2 » avec des bras de dynéine, qu'il était doté de motilité et probablement d'une fonction sensorielle (Mitchell, 2017). La présence et l'absence de cil a été répertoriée chez les eucaryotes dans un effort de classification des espèces, permettant ainsi d'avoir une vue globale des évènements de perte ayant eu lieu au cours de l'évolution (Adl et al., 2012). Ainsi, il apparaît clair que le cil a été perdu de façon indépendante dans plusieurs taxons, notamment chez la plupart les plantes terrestres (gymnospermes et angiospermes), quelques alvéolés et amibozoaires ainsi que la majorité des champignons (*Figure 1-4*). Chez la plupart des espèces ayant conservé le cil, celui-ci est sous sa forme motile et sert au déplacement de l'organisme. En revanche, l'apparition de tissus et d'organes multicellulaires permettant, entre autre, la locomotion chez les métazoaire a rendu la capacité motrice du cil obsolète dans une grande partie de leurs cellules, il y a donc eu une simplification de la structure pour ne garder que les fonctions sensorielles, donnant ainsi lieu au cil primaire (Bornens, 2018). Le cil motile reste majoritairement utilisé pour le déplacement du gamète chez ces espèces, on retrouve néanmoins des exceptions comme chez les nématodes (dont l'organisme modèle *Caenorhabditis elegans*), où l'on ne retrouve le cil que sous sa forme primaire et sensorielle.

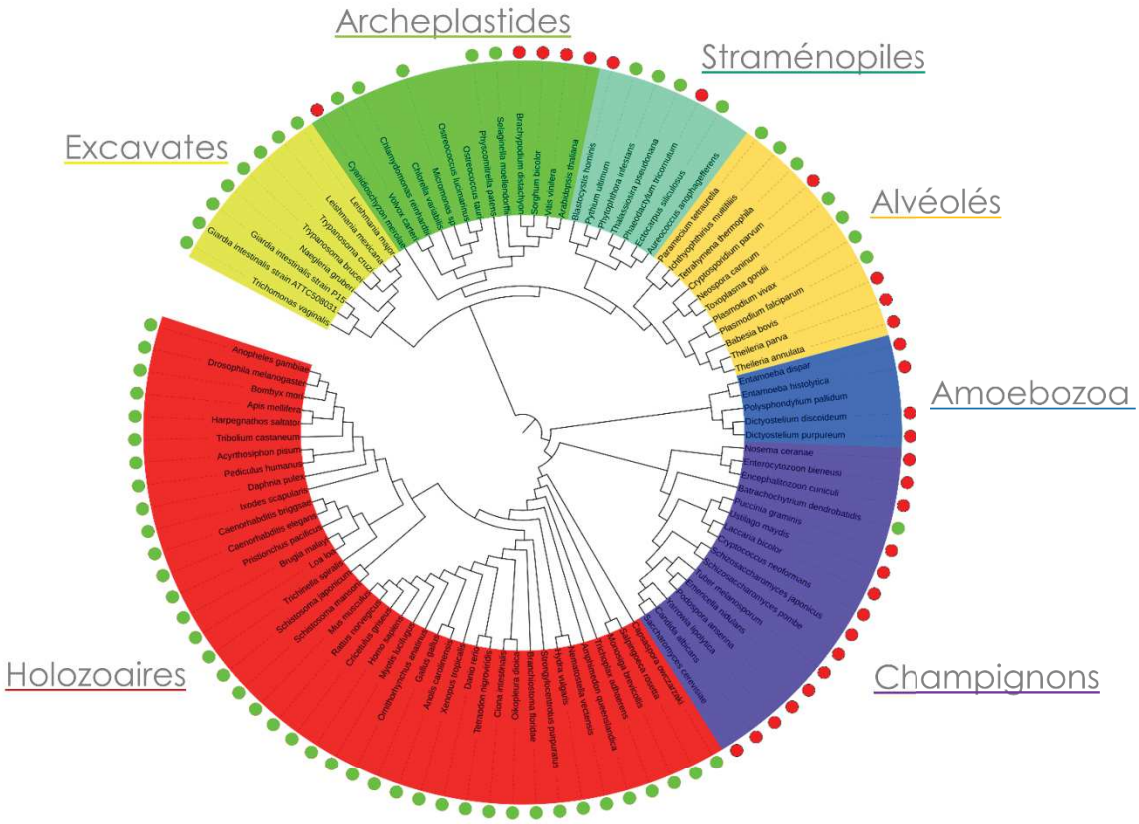

Figure 1-4: Distribution du cil dans 100 espèces représentatives des eucaryotes. Les différentes colorations représentent les taxons majeurs eucaryotes. La présence du cil est indiquée par un cercle vert, son absence par un cercle rouge. L'absence de cercle indique que le phénotype n'est pas connu. Les Holozoaires correspondent aux Métazoaires et à leurs plus proches parents unicellulaires. Figure adaptée de Nevers, communication personnelle.

Chapitre 2 : La multiciliation

L'étude du cil provoque depuis déjà plusieurs décennies un engouement certain, mais ce n'est que récemment que des travaux portant spécifiquement sur les cellules multiciliées ont vu le jour, et la majorité des grandes découvertes concernant la multiciliation ont été réalisées au cours des dix dernières années. Malgré l'importance physiologique des cellules multiciliées et l'existence de pathologies qui leur sont associées, il reste aujourd'hui de nombreux points à élucider, notamment concernant leurs mécanismes de régulation. Le nombre d'études portant sur la multiciliation croit constamment, mais comme c'est souvent le cas, chaque nouvelle découverte apporte son lot de questions à résoudre.

Dans ce chapitre, nous verrons tout d'abord la multiciliation sous l'œil de l'évolution, avant d'étudier les différents rôles des cellules multiciliées dans divers organismes. Nous discuterons ensuite des principaux mécanismes de régulation de la multiciliogénèse, et les conséquences de dysfonctionnements de cette dernière.

1. Evolution de la multiciliation

Nous avons vu précédemment que la présence et l'absence du cil eucaryote avaient été largement étudiées, il n'en est en revanche pas de même pour la présence de cils multiples. Il est d'autant plus complexe de construire une distribution précise de la multiciliation que l'absence de ce trait est rarement établie avec certitude dans la littérature. Elle a été observée à maintes reprises chez les Vertébrés, en particulier chez les Mammifères, mais la présence de cellules multiciliées (*MCCs : Multiciliated cells*) ne reste que très peu documentée au-delà de ces espèces.

Au sein du règne animal, il est difficile de caractériser les divers évènements évolutifs ayant abouti à la répartition actuelle des MCCs (*Figure 2-1*). Elles semblent avoir été présentes chez l'ancêtre des Spiralia avant d'être perdues dans certains clades, mais il n'est en revanche pas possible de dire si l'absence de multiciliation chez les Ecdysozoaires reflète l'état ancestral des Protostomia ou s'il s'agit d'une conséquence de la perte de ciliation chez les Ecdysozoaires. Le cas des Deuterostomia est tout aussi complexe, avec une présence de MCCs dans certains embranchements seulement, et une ambiguïté chez les Hémichordés avec un clade strictement monocilié et un clade multicilié (Nielsen, 2012). Il est donc possible que la multiciliation soit apparue à plusieurs moments au cours de l'évolution des métazoaires et que les mécanismes responsables de la génération de multiples cils soient entièrement différents entre les divers embranchements animaux. Ceci est notamment suggéré par l'existence d'un centriole accessoire associé à chaque corps basal dans les MCCs de certaines éponges, alors que nous avions vu précédemment que les cils motiles ne présentent habituellement pas cette caractéristique (Boury-Esnault et al., 1999).

En dehors des métazoaires, la présence de cils multiples a été observée dans des taxons très diversifiés, notamment chez des unicellulaires comme *Multicilia marina*, une espèce d'amibozoaire (Nikolaev, 2006), ou chez les Ciliophora comme *Tetrahymena* et *Paramecium* (Allen, 1969; Machemer, 1972). Du côté des organismes multicellulaires, on note la présence de multicils au niveau des spermatozoïdes de certaines plantes, comme les fougères du genre *Marsilea*, les espèces du genre *Zamia* (Mizukami and Gall, 1966), et d'autres fougères et gymnospermes (Hodges et al.,

2012). Là encore, les processus de génération des corps basaux nécessaires à la formation de multiples cils semblent tout à fait différents de ceux retrouvés chez l'Homme.

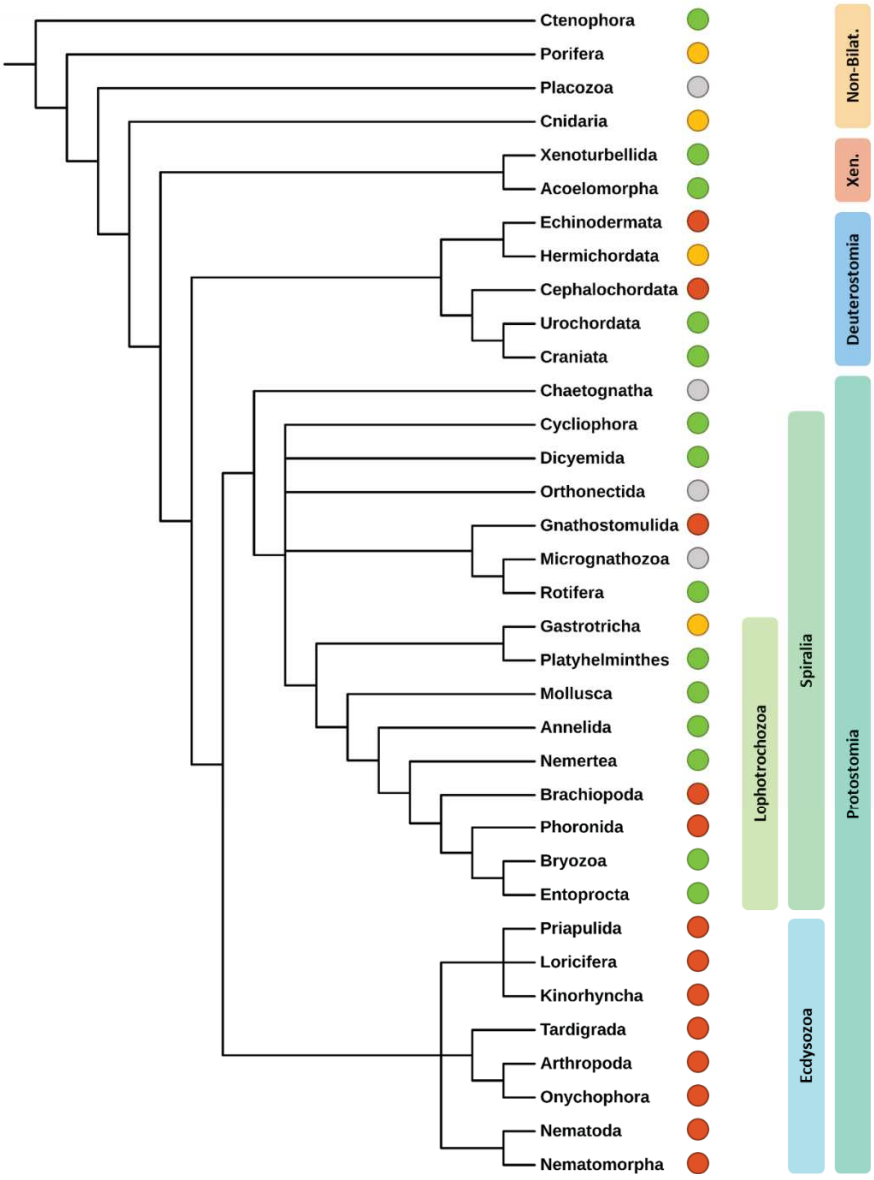

Figure 2-1: Distribution de la multiciliation dans les différents embranchements métazoaires. Un cercle vert indique la présence de MCCs dans le taxon considéré, un cercle rouge indique l'absence de MCCs. Un cercle jaune indique la coexistence de clades strictement monociliés et de clades multiciliés dans un même phylum. L'absence d'information quant à la condition multiciliée est indiquée par un cercle gris. Xen. : Xenacoelomorpha ; Non-Bilat : Non-Bilatérien. Phylogénie basée sur les résultats de (Giribet, 2016).

2. Les cellules multiciliées et leurs rôles chez les métazoaires

Les MCCs sont retrouvées dans différents taxons animaux adaptés à divers modes de vie, par conséquent, leur localisation et leurs fonctions peuvent varier, bien que leur structure et leur mode de fonctionnement restent similaires.

2.1. Organisation d'une cellule polarisée

Les MCCs sont caractérisées par la présence de plusieurs dizaines voire plusieurs centaines de cils motiles à leur surface, dont le battement coordonné et directionnel permet la génération de

force hydrodynamique. Pour permettre une efficacité optimale du flux créé par les MCCs, il est impératif que ces dernières soient orientées de façon précise et coordonnée, à la fois au niveau de chaque cellule, mais également au niveau du tissu entier dans le cas d'un organisme pluricellulaire. Il est également important que le battement engendré par les cils soit effectué de façon métachrone (voir section 2.1.2 plus bas), pour assurer le déplacement progressif et continu du fluide en contact avec les cellules.

2.1.1. Polarisation planaire

Des études réalisées chez le xénope ont montré que la polarisation des MCCs est régulée par la cascade de signalisation de polarité planaire PCP (*Planar Cell Polarity*), et implique deux mécanismes distincts, l'un mettant en place la polarité planaire dite rotationnelle, l'autre la polarité planaire tissulaire. La première désigne l'orientation et l'espacement de chaque corps basal au sein de la cellule, l'autre la polarisation coordonnée des MCCs d'un même tissu (Wallingford, 2010).

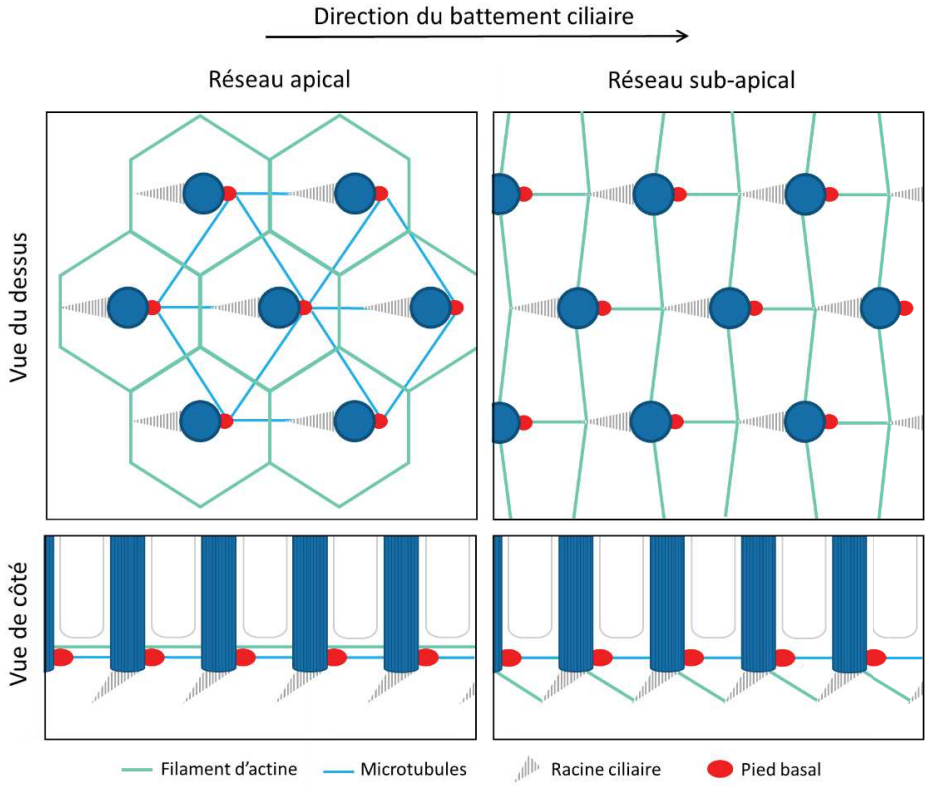

Figure 2-2: Représentation schématique des réseaux d'actine et de microtubules du pôle apical des cellules multiciliées.

Les corps basaux sont ancrés au niveau de la membrane apicale de façon polarisée par un pied basal, dirigé dans le sens du battement, et par la racine ciliaire, orientée dans la direction opposée. Des études réalisées chez le xénope ont montré qu'il existait dans les MCCs deux réseaux d'actine interconnectés, un apical et un sub-apical. Le premier est retrouvé sous forme de réseau régulier autour des corps basaux, le second, situé juste en dessous, est composé de ponts d'actine liant entre eux les corps basaux *via* leur racine ciliaire (*Figure 2-2*). Il a également été montré que les microtubules, associés aux pieds basaux, jouent un rôle dans l'orientation des corps basaux de façon locale, contrairement aux réseaux d'actine qui semblent impacter la localisation des corps basaux de façon plus globale au sein de la cellule (Herawati et al., 2016; Werner et al., 2011).

En ce qui concerne la polarisation tissulaire, permettant la coordination du battement sur l'ensemble du tissu, des études réalisées chez la souris ont montré l'existence d'une répartition asymétrique des protéines PCP. Ces protéines sont ensuite capable de transmettre un signal aux cellules voisines et génèrent, à l'aide de microtubules, une polarisation sur l'ensemble du tissu (Vladar et al., 2012).

2.1.2. *Battement métachrone*

Le rôle des MCCs est de créer un mouvement de fluide, soit pour faciliter la locomotion de l'organisme, soit pour permettre la circulation d'un liquide biologique. De ce fait, les cils sont soumis à des forces visqueuses nécessitant des adaptations afin d'optimiser l'efficacité de leur mouvement. C'est en 1981 que les caractéristiques bi-phasique et métachrone du battement ciliaire ont été mises en évidence dans l'épithélium de trachée de lapin (Sanderson and Sleigh, 1981). En effet, le battement ciliaire s'effectue en deux phases successives, une première phase effectrice rapide, durant laquelle le cil se place de façon perpendiculaire à l'épithélium, puis une seconde phase de rétablissement plus lente, pendant laquelle le cil sera courbé et dans un axe parallèle à l'épithélium pour revenir dans sa position initiale (*Figure 2-3*). A la surface d'une même cellule, on peut observer un décalage de phase entre les cils dans l'axe du battement créant ainsi une onde métachrone dont le but est d'augmenter l'efficacité jusqu'à 10 fois par rapport à une onde synchone (Elgeti and Gompper, 2013).

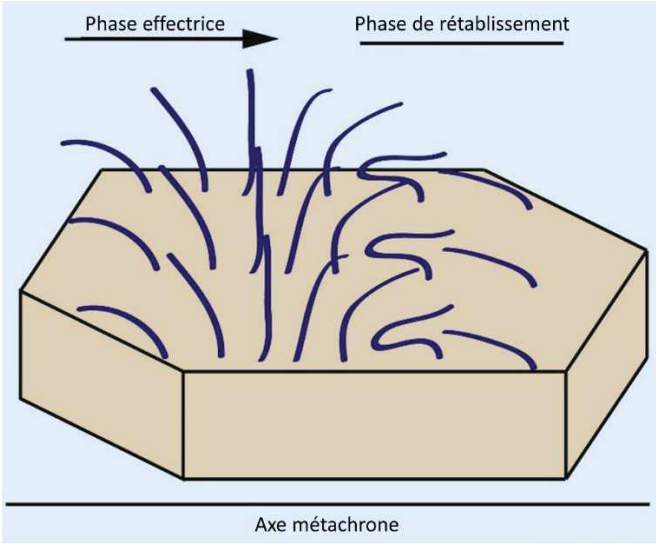

Figure 2-3: Représentation schématique de l'onde métachrone du battement ciliaire. (Figure adaptée de Brooks and Wallingford, 2014)

2.2. Rôles physiologiques de la multiciliation chez les mammifères

Chez les mammifères, les MCCs sont retrouvées au niveau du système respiratoire, où elles participent à la clairance du mucus, dans les ventricules du cerveau, où elles permettent la circulation du liquide cérébrospinal, ainsi que dans le système reproducteur (Spassky and Meunier, 2017).

2.2.1. Système respiratoire

L'épithélium du système respiratoire est doté de cellules sécrétrices capables de fabriquer un liquide visqueux appelé mucus, dont le rôle est de piéger les particules inhalées telles que les pathogènes, les poussières et autres macromolécules. Ces cellules sont accompagnées de MCCs, dont le battement ciliaire coordonné va permettre l'évacuation du mucus et des particules, au cours d'un mécanisme appelé clairance muco-ciliaire. On retrouve cet épithélium au niveau de la cavité nasale, de la trachée, des bronches et des bronchioles. La proportion de MCCs est forte à l'entrée du système respiratoire, et diminue à mesure que l'on s'approche des alvéoles (*Figure 2-4*).

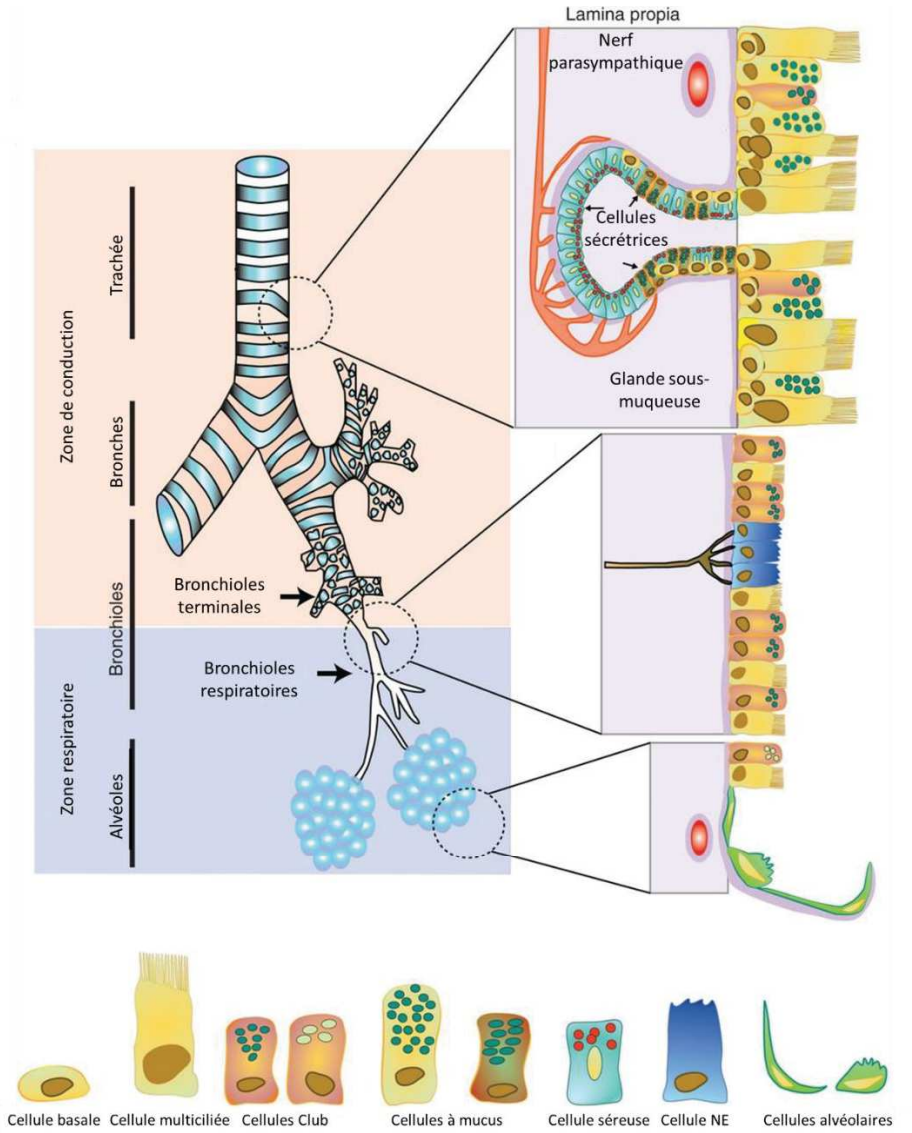

Figure 2-4: Représentation schématique de l'épithélium respiratoire dans les différentes régions du système respiratoire. Cellule NE : Cellule neuroendocrine. Adapté de (Bustamante-Marin and Ostrowski, 2017).

Les MCCs du système respiratoire présentent plus de 200 cils à leur surface, d'une longueur d'environ 7µm. Leur longueur est limitée dans ce tissu pour pouvoir maintenir une rigidité de l'organelle suffisante pour mouvoir le mucus. Afin d'optimiser le battement de ces cils, le système muco-ciliaire est composé de deux couches de fluides : une couche périciliaire lubrifiante autour des cils, et une couche de mucus dont la consistance est semblable à un gel, localisé au-dessus des cils

(*Figure 2-5*). Le battement des cils est dirigé vers le pharynx, où le mucus est expulsé, et sa fréquence est variable selon les conditions physiologiques (Fahy and Dickey, 2010).

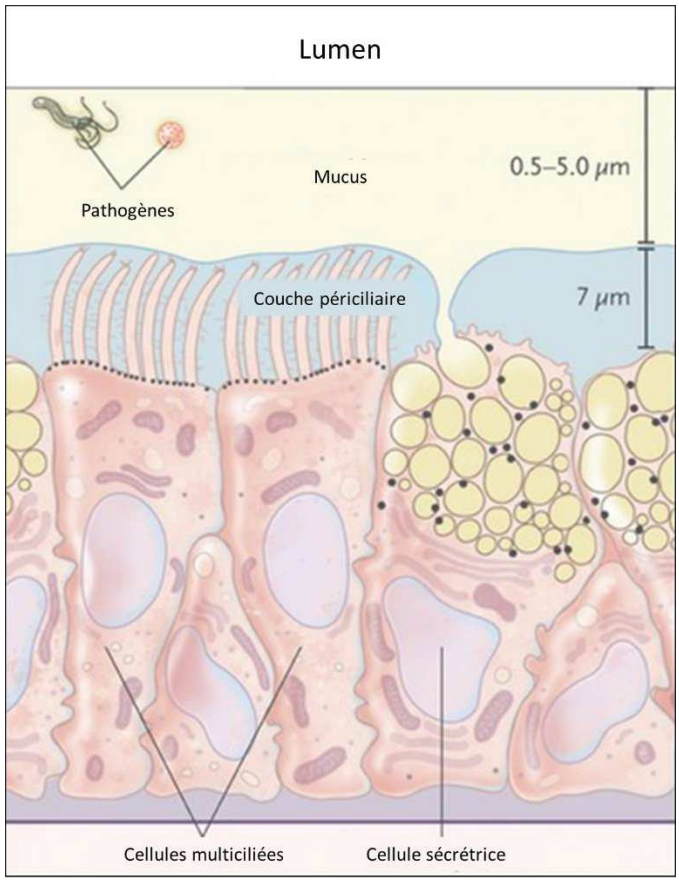

Figure 2-5: Représentation schématique d'une partie de l'épithélium respiratoire, composé de cellules ciliées et de cellules sécrétrices. Les cils sont entourés d'une couche de liquide périciliaire, au-dessus de laquelle se trouve le mucus produit par les cellules sécrétrices, dont le rôle est de piéger les pathogènes. (Figure adaptée de Fahy and Dickey, 2010)

2.2.2. *Système nerveux*

Chez les mammifères, le système nerveux est entouré d'un liquide biologique appelé liquide cérébrospinal, dont les rôles incluent la protection mécanique des structures nerveuses, le transport de molécules (hormones, nutriments…) et l'évacuation des déchets. Ce liquide est synthétisé par les plexus choroïdes, structures particulières localisées dans les ventricules du cerveau, par filtration du sang artériel. Il est composé en majeure partie d'eau, de protéines, d'ions, de glucose et de neurotransmetteurs. On retrouve dans le cerveau des mammifères quatre ventricules, dans lesquels circule le liquide cérébrospinal : il passe d'abord dans les ventricules latéraux, descend dans le 3ème ventricule, puis dans le 4ème, à partir d'où il peut rejoindre le canal de l'épendyme ou l'espace subarachnoïdien où il sera réabsorbé (Khasawneh et al., 2018).

La circulation du liquide cérébrospinal est permise par différents mécanismes, le principal étant les contractions du système cardiovasculaire, créant ainsi un flux pulsatile. Il existe en revanche une circulation spécifique laminaire proche des parois des ventricules réalisée par un ensemble de cellules épithéliales multiciliées appelées épendymocytes. Ces cellules cuboïdes présentent à leur surface des cils motiles d'une longueur d'environ 8 à 15 µm, organisés en une petite zone recouvrant moins de 35% du pôle apical de la cellule. Ce groupement contient entre 30 et 70 cils et est localisé

en aval du flux dû au battement ciliaire (*Figure 2-6*). Tout comme dans le système respiratoire, la fréquence du battement est modulable selon les conditions.

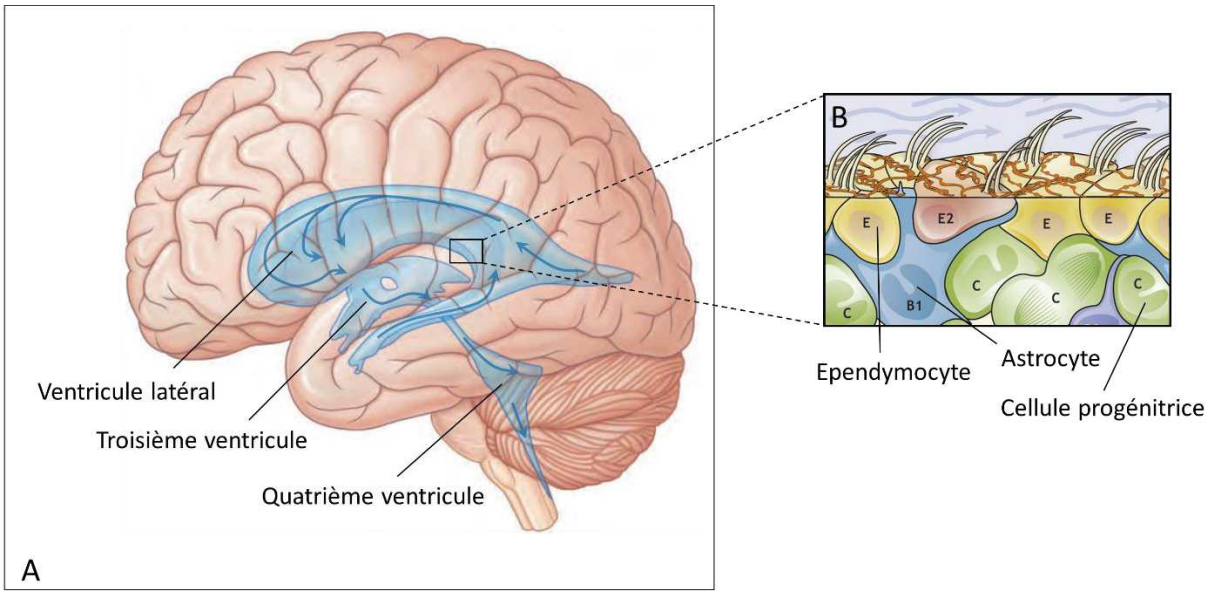

Figure 2-6: Schéma de la multiciliation dans le système nerveux. ***A.*** *Anatomie du cerveau et des ventricules, le flux du liquide cérébrospinal est représenté par des flèches. Adapté de (Standring, 2016)* ***B.*** *Epithélium des ventricules du cerveau, contenant des épendymocytes multiciliés, des astrocytes et des cellules progénitrices. Adapté de (Obernier and Alvarez-Buylla, 2019)*

2.2.3. Système reproducteur

En ce qui concerne le système reproducteur, les MCCs sont majoritairement retrouvées chez la femme, le long des trompes de Fallope et au niveau de l'épithélium utérin. Le rôle des cils présents dans les trompes est d'acheminer l'ovule, alors que celui des MCCs de l'utérus n'est pas bien connu. La proportion de cellules ciliées varie en fonction de la localisation le long du tractus, allant de 30% au niveau de l'isthme jusque 80% au niveau du pavillon. Elle varie dans l'épithélium de l'utérus en fonction du cycle ovarien, atteignant 20% au moment de l'ovulation. Ces proportions sont susceptibles de changer en fonction de l'état physiologique, notamment lors d'une grossesse, où le nombre de MCCs diminue.

Chez l'homme, les MCCs sont uniquement retrouvées au niveau des canaux efférents, et leur fonction reste encore peu étudiée. Le rôle des canaux efférents est de transporter fluide et spermatozoïdes depuis le testicule jusqu'à l'épididyme, tout en augmentant la concentration du fluide par des mécanismes de réabsorption. Une étude récente réalisée chez la souris a mis en évidence un battement des cils particulier, différent de celui que l'on peut observer dans les autres tissus présentant un épithélium multicilié. En effet, contrairement au battement coordonné unidirectionnel des autres tissus, les cellules des canaux efférents semblent battre dans des directions différentes, générant ainsi une force centripète dont le but semble être de permettre le mouvement constant des spermatozoïdes et leur suspension dans le fluide sans créer d'obstruction dans le canal. Les MCCs ne semblent en revanche pas être responsables de l'avancement des spermatozoïdes et du fluide le long des canaux, il s'agirait plutôt de l'action du péristaltisme des muscles lisses environnants (Yuan et al., 2019).

2.2.4. Autre tissus

Au-delà des systèmes décrits précédemment, des MCCs peuvent être observées de façon anormale ou de façon transitoire dans différents tissus, tels que les reins ou l'œsophage.

Système rénal. Il n'est pas anormal d'observer des MCCs dans le système excréteur des amphibiens ou des poissons, mais de manière physiologique, elles sont absentes du système rénal des mammifères. On les retrouve de façon transitoire lors du développement de l'appareil urinaire chez le fœtus, au niveau des tubules du pronephros notamment. Le rôle des MCCs au cours du développement n'a pas encore été élucidé, mais il est probable qu'elles facilitent l'excrétion, comme c'est le cas chez les autres vertébrés. Chez l'adulte, la présence de cils dans le néphron est pathologique, et est associée avec des troubles rares tels que la glomérulonéphrite membranoproliférative, le syndrome néphrotique congénital ou encore le lupus érythémateux systémique (Katz and Morgan, 1984).

Œsophage. Aux alentours de 8 semaines de développement, certaines cellules épithéliales de l'œsophage de l'embryon humain se transforment en MCCs, avant de disparaître à partir de la 17ème semaine. Une partie de ces cellules semblent être à l'origine des glandes sous muqueuses œsophagiennes, mais ces structures ne sont pas retrouvées chez la souris, qui possède également des MCCs lors de son développement embryonnaire. Le rôle de ces cellules reste encore à élucider. Il est à noter que tout comme pour le néphron, des cas pathologiques ont été observés où des MCCs se développaient dans l'œsophage chez des patients présentant une inflammation de longue durée due à des reflux gastro-œsophagiens (Que, 2015).

2.3. Cas des Vertébrés modèles

Chez les métazoaires non mammifères, les MCCs sont trouvées dans un ensemble varié de tissus, nous n'aborderons ici que les cas de deux organismes fréquemment utilisés comme modèles pour l'étude des MCCs : le xénope (*Xenopus laevis*) et le poisson zèbre (*Danio rerio*).

2.3.1. Xénope

Le têtard du xénope est particulièrement intéressant pour l'étude de la multiciliation puisqu'il dispose de plusieurs épithéliums multiciliés, notamment au niveau du système digestif (œsophage et estomac), des fosses nasales, du pronephros, et en moindre quantité au niveau de la trachée. Le plus utilisé pour la recherche reste l'épiderme mucociliaire du têtard, très proche de l'épithélium respiratoire des mammifères à la fois en termes de composition cellulaire, de morphologie et de fonction. Il présente des cellules ciliées et des cellules sécrétrices de mucus, dont les rôles sont d'empêcher l'infection de l'embryon par des bactéries grâce à la production de substances antimicrobiennes, et de permettre une bonne oxygénation de l'organisme par mouvements d'eau autour du têtard.

L'utilisation du xénope comme modèle de la multiciliation est avantageuse car elle permet l'étude *in vivo* du développement des MCCs et la maturation de l'épiderme est très rapide par rapport aux modèles mammifères disponibles. De plus, les différents mécanismes de régulation de la multiciliogénèse observés chez le xénope tendent à être conservés dans l'épithélium respiratoire des

mammifères, ce qui a permis d'étudier de nombreux mécanismes liés à la mise en place de cils multiples tels que les voies de signalisation de la polarité planaire, les cascades transcriptionnelles contrôlant la multiciliogénèse, ou encore les mécanismes d'amplification des corps basaux (Walentek and Quigley, 2017).

2.3.2. *Danio rerio*

Chez le poisson zèbre, on retrouve des MCCs majoritairement dans deux tissus : le pronephros et les placodes olfactives. Les placodes olfactives ne sont généralement pas étudiées dans le cadre de la multiciliation car elles contiennent en grande partie des cellules sensorielles avec des cils multiples de structure « 9+2 » dépourvus de bras de dynéine, et peu de MCCs à proprement parler (Hansen and Zielinski, 2005). Au niveau du pronephros, la majorité des cellules sont mono-ciliées mais il existe une population au niveau du canal pronéphrotique capable de générer jusqu'à 16 cils motiles, dont le rôle pourrait être de faciliter le flux de fluide dans le néphron primitif (Kramer-Zucker et al., 2005).

3. Les cellules multiciliées chez les autres Eucaryotes

Nous avons vu précédemment que les MCCs étaient retrouvées en dehors du règne animal, notamment chez des organismes unicellulaires et au niveau des spermatozoïdes de certaines plantes. Nous décrirons ici succinctement quelques espèces multiciliées.

3.1. Protistes

Les ciliés, tels que *Tetrahymena* et *Paramecium,* sont des protistes ayant longtemps été utilisés comme modèles pour l'étude de la génération des corps basaux et des centrioles. Ils possèdent respectivement plus de 750 et 4000 cils à leur surface, dédiés à la locomotion et à la nutrition, faisant d'eux d'excellents modèles pour l'étude du cil motile. Les corps basaux à l'origine de ces cils ont une structure très similaire à ceux retrouvés chez les mammifères, et leur organisation à la surface de la cellule est coordonnée grâce à des réseaux de fibres. Les études réalisées sur *Tetrahymena* et *Paramecium* ont permis d'identifier différents composants des corps basaux et des cils motiles (Bayless et al., 2019).

L'amibozoaire *Multicilia marina* possède entre 20 et 30 flagelles à sa surface, connectés par leurs corps basaux grâce à des fibres. Curieusement, les mouvements de ces cils ont été caractérisés de faibles et oscillatoires, avec un manque de coordination, conférant à *Multicilia* une locomotion lente et rotatoire, sans direction précise.

3.2. Plantes

Les quelques cas rapportés dans la littérature de multiciliation chez les plantes se présentent de manière variées selon les espèces, avec une organisation et un nombre différent, mais le rôle des cils reste constamment celui de faciliter la locomotion des spermatozoïdes. Ainsi, on retrouve chez les membres du genre *Isoetes* une moyenne de 11 cils s'effilant à leur extrémité le long d'une cellule en forme de spirale d'environ 30 µm de long. Trois d'entre eux sont dirigés vers l'avant tandis que les autres génèrent le mouvement le long du spermatozoïde, mais leur battement ne semble pas coordonné (Yuasa, 1933). Chez les individus du genre *Ginkgo*, les spermatozoïdes présentent à leur

surface plus de 1000 cils disposés sous forme d'une spirale couvrant 15 à 20% de la surface cellulaire. Le spermatozoïde se déplace la spirale en avant, grâce au battement bi-phasique des cils, semblable à celui que l'on peut observer chez les mammifères (Ridge et al., 1997).

4. Régulation de la multiciliogénèse chez les Vertébrés

La régulation de la multiciliogénèse chez les Vertébrés est un processus complexe, et nos connaissances actuelles sur le sujet restent encore parcellaires. De manière intéressante, cette régulation fait appel à de nombreuses protéines impliquées dans la régulation du cycle cellulaire et réutilise un grand nombre de protéines spécifiques de la duplication centrosomale. L'ensemble des informations dont nous disposons ne nous permettent pas d'affirmer avec certitude si les mécanismes impliqués sont partagés entre les différentes espèces voire entre les différents tissus d'une même espèce.

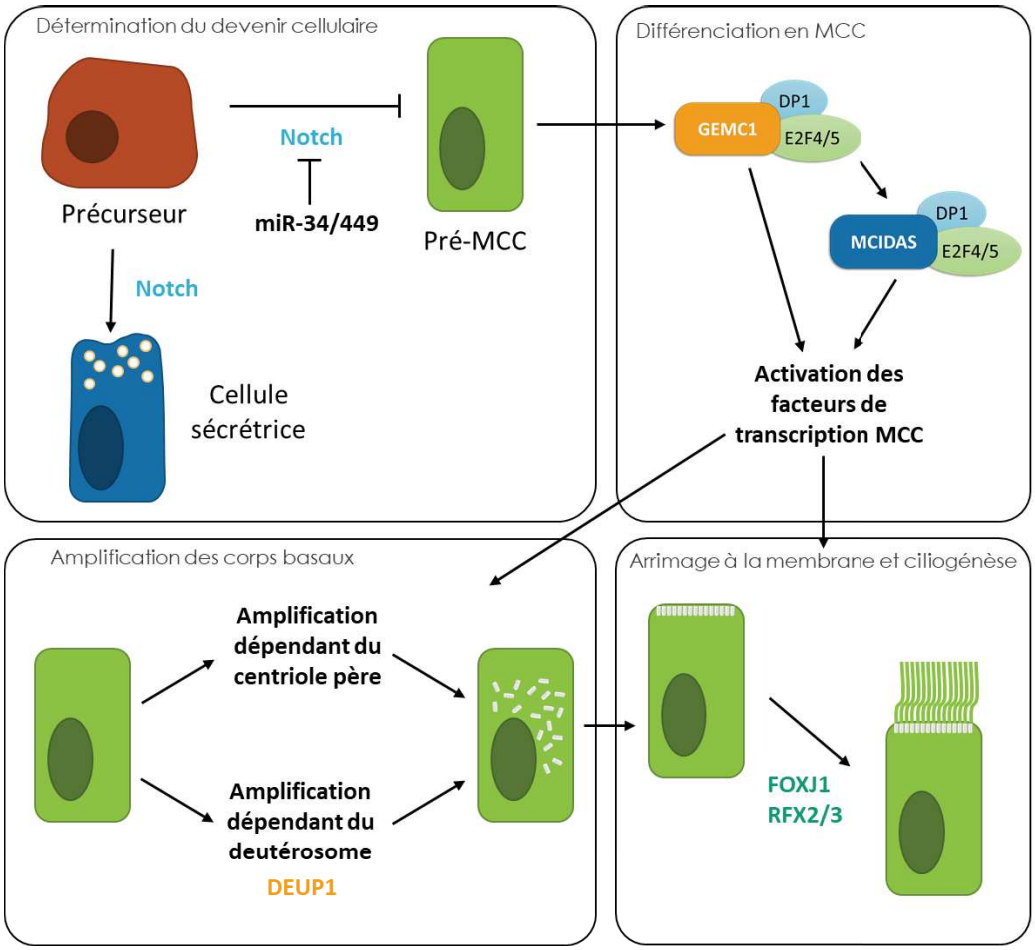

Figure 2-7: Schéma résumant les différentes étapes de la multiciliogénèse. Les flèches à extrémité plate illustrent une répression, une flèche classique représente une activation.

Le processus de multiciliogénèse peut être divisé en plusieurs grandes étapes : l'acquisition de l'identité multiciliée, la spécification du devenir des MCCs, l'amplification des corps basaux et enfin l'arrimage de ces corps basaux à la membrane accompagné de la ciliogénèse (*Figure 2-7*). Nous décrirons ici de façon succincte les différentes étapes de la multiciliogénèse ainsi que les différents acteurs impliqués.

4.1. Détermination du devenir cellulaire

La première étape de la différenciation en MCC est l'acquisition de l'identité multiciliée, celle-ci étant majoritairement régulée par la voie de signalisation Notch et par la famille de microARN miR-34/449.

4.1.1. Progéniteurs

Chaque population de MCCs observées chez les mammifères se développe à partir d'un progéniteur différent, et ce au cours du développement (système nerveux) ou au cours de la vie durant le maintien de l'homéostasie (systèmes respiratoire et reproductif). Des études ont montré que dans le système respiratoire, les MCCs se différencient à partir de cellules basales exprimant p63, tandis que dans le cerveau, les MCCs sont des dérivés des cellules gliales radiaires (Rock et al., 2011; Spassky, 2005). Dans le système reproducteur féminin, l'origine cellulaire des MCCs reste floue, mais des travaux ont montré que les cellules sécrétrices exprimant PAX8 pouvait se différencier en MCC, à la fois au cours du développement embryonnaire et pendant la vie adulte (Ghosh et al., 2017).

4.1.2. micro-ARNs

De nombreuses études réalisées sur différents organismes et dans différent tissus ont mis en évidence un rôle crucial de la famille de micro-ARN (miARNs) miR-34/449 dans la différenciation des MCCs. Des travaux ont montré l'existence d'une surexpression des différents homologues de miR-34/449 dans les MCCs du système respiratoire à la fois chez l'homme et chez la souris, ainsi qu'au niveau du cerveau, du système reproducteur féminin et des canaux efférents chez cette dernière (Marcet et al., 2011; Song et al., 2014; Yuan et al., 2019). miR-34/449 a également été détecté dans l'épiderme de xénope et dans le pronéphros de *Danio rerio* (Wang et al., 2013). L'inactivation de ces miARN dans ces différents tissus est à l'origine d'une large diminution voire d'une absence de MCCs dans les épithéliums concernés, à l'exception notable de l'épithélium du système nerveux de la souris, où les épendymocytes semblent se développer normalement (Yuan et al., 2019). Il a été montré que miR-34/449 permettais d'induire la différenciation en MCC en promouvant la sortie du cycle cellulaire et en inhibant le récepteur Notch1 (Marcet et al., 2011). Certaines études ont également mis en évidence d'autres rôles potentiels de ces miARN, notamment lors de la maturation et l'arrimage des corps basaux, la formation du réseau apical d'actine, ainsi que le maintien du phénotype multicilié au cours du temps, soulignant l'apparente complexité du processus régulatoire de la multiciliogénèse (Chevalier et al., 2015; Song et al., 2014).

Chez les mammifères, on retrouve six homologues dans cette famille, présentant une vraisemblable redondance fonctionnelle : miR-34a, miR-34b, miR-34c, miR-449a, miR-449b et miR-449c. Ils occupent généralement 3 loci génomiques : miR-34a, miR-34b/c et miR-449a/b/c, ce dernier cluster étant localisé au niveau du second intron de son gène hôte *CDC20B*. Le xénope *Xenopus tropicalis* présente une organisation similaire, à la différence près qu'il possède deux clusters miR-34b/c, probablement suite à une duplication du locus. Chez *Danio rerio*, seuls mir-34b et miR-34c ont été retrouvés, alors que d'autres résultats semblent indiquer la présence du cluster miR-449 chez les autres poissons osseux (Lv et al., 2019; Marcet et al., 2011).

4.1.3. Notch

La voie Notch est une cascade de signalisation conservée chez les métazoaires impliquée dans de nombreux processus cellulaires au cours du développement, en particulier dans la différenciation des différentes lignées de cellules. Cette voie de signalisation permet une communication de proximité entre cellules adjacentes basée sur l'interaction entre les récepteurs transmembranaires Notch et leurs ligands, localisés sur des cellules voisines. Ainsi, la fixation du ligand sur le récepteur Notch va induire l'activation de ce dernier grâce à deux clivages successifs, libérant de cette manière un fragment intracytoplasmique (NCID : *Notch Intracellular Domain*) capable de réguler l'expression de certains gènes (Henrique and Schweisguth, 2019).

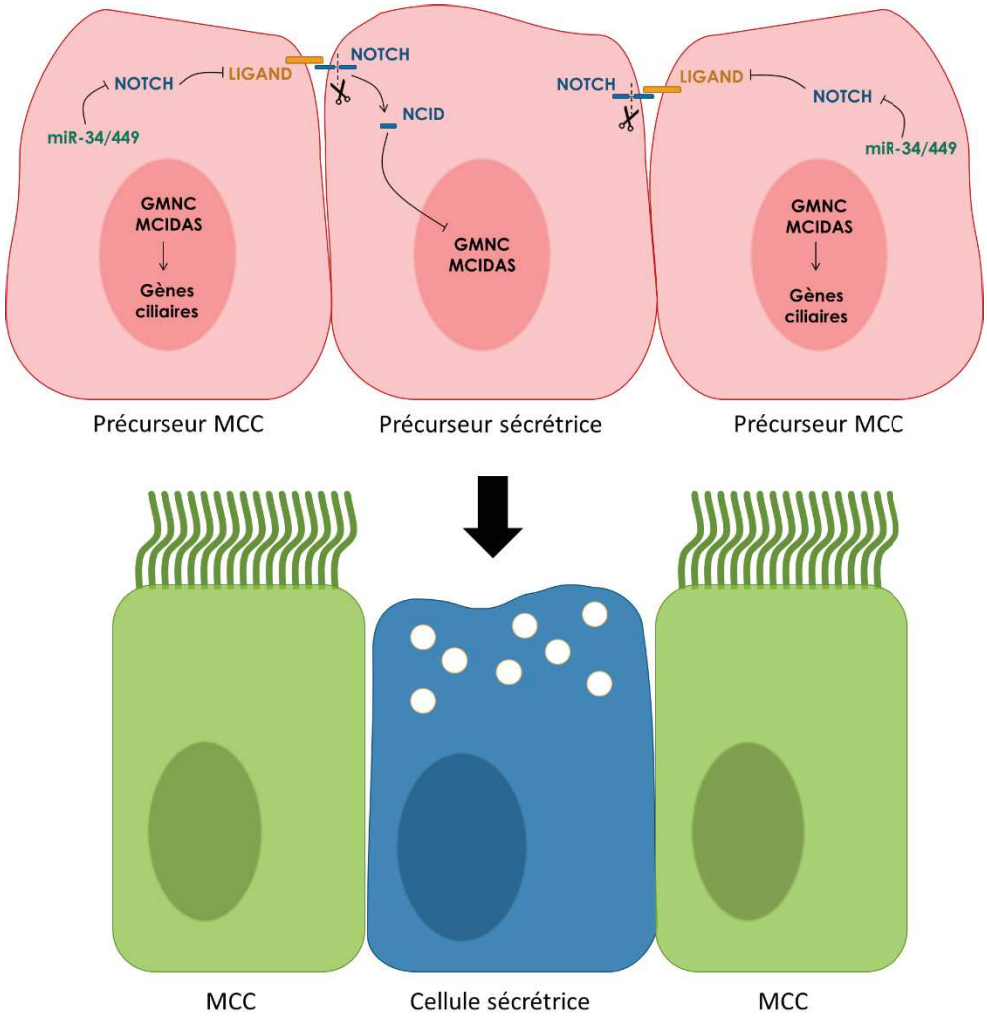

Figure 2-8: Schéma du fonctionnement de l'inhibition latérale de Notch.

Dans les épithéliums multiciliées, des études ont montré que l'absence de Notch dans les précurseurs permettait une différenciation en MCC, et ce dans les systèmes nerveux, respiratoire et reproductif de la souris, dans l'épiderme du xénope, ainsi que dans le pronéphros de *Danio rerio* (Kessler et al., 2015; Kyrousi et al., 2015; Liu et al., 2007b; Marcet et al., 2011). Au niveau de ces tissus, on retrouve une répartition des cellules multiciliées dite en 'poivre et sel', en alternance avec les autres types cellulaires présents, et ce grâce à un mécanisme d'inhibition latérale médié par la voie Notch (*Figure 2-8*). En effet, les ligands de Notch tels que jagged1 ou DLL1 sont exprimés

uniquement dans les MCCs, tandis que Notch est exprimé dans les cellules voisines, son activation inhibant l'expression de ses différents ligands ainsi que les gènes permettant la différenciation en MCCs, *GMNC* et *MCIDAS* (Kyrousi et al., 2015; Liu et al., 2007b). Tout comme les miARN mentionnés précédemment, il semble que les associations ligand/Notch soient variables selon les espèces et les tissus considérés.

4.1.4. STK11

Connu pour son rôle en tant que suppresseur de tumeur, la sérine/thréonine kinase STK11 n'a que récemment été liée à la multiciliogénèse grâce à son expression abondante dans les MCCs en cours de différenciation dans le système respiratoire de souris. Sa suppression dans les progéniteurs de l'épithélium embryonnaire de poumon diminue très largement la présence de MCCs, suggérant un rôle important dans l'établissement du devenir cellulaire et dans le maintien de l'engagement dans la voie multiciliée. Des études approfondies ont permis de mettre en évidence que STK11 contribue au programme multicilié vraisemblablement en inhibant la prolifération cellulaire et en arrêtant le cycle cellulaire. Pour ce faire, STK11 phosphoryle MARK3, dont le rôle est ensuite d'inhiber la cascade de signalisation pro-mitotique ERK1/2. Des résultats de séquençage ARN dans des souris déficientes en STK11 montrent que cette cascade STK11/MARK3/ERK1/2 agit en amont du facteur de transcription MCIDAS, mais sa relation avec GEMC1 reste pour l'instant inconnue (Chu et al., 2019).

4.2. GMNC et MCIDAS, régulateurs centraux de la multiciliogénèse

Une fois l'identité multiciliée spécifiée par l'absence de Notch, la différenciation en MCC se fait à l'aide de deux médiateurs principaux, *GMNC*, codant pour la protéine GEMC1 et *MCIDAS*, codant pour la protéine MCIDAS/Multicilin, chargés d'activer divers facteurs de transcription et gènes nécessaires à la multiciliogénèse. Les protéines codées par ces deux gènes, dont l'expression peut être inhibée par Notch, font partie de la famille des Geminin, caractérisée par la présence d'un domaine *coiled-coil* conservé.

L'inactivation de GEMC1 chez la souris, le xénope et le poisson zèbre est à l'origine d'une absence totale de MCC dans tous les tissus où elles sont normalement retrouvées (Terré et al., 2016; Zhou et al., 2015). De même, l'inactivation de MCIDAS induit une absence de cils multiples à la surface des cellules épithéliales chez la souris et le xénope (Stubbs et al., 2012). Une étude récente a pu mettre en évidence l'existence de deux étapes lors de la spécification et de la différenciation des MCCs chez les mammifères. En effet, GEMC1 va tout d'abord activer MCIDAS ainsi que d'autres facteurs de transcription pour permettre la spécification des MCCs (*Figure 2-9*), puis MCIDAS va activer les gènes requis pour la multiciliogénèse, en particulier ceux impliqués dans l'amplification des corps basaux (Lu et al., 2019).

Pour activer les différents gènes nécessaires à l'établissement de la multiciliation, GEMC1 et MCIDAS, qui ne possèdent pas de domaine d'interaction à l'ADN, doivent s'associer à l'un des facteurs de transcription, E2F4 ou E2F5, ainsi qu'à leur partenaire obligatoire, DP1. Ces interactions se font *via* un domaine TIRT conservé, d'environ 45 acides aminés, se trouvant dans la région C-terminale de MCIDAS et de GEMC1 (Ma et al., 2014; Terré et al., 2016). On peut noter une préférence de GEMC1 pour E2F5, tandis que MCIDAS s'associe à E2F4 et E2F5 avec une affinité semblable (Lu et al., 2019). Ainsi, une fonction cruciale des complexes ternaires EDM (E2F4/5-DP1-

Multicilin) et EDG (E2F4/4-DP1-GEMC1) est l'activation de facteurs de transcriptions tels que MYB et p73, deux autres régulateurs importants, ainsi que CCNO, impliqué dans la régulation du deutérosome, ou encore FOXJ1, RFX2, RFX3 et FOXN4, dont nous détaillerons les rôles plus loin.

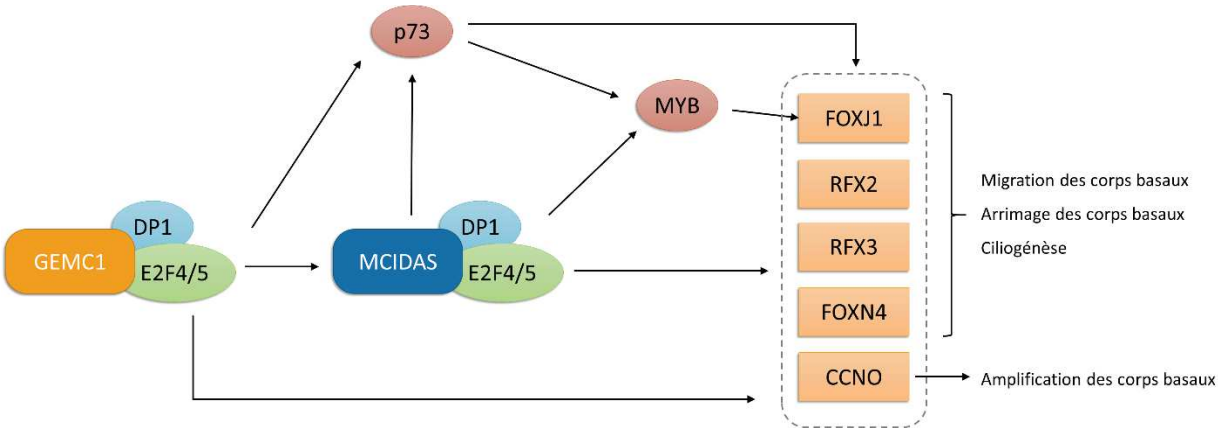

Figure 2-9: Schéma de la régulation de la multiciliogénèse par GEMC1 et MCIDAS

p73 est un facteur de transcription appartement à la famille de p53 dont les rôles ont longtemps été ignorés, mais qui, récemment, a été impliqué dans la régulation des MCCs chez la souris. Activé par GEMC1 et MCIDAS, p73 module directement l'activité de FOXJ1, RFX2, RFX3, miR-34b/c et MYB et est par conséquent un régulateur indispensable à la multiciliogénèse (Marshall et al., 2016; Nemajerova et al., 2016). Un autre facteur de transcription crucial est MYB, dont la place dans le programme de la multiciliation est centrale : activé par MCIDAS et p73, il a pour rôle de contrôler l'expression de FOXJ1 et d'induire l'amplification des centrioles (Tan et al., 2013). De manière intéressante, les différents facteurs de transcription dont nous venons de discuter semblent partager des cibles en commun, posant la question du rôle et de la spécificité de chacun au cours d'un programme transcriptionnel d'apparence redondante, qui restent encore à ce jour à déterminer.

4.3. Amplification des corps basaux

Les MCCs retrouvées chez les Vertébrés possèdent à leur surface jusqu'à 300 cils motiles, qu'il est nécessaire de générer de manière efficace. Il existe ainsi deux modes d'amplification des corps basaux, l'un canonique dépendant du centriole père (*MCD : Mother centriole dependent*) l'autre *de novo via* des structures appelées deutérosomes (*DD : Deuterosome dependent*). La relation entre ces deux modes de création de corps basaux reste encore floue : il était d'abord pensé que les deutérosomes étaient générés par le centriole fils dans les cellules épendymaires (Al Jord et al., 2014), mais des études récentes sur des cellules de trachées de souris ont montré que l'absence des centrioles parents n'inhibait pas la formation de ces structures. De plus, il semblerait qu'en l'absence de l'un ou l'autre mode d'amplification, le second est capable de prendre le relais pour induire un phénotype multicilié normal. De manière encore plus surprenante, il semblerait que l'abolition des deux voies d'amplification jusqu'alors identifiées n'empêche en rien la génération de multiples procentrioles, qui se produit grâce à un mécanisme qui reste encore à élucider, mais pourrait inclure le matériel péricentriolaire (Mercey et al., 2019; Nanjundappa et al., 2019).

4.3.1. Voie d'amplification MCD

La voie d'amplification canonique est d'ordinaire utilisée pour la duplication du centrosome en vue de la mitose, mais sa participation à la génération des corps basaux de MCCs a été mise en évidence à plusieurs reprises. Bien que la majorité des corps basaux soit générée *de novo*, la voie MCD semble être responsable d'environ 10% de la production.

A l'initiation de la duplication, deux protéines centrosomales, Cep63 et Cep152, interagissent et forment une structure en anneau au niveau de l'extrémité proximale du centriole père (*Figure 2-10*). Cep152 recrute ensuite une kinase, Plk4 au niveau du centriole, dont un des rôles semble être de spécifier la position du futur procentriole. Sas-6 et STIL forment ensuite une structure en roue de charrette au niveau de Plk4, qui sert alors de base pour l'assemblage du centriole (Zhao et al., 2013). L'ancrage des microtubules à cette roue est médié par CPAP, qui se fixe à STIL *via* sa région C-terminale et interagit avec les dimères de tubuline *via* sa région N-terminale (Tang et al., 2011; Zheng et al., 2016). Enfin, une fois l'élongation commencée, CCP110 contrôle la longueur des procentrioles en se fixant à leur extrémité lorsqu'ils atteignent la taille désirée pour inhiber leur croissance (Schmidt et al., 2009).

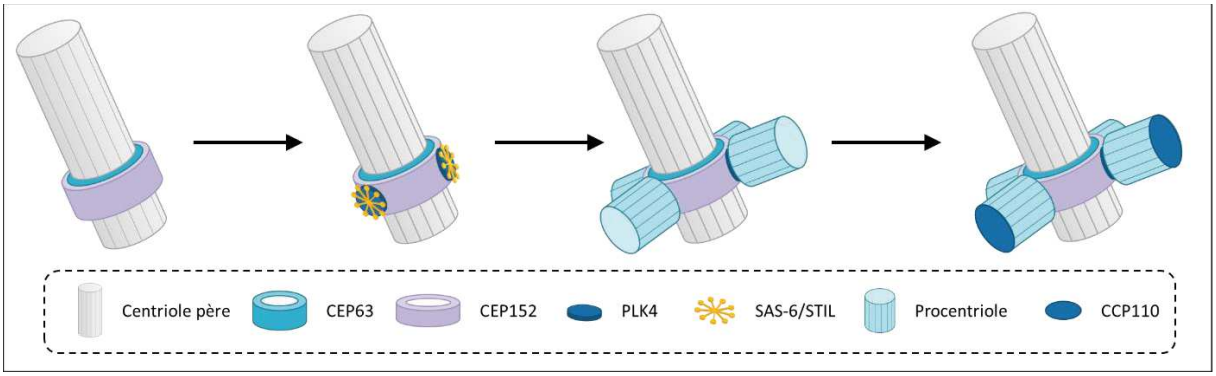

Figure 2-10: Voie d'amplification des corps basaux médiée par le centriole père.

4.3.2. Voie d'amplification DD

L'amplification *de novo* des corps basaux se fait par l'intermédiaire de structures sphériques, transitoires et denses aux électrons appelés deutérosomes, que l'on retrouve uniquement chez les vertébrés. Ces structures ont été observées pour la première fois dans les années 1960, mais ce n'est qu'en 2013 que DEUP1, l'un de leurs composants majeurs, a été identifié (Zhao et al., 2013). DEUP1 est un paralogue de CEP63, lui-même acteur majeur de la voie d'amplification MCD, issu d'une duplication ayant vraisemblablement eu lieu au cours de l'évolution des vertébrés. Il ne s'agit pas là de l'unique ressemblance entre ces deux voies d'amplification, puisque des études ont permis de mettre en évidence que certains composants clés de la voie MCD étaient également requis pour le fonctionnement de la voie DD. En effet, tout comme son paralogue, DEUP1 est capable d'interagir avec CEP152, mais requiert la présence de CCDC78 pour permettre sa localisation au deutérosome. Plk4 et Sas-6 ont également étés identifiés comme nécessaire au fonctionnement des deux voies (Klos Dehring et al., 2013).

D'un point de vue structural, la voie DD est très semblable à la voie MCD : DEUP1 forme une structure en anneau, entourée par CEP152 (*Figure 2-11*). Au niveau de ces deutérosomes, il est

possible d'observer la présence de foyers de Plk4 et de Sas-6, indicateurs de la présence de procentrioles en formation. On retrouve autour des deutérosomes une couronne contenant entre autre les protéines Pericentrin et γ-tubuline, toute deux composants majeurs du matériel péricentriolaire, ainsi que CDC20B, gène hôte des miR-449 mentionnés précédemment. Ce dernier semble avoir un rôle dans la procédure de désengagement des corps basaux du deutérosome, comme a pu le montrer une étude récente (Revinski et al., 2018).

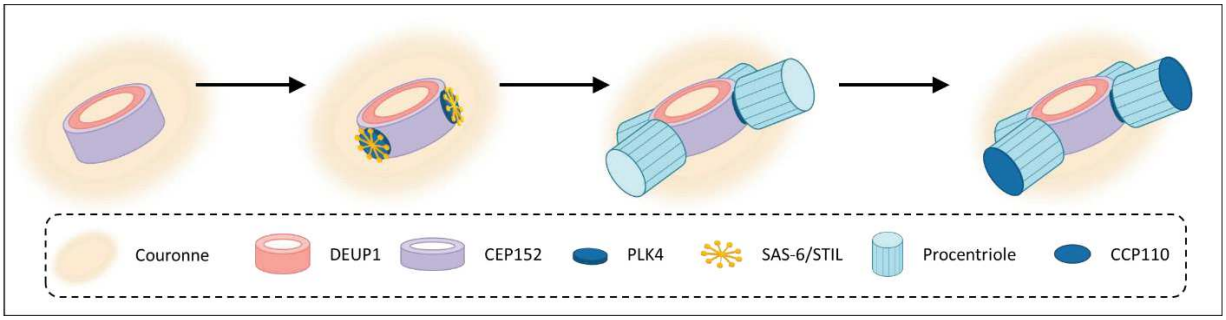

Figure 2-11: Voie d'amplification des corps basaux dépendant du deutérosome.

En ce qui concerne la régulation de cette voie d'amplification, il a été montré qu'une surexpression de MCIDAS, mais pas de GEMC1 pouvait induire une surexpression de DEUP1, CCNO et CDC20B, témoignant ainsi d'une spécificité du rôle régulatoire de MCIDAS envers la génération massive de corps basaux (Lu et al., 2019). La nature de l'interaction entre MCIDAS et DEUP1 est encore floue, mais des résultats indiquent que la présence de CCNO est nécessaire à la formation adéquate des deutérosomes, en termes de nombre, taille et morphologie, bien que son mode d'action reste, lui aussi, indéterminé (Funk et al., 2015).

4.3.3. Désengagement des corps basaux

L'ultime étape de l'assemblage des corps basaux avant qu'ils ne puissent être transportés jusqu'à la membrane est le désengagement des procentrioles, phase encore peu étudiée du processus. A l'heure actuelle, seules quelques protéines ont été identifiées comme participant à la libération des centrioles fixés aux deutérosomes : CDC20B, PLK1, SPAG5 et Separase. Seul CDC20B semble être spécifique au processus de désengagement impliquant les deutérosomes, les trois autres protéines étant impliquées dans la duplication des centrioles au moment de la mitose. Tout comme pour les deux voies d'amplification des corps basaux, il semblerait que la multiciliogénèse réemploie la machinerie déjà en place servant à la duplication du centrosome en l'adaptant spécifiquement à ses besoins à l'aide d'un ou plusieurs gènes, en l'occurrence CDC20B. Ainsi, il n'est pas surprenant qu'une étude montre l'implication de diverses protéines du cycle cellulaire habituellement employées lors de la mitose, telles que le complexe APC/C, dans le processus de désengagement des corps basaux dans les MCCs (Al Jord et al., 2017).

Le mode de fonctionnement exact du désengagement et en particulier du complexe CDC20B/PLK1/SPAG5 et Separase reste encore à déterminer, mais des résultats préliminaires suggèrent que CDC20B interagit avec PLK1, une kinase dont une des cibles est SPAG5. Ce dernier peut ensuite activer ou inhiber la protéase Separase selon son état de phosphorylation (Revinski et al., 2018).

4.4. Arrimage à la membrane et ciliogénèse

Une fois les centrioles formés et désengagés, ceux-ci vont migrer vers la membrane apicale en se différenciant en corps basaux, s'y ancrer à l'aide d'un réseau d'actine et démarrer leur ciliogénèse. Le déplacement des centrioles jusqu'à la membrane dépend d'une interaction actine/myosine, une étude sur les oviductes de cailles ayant montré que l'inhibition pharmacologique de la polymérisation des filaments d'actine empêchait la migration des corps basaux, tout comme l'inhibition des interactions actine/myosine (Lemullois et al., 1988).

L'un des premiers facteurs de transcription identifiés comme étant indispensable à la multiciliation est FOXJ1, dont les gènes cibles sont nombreux, et dont l'activité est nécessaire à la migration des corps basaux et à leur arrimage, ainsi qu'à la ciliogénèse en elle-même. En effet, l'activité de FOXJ1 permet de réguler la mise en place du réseau apical d'actine, indispensable à l'arrimage des corps basaux, *via* la localisation de l'ezrine au niveau apical, et l'activation de la GTPase RhoA (Pan et al., 2007; You et al., 2004). L'ezrine est localisée au niveau des corps basaux, et son rôle est d'accompagner la migration des centrioles et leur arrimage, tandis que RhoA a pour but d'organiser les filaments d'actines au niveau de la membrane apicale (Epting et al., 2015).

Par ailleurs, plusieurs études ont montré que l'activité de FOXJ1 pouvait être modulée par les facteurs de transcriptions RFX2 et RFX3, et que son activité semblait être redondante avec celle de FOXN4 (Didon et al., 2013; Quigley and Kintner, 2017). Ensemble, ces 4 facteurs de transcription contrôlent l'expression d'un grand nombre de gènes impliqués dans l'arrimage des corps basaux ainsi que dans la formation des cils motiles, bien qu'ils aient chacun leur spécificité. Ainsi, RFX2 est un régulateur central de la voie PCP, rôle qui n'a pour l'instant pas été attribué à son homologue RFX3, malgré un évident chevauchement de leurs cibles (Chung et al., 2014). De même, FOXN4 présente des cibles remarquablement similaires à celles de FOXJ1, mais très peu sont uniquement spécifiques à FOXN4. De plus, FOXJ1 semble avoir une expression continue dans les MCCs, très certainement dans un but régulatoire, tandis que l'expression de FOXN4 reste transitoire et spécifique à la phase de formation des cils motiles (Campbell et al., 2016).

5. Pathologies de la multiciliation

Nous avons vu précédemment que le dysfonctionnement des cils motiles était à l'origine de pathologies connues sous le nom de dyskinésies ciliaires primitives (PCD : *Primary ciliary dyskinesia*), caractérisée par des anomalies de la motilité des cils. Il existe une classe spécifique de ciliopathies liées à des défauts de la multiciliation que l'on appelle RGMC (*Reduced generation of multiple motile cilia*), et qui se présentent sous la forme d'une hypoplasie ou d'une aplasie ciliaire. Elle concerne 1 à 6 % des patients atteints de PCD et en partage la majorité des symptômes, bien que les causes génétiques ne soient pas encore totalement élucidées.

5.1. Physiopathologie

L'atteinte des MCCs chez les mammifères, et en particulier chez l'homme, va être responsable d'une variété de symptômes pathologiques atteignant les voies respiratoires, la circulation du liquide cérébrospinal ainsi que le transport des gamètes dans les tractus génitaux féminin et masculin. Selon l'importance des gènes impliqués, les symptômes développés peuvent être plus ou moins prononcés et présenter un danger pour l'individu.

Atteintes respiratoires. L'un des symptômes majeurs retrouvés chez les patients atteints de pathologies des cellules ciliées est le défaut de clairance mucociliaire, qui se traduit par des infections respiratoires dès l'enfance. En effet, l'inhibition du flux de mucus peut causer des obturations dans les plus petits conduits respiratoires, créant ainsi des environnements propices au développement des bactéries piégées dans le mucus. La chronicité de ces infections provoque, à terme, une hypertrophie des glandes à mucus ainsi que des dommages au niveau de l'épithélium, réduisant d'autant plus la clairance. La quasi-totalité des patients développe au cours de leur vie une bronchectasie (dilatation des bronches) avec une réduction des capacités pulmonaires, qui, dans les cas les plus graves, peut nécessiter une transplantation de poumons ou mener à un décès prématuré (Knowles et al., 2013a; Munkholm and Mortensen, 2014).

Hydrocéphalie. Cette pathologie se définit par l'accumulation de liquide cérébrospinal dans les ventricules, provoquant ainsi une augmentation de la pression et une dilatation des ventricules ayant des répercutions diverses sur le cerveau. Beaucoup plus rare que les atteintes respiratoires, l'hydrocéphalie a néanmoins été associée à des cas d'aplasie ciliaire par défaut de circulation du liquide cérébrospinal (Barlocco et al., 1991; Berlucchi et al., 2012). La prévalence semble être moindre chez les humains affectés par des RGMC que chez les petits mammifères, telles que les souris, les cils jouant un rôle plus ou moins important dans la circulation du fluide cérébrospinal selon les espèces (Spassky and Meunier, 2017).

Infertilité. Contrairement aux PCD, les RGMC n'affectent pas la motilité des spermatozoïdes, en revanche, elles peuvent impacter la circulation de ces derniers dans les canaux efférents. Des études réalisées chez la souris ont montré que, malgré la présence de spermatozoïdes fonctionnels, le dysfonctionnement des MCCs provoque une stérilité due à une obstruction des canaux efférents (Yuan et al., 2019). Chez l'homme, un cas similaire a été décrit chez un jeune homme présentant une aplasie ciliaire associée à des infections respiratoires chroniques, un *situs inversus* ainsi qu'une azoospermie malgré des spermatozoïdes normaux (Matwijiw et al., 1987).

Chez la femme, l'infertilité dans le cadre d'une RGMC a peu été étudiée : deux cas ont été rapportés où les patientes ont fait appel à une fertilisation *in vitro* pour parvenir à une grossesse, mais les patientes sont généralement trop jeunes pour permettre des analyses de fertilité (Amirav et al., 2016; Wallmeier et al., 2014). De manière moins spécifique, il a été montré que les PCD pouvaient induire une stérilité ou une fertilité amoindrie, bien que ce symptôme ne soit pas systématique et que certaines patientes présentent une fertilité normale malgré une absence de motilité des cils (Jean et al., 1979). Curieusement, un cas diagnostiqué comme étant une PCD mais présentant une hypoplasie évidente a été associé à une infertilité, malgré la présence d'un battement lent des cils présents, soulevant la question de l'importance relative du nombre de cils par rapport à leur motilité dans le cas du transport des ovules dans le tractus féminin (Halbert et al., 1997).

5.2. Causes génétiques

A l'heure actuelle, très peu de gènes ont été identifiés comme étant responsables de RGMC : MCIDAS, CCNO, et potentiellement FOXJ1, bien que son dysfonctionnement atteigne le cil nodal en plus des MCCs, symptôme plus suggestif d'une PCD. Malgré le fait qu'aucun rôle n'a jusqu'alors été

attribué à ce gène dans le processus de multiciliation, une étude récente a mis en évidence l'implication de NEK10 dans des pathologies rappelant les RGMC.

MCIDAS. Après la découverte d'une mutation homozygote de MCIDAS chez un patient présentant une aplasie ciliaire, Boon et collaborateurs ont séquencé les séquences exoniques de MCIDAS chez 59 familles présentant une hypoplasie, résultant en une identification d'une mutation bi-allélique chez 8 individus. Tous ces individus présentaient un défaut de clairance mucociliaire associé à des infections respiratoires chroniques et une réduction de la fonction pulmonaire, allant parfois jusqu'au décès de la personne. Certaines des mutations identifiées sont localisées au niveau du domaine TIRT, servant à l'interaction avec E2F4/E2F5, suggérant ainsi une perte de fonction de MCIDAS dans ces cas (Boon et al., 2014).

CCNO. Le séquençage de l'exome de plusieurs membres d'une famille consanguine présentant des infections respiratoires chroniques a permis d'identifier des mutations homozygotes de type perte de fonction chez CCNO. Tous les individus souffraient de difficultés respiratoires dues à des infections, avec une bronchectasie précoce et dans deux cas graves une transplantation de poumon a été nécessaire. Sur un total de 16 patients diagnostiqués, 12 présentaient une détresse respiratoire à la naissance, une présentait une stérilité, mais aucun cas de *situs inversus* n'a été recensé. En moyenne, seuls 1 à 2 cils ont été détectés par MCC (Wallmeier et al., 2014).

FOXJ1. L'analyse génétique d'une cohorte de 6 individus présentant un défaut de clairance mucociliaire associé à une hydrocéphalie a permis l'identification d'une mutation hétérozygote de type perte de fonction dans le gène de FOXJ1 chez 3 patients sans lien de parenté, puis chez 3 autres patients issus de différentes cohortes. Sur les 6 individus, tous présentent une hydrocéphalie et des troubles infectieux de l'appareil respiratoire, et la moitié d'entre eux ont un *situs inversus*, indiquant une influence de FOXJ1 sur le fonctionnement du cil nodal. Le nombre de cils retrouvés à la surface des MCCs de ces patients varie de 0 à un nombre presque normal, la majorité des cellules étant tout de même hypoplasiques malgré un nombre vraisemblablement normal de corps basaux retrouvés dans le cytoplasme. Ces résultats suggèrent donc qu'un défaut d'expression de FOXJ1 induit un défaut d'arrimage des corps basaux, et l'existence d'une éventuelle haploinsuffisance, au vu du statut hétérozygote des patients et de la variabilité des phénotypes cellulaires observés (Wallmeier et al., 2019).

NEK10. Le diagnostic chez une patiente présentant une insuffisance respiratoire accompagnée d'une bronchectasie d'une mutation homozygote dans la séquence de NEK10 a permis d'identifier la cause génétique chez six autre patients dont la dilatation des bronches n'était pas expliquée. De manière intéressante, NEK10 n'a jusqu'ici jamais été associé à la multiciliation, mais des analyses approfondies ont mis en évidence une participation de NEK10 dans la ciliogénèse de par son activité de kinase. En effet, un défaut de NEK10 se traduit par la présence de cils raccourcis à mouvement réduit, malgré une fréquence de battement normale, ayant pour conséquence une diminution de la clairance mucociliaire (Chivukula et al., 2020). D'autres travaux seront en revanche nécessaires pour évaluer avec précision la place de NEK10 dans le programme de la multiciliogénèse.

Il apparait ici évident que la multiciliation est un processus très complexe, impliquant de nombreux gènes, et dont nous n'avons encore qu'une vision très parcellaire. Malgré un nombre

croissant de travaux sur le sujet, beaucoup de questions ont toujours besoin de réponses ; des acteurs restent encore à identifier, et le rôle de certains est à préciser. Bon nombre des progrès réalisés récemment dans la connaissance du cil et de la multiciliation sont basés sur la génomique au sens large du terme, grâce à des approches telles que la génomique comparative, la génomique médicale, la génomique fonctionnelle et plus généralement l'ensemble des 'omiques'. Ces différentes disciplines, qui ont révolutionné la biologie ces dernières décennies, seront présentées dans leurs grandes lignes dans le chapitre suivant ainsi que la manière dont elles ont contribué à notre connaissance du cil et de la multiciliation.

Chapitre 3 : La génomique à l'ère des approches intégratives

C'est en 1987 qu'apparait pour la première fois le terme 'génomique', suite à l'engouement provoqué par l'arrivée des méthodes de séquençage ADN de première génération en 1977. Il désigne la discipline regroupant la cartographie et le séquençage de l'ADN, la caractérisation fonctionnelle ainsi que l'analyse de ces informations (McKusick and Ruddle, 1987). Jusqu'en 2005, la méthode de séquençage de Sanger reste la plus utilisée, mais ses limitations ont poussé à la mise au point de nouvelles démarches, plus adaptées au séquençage routinier. C'est à ce moment-là qu'émergent les techniques de séquençage dites de « nouvelle génération » (NGS : *Next Generation Sequencing*), dont la particularité est de permettre de générer un volume de données considérable à moindre coût (Metzker, 2010).

Les avancées technologiques de ces dernières années sont donc à l'origine d'une augmentation exponentielle de la quantité de données biologiques disponibles à l'exploitation, ce qui a donné lieu à l'avènement des disciplines à haut-débit, connues sous le nom de '-omiques'. Depuis lors, de nombreuses disciplines ont vu le jour pour faciliter l'analyse de systèmes biologiques, notamment par l'étude des transcrits, des protéines, des modifications épigénétiques, ou encore des interactions entre les molécules biologiques. De plus en plus souvent, ces différents domaines sont combinés par des approches dites 'intégratives' pour permettre l'étude des systèmes biologiques dans leurs ensembles.

Dans ce chapitre, nous décrirons de manière succincte les différentes branches de la génomique et la manière dont ces différentes approches ont contribué à une meilleure compréhension des processus régissant la mise en place du cil et de la multiciliation. Nous verrons ensuite comment il est possible de combiner l'ensemble de ces techniques pour caractériser de manière approfondie le fonctionnement des gènes.

1. Génomique comparative

La génomique comparative est une branche de la génomique dédiée à la comparaison des génomes de plusieurs espèces, à la fois en termes de contenu en gènes, d'organisation, de structure et de séquence. Elle a pour but d'aider à la compréhension des processus évolutifs ayant contribué à la biodiversité actuelle, ainsi qu'à la caractérisation des fonctions des différents éléments génomiques étudiés. Cette discipline a récemment pu se développer grâce aux efforts réalisés pour améliorer les techniques de séquençage d'ADN, qui sont à l'origine d'une hausse exponentielle des projets de séquençage de génomes disponibles dans les bases de données publiques. Ainsi, la base de données GOLD (*Genome Online Database*) recense aujourd'hui près de 350 000 projets de séquençage, répartis entre les bactéries, les eucaryotes, les archées, les virus et des études de métagénomique (Mukherjee et al., 2019) (*Figure 3-1*). Bien que de nombreux projets parmi ceux-ci soient incomplets et gardent le statut d'ébauche permanente, la diversité qu'ils représentent permet d'ores et déjà d'entreprendre des études de grande ampleur pour améliorer notre connaissance du monde vivant.

Dans cette partie, nous aborderons le principe fondamental de la génomique comparative qu'est l'homologie, ainsi que les différentes façons dont il est possible de s'appuyer sur ce concept pour comparer les génomes de plusieurs espèces. Ces comparaisons peuvent se faire à plusieurs niveaux : au niveau du répertoire de gènes, au niveau de l'organisation des gènes, et au niveau de la séquence de chaque gène. Nous verrons brièvement chacune de ces approches dans les parties suivantes ainsi que la manière dont elles ont pu contribuer à l'analyse du cil et de la multiciliation. L'ensemble des notions abordées ici sont décrites de manière succincte, pour une description plus approfondie voir le chapitre de livre intitulé *Orthology: promises and challenges* (Nevers et al., In press) auquel j'ai contribué, présenté en annexe.

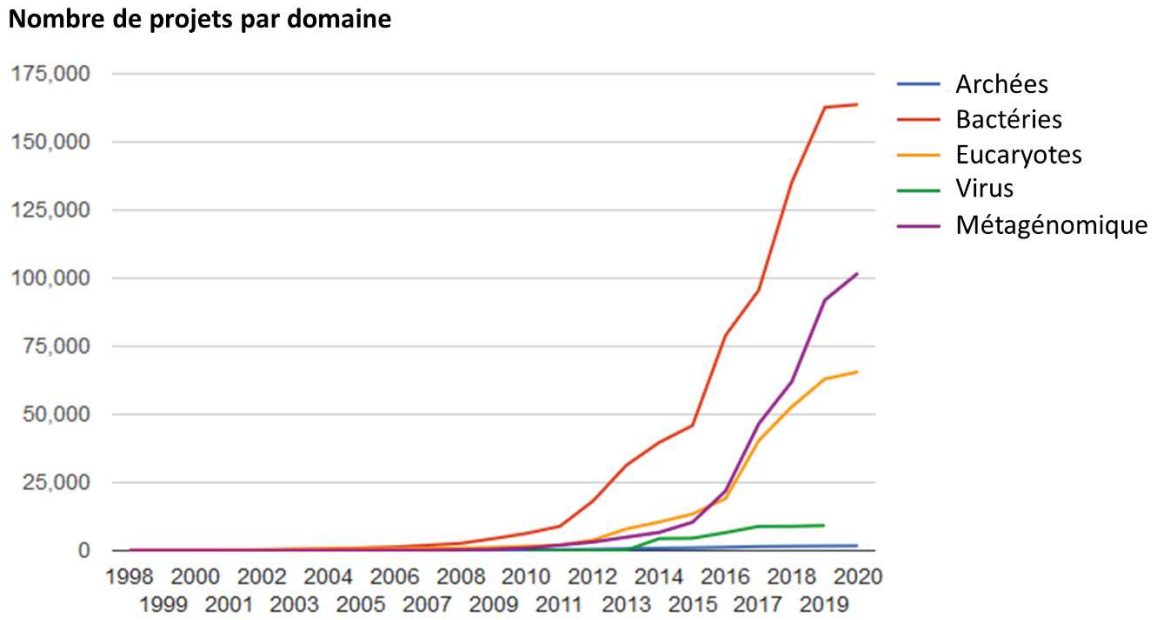

Figure 3-1: Nombre de projets de séquençage par domaine du Vivant et par année sur la base de données GOLD. Tiré de https://gold.jgi.doe.gov/statistics (Page consultée le 21/06/2020)

1.1. L'homologie, principe fondamental de la génomique comparative

Pour permettre la comparaison de deux entités, il est tout d'abord nécessaire de définir la relation liant les différents éléments que l'on souhaite comparer ; c'est là qu'intervient le principe d'homologie. Utilisée en premier lieu pour décrire les similarités entre deux caractères physiques, l'homologie désigne la similarité existant entre deux traits partageant un lien évolutif direct. Au contraire, on parle d'analogie lorsque deux traits similaires ont évolué de manière indépendante. Dans le contexte de la génomique comparative et des séquences d'ADN, on parle ainsi de deux gènes homologues lorsque ceux-ci partagent un même ancêtre commun.

Ce principe d'homologie est central aux études de génomique comparative, puisqu'il permet de mettre en évidence, *via* comparaison, les éléments conservés et les éléments modifiés, acquis ou perdus au cours de l'évolution, donnant ainsi un premier aperçu des informations génétiques indispensables à la vie.

1.1.1. Notions d'orthologie et de paralogie

Depuis 1970, le principe d'homologie est séparé en deux notions distinctes, permettant de mieux caractériser les évènements évolutifs ayant mené à l'existence des gènes comparés (Fitch, 1970). On parle ainsi d'orthologie lorsque deux gènes sont issus d'un évènement de spéciation et sont retrouvés chez deux organismes différents, tandis que la paralogie désigne deux gènes séparés par un évènement de duplication (*Figure 3-2*).

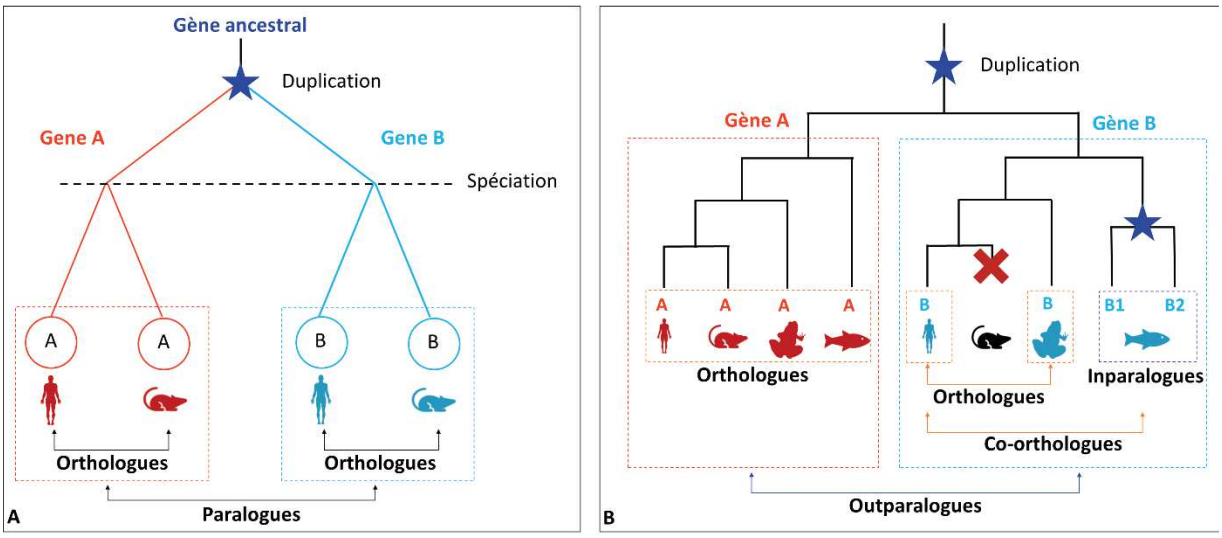

Figure 3-2: Représentation schématique des relations d'homologie. ***A. Illustration des notions d'orthologie et de paralogie.*** *Les gènes A retrouvés chez la souris et l'homme sont orthologues entre eux puisque séparés par un évènement de spéciation. Il en est de même pour les gènes B chez ces espèces. Les gènes A et B sont paralogues entre eux puisqu'ils sont issus d'un évènement de duplication.* ***B. Illustration des notions d'inparalogie et d'outparalogie.*** *Pour les espèces considérées, l'évènement de duplication à l'origine des gènes A et B s'est produit avant les différents évènements de spéciation, les gènes A et B sont dont outparalogues entre eux. Au cours de l'évolution du gène B, une duplication a eu lieu chez les poissons après un évènement de spéciation, B1 et B2 sont donc inparalogues entre eux, et co-orthologues avec les gènes B de l'homme et de la grenouille. Adapté de (Nevers et al., In press).*

D'un point de vue fonctionnel, il est largement supposé que les orthologues tendent à conserver la même fonction tandis que les paralogues tendent à développer de nouvelles fonctions (Koonin, 2005). Cette 'conjecture de l'orthologie', bien qu'elle ne soit pas avérée dans tous les cas, est fréquemment utilisée pour prédire la fonction de gènes ou pour annoter les séquences de génomes nouvellement séquencés. Il est néanmoins important de considérer à la fois la distance évolutive séparant deux espèces ainsi que l'ensemble des évènements de duplication et de spéciation à l'origine des gènes comparés. En effet, deux orthologues retrouvés chez des espèces dont la séparation est « ancienne » peuvent présenter des fonctions différentes, tandis que deux gènes paralogues issus d'une duplication « récente » sont susceptibles de garder une fonction similaire.

Pour permettre d'évaluer plus simplement ces distances dans les relations d'homologie, la définition de Fitch de 1970 a été complétée par l'introduction de deux nouveaux termes : inparalogie et outparalogie (Sonnhammer and Koonin, 2002). Ces deux définitions ont pour vocation de caractériser l'ordre des évènements de duplication à l'origine de paralogues par rapport à un évènement de spéciation donné. On parle ainsi d'outparalogie lorsque l'évènement de duplication précède la spéciation, et d'inparalogie lorsque l'évènement de duplication succède à la spéciation

(*Figure 3-2*). Ces deux notions nécessitent également la définition du terme « co-orthologues », qui désigne la relation entre des inparalogues et les autres gènes issus de l'évènement de spéciation considéré.

1.1.2. Prédiction des relations d'orthologie

De manière générale, la génomique comparative repose essentiellement sur les comparaisons entre orthologues, il est donc nécessaire d'avoir préalablement défini au mieux les relations d'homologie. Il existe à ce jour un grand nombre de méthodes visant à prédire les relations d'orthologie avec toujours plus de précision, nous ne détaillerons ici que les principes de base sur lesquels s'appuient la majorité des méthodes. Brièvement, il existe deux types principaux de méthodes pour prédire des relations d'orthologie : celles basées sur les graphes, et celles basées sur les arbres.

Dans les méthodes basées sur les graphes, les gènes sont représentés par des nœuds, tandis que leurs relations évolutives sont représentées par les arrêtes. La notion principale sur laquelle se basent ces prédictions d'orthologie est la similarité de séquence. En effet, deux gènes orthologues étant issus du même ancêtre, ils doivent en théorie présenter une haute similarité de séquence. Ainsi, lors d'une recherche de similarité entre deux espèces, on émet l'hypothèse selon laquelle deux orthologues auront plus de similarité entre eux qu'avec tous les autres gènes présents dans ces espèces, bien que dans les faits ces relations soient parfois plus complexes. Les graphes sont alors construits par la comparaison des gènes de deux génomes complets, et des groupes d'orthologues peuvent être prédits par *clustering*.

Les méthodes basées sur les arbres reposent sur l'histoire évolutive des gènes et la réconciliation de l'arbre du gène avec l'arbre des espèces considérées. Un arbre est tout d'abord généré à partir d'un ensemble de séquences homologues, puis chaque nœud de l'arbre est annoté comme étant un évènement de spéciation ou de duplication, selon une comparaison à l'arbre des espèces. Il est ensuite possible d'annoter les relations de paralogie et d'orthologie entre les différents gènes présents dans l'arbre. Bien que plus précises et informatives que les graphes, ces méthodes de prédiction sont très couteuses en ressources informatiques, et sont donc généralement appliquées à de plus petits jeux de données.

1.2. Comparaison de génomes complets

La comparaison de génomes complets est une approche majeure de la génomique comparative qui permet de s'appuyer sur les relations d'homologie pour essayer d'améliorer notre compréhension des mécanismes évolutifs et des relations génotype/phénotype. Cette approche se base sur la comparaison de répertoires de gènes à plusieurs échelles, et nous verrons ici qu'elle a notamment permis à plusieurs reprises d'identifier des gènes impliqués dans le cil.

1.2.1. Comparaison de répertoires de gènes

La comparaison de répertoires de gènes entre deux ou plusieurs espèces se base sur l'hypothèse que des gènes conservés entre les espèces proviennent du génome de leur ancêtre commun et sont responsables des phénotypes communs de ces espèces, tandis que les gènes

variables sont spécifiques à l'histoire évolutive de chaque organisme. Il est donc possible de mettre en évidence les gènes conservés au cours de l'évolution et nécessaires aux fonctions de base des espèces vivantes, ainsi que les gènes spécifiques à chaque espèce ou clade considérés. De cette manière, la comparaison des génomes de *Saccharomyces cerevisiae, Caenorhabditis elegans* et *Drosophila melanogaster* a montré que 20% des gènes codant pour des protéines de la drosophile présentaient des orthologues à la fois chez le nématode et chez la levure, bien que la séparation de ces espèces soit très ancienne (Rubin et al., 2000).

Il est également possible d'utiliser la comparaison de répertoires de gènes pour étudier un phénotype spécifique grâce aux corrélations génotype/phénotype *via* une approche soustractive. Le principe est le suivant : la comparaison se fait entre au moins 3 espèces, parmi lesquelles deux sont éloignées et présentent le phénotype d'intérêt, et la troisième ne présente pas le phénotype et est proche taxonomiquement d'une des deux autres espèces. La théorie veut alors que les gènes communs entre les deux premières espèces et absents de la dernière soient potentiellement liés au phénotype étudié.

Le développement des techniques de séquençage de nouvelle génération a permis d'augmenter considérablement le catalogue d'espèces disponibles pour les comparaisons de génomes, créant ainsi l'opportunité d'étudier les relations génotype/phénotype avec plus de précision en comparant un plus grand nombre d'espèces à la fois, grâce à la méthode du profilage phylogénétique.

1.2.2. Profilage phylogénétique

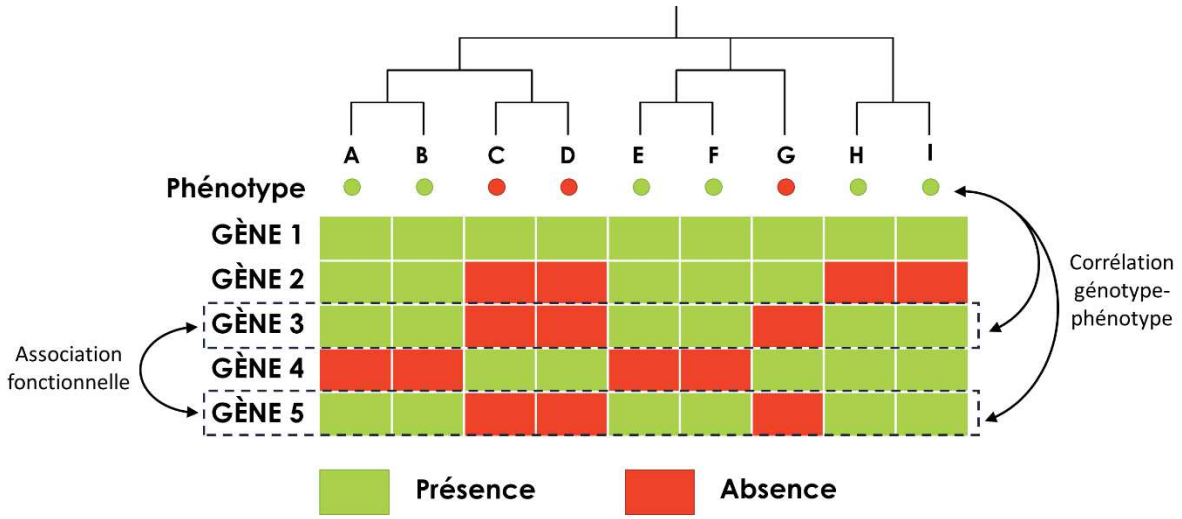

Figure 3-3: Profilage phylogénétique. Profils de présence et d'absence d'orthologues de 5 gènes dans 9 espèces (A-I) pour lesquelles la présence et l'absence d'un phénotype d'intérêt est connue. Les gènes 3 et 5 présentent une distribution similaire, on suppose donc une association fonctionnelle entre eux. Ils présentent également la même distribution que le phénotype d'intérêt, on suppose alors une corrélation entre ces gènes et le phénotype.

Le profilage phylogénétique se base sur l'établissement de profils de présence et d'absence d'orthologues de gènes dans un ensemble d'espèces, et a pour but d'inférer des relations fonctionnelles entre gènes, ou d'associer un ou plusieurs gènes à un processus spécifique. L'hypothèse sous-jacente est que des gènes fonctionnellement liés évoluent de façon similaire (Pellegrini et al., 1999). Ainsi, leurs orthologues devraient être présents et absents dans les mêmes espèces. De même, des gènes associés à un phénotype devraient être retrouvés spécifiquement dans

des espèces présentant ce phénotype (*Figure 3-3*). Tout comme pour l'approche soustractive à petite échelle vue au point précédent, l'étude des relations génotype/phénotype par le profilage phylogénétique se base sur l'exploitation de pertes spécifiques dans certaines espèces.

1.3. Comparaison de l'organisation génomique

Chez les procaryotes, il n'est pas surprenant de voir des gènes liés fonctionnellement regroupés sous la forme d'opéron. Bien que ces structures ne soient que très peu retrouvées chez les eucaryotes, il a néanmoins été montré à plusieurs reprises que l'ordre des gènes n'était pas complètement aléatoire chez ces espèces (Hurst et al., 2004). En effet, des études menées chez plusieurs organismes modèles eucaryotes ont montré que des gènes impliqués dans la même voie de signalisation avaient tendance à se regrouper dans le génome (Lee and Sonnhammer, 2003). De même, il est possible d'observer une co-localisation de gènes présentant une expression similaire, comme par exemple des gènes exprimés au même stade du cycle cellulaire (Cho et al., 1998).

Lors de la comparaison de plusieurs génomes, il peut donc être intéressant de comparer l'organisation des gènes pour mettre en évidence des régions conservées pendant l'évolution. La notion de synténie désigne la conservation de l'ordre des gènes entre deux ou plusieurs espèces, et peut traduire une association fonctionnelle de ces gènes ou un mécanisme de régulation commun.

1.4. Comparaison de séquences homologues

Nous avons vu précédemment que l'exploitation des relations d'homologie permettait de caractériser fonctionnellement des gènes, d'établir des liens entre un phénotype et un gène, ou encore d'inférer des interactions entre plusieurs gènes, mais elle permet également de caractériser chaque gène de manière beaucoup plus précise, et ce par l'étude des familles de gènes ou de protéines.

La comparaison de séquences homologues peut être réalisée à plusieurs niveaux : elle peut concerner des régions génomiques étendues, tout comme elle peut concerner des séquences individuelles. Cette comparaison se fait à l'aide d'alignements multiple de séquences (MSA : *Multiple Sequence Alignment*), où chaque séquence est alignée de façon à ce que chaque nucléotide ou résidu dérivant de la même position sur la séquence ancestrale soit aligné dans la même colonne. De ce fait, l'exploitation des alignements multiples de familles d'homologues permet de replacer les gènes dans leur contexte évolutif tout en obtenant des informations pertinentes sur leurs fonctions biologiques. Il est ainsi possible de : (1) identifier des nucléotides ou résidus conservés liés à la fonction, le repliement tridimensionnel ou l'interaction avec d'autres protéines, (2) retracer l'histoire évolutive des gènes en récapitulant les mutations ayant eu lieu dans les différents taxons, (3) prédire la structure tridimensionnelle des protéines, (4) mettre en évidence les résidus ou nucléotides responsables des différences ou similitudes fonctionnelles entre des familles de paralogues, (5) évaluer les vitesses d'évolution des espèces ou encore (6) reconstruire la séquence ancestrale d'un gène ou d'une protéine.

1.5. Applications au cil et à la multiciliation

Les méthodes de génomique comparative ont été employées à plusieurs reprises pour tenter d'éclaircir les mécanismes régissant la ciliation et la multiciliation. La comparaison de répertoires de gènes

a été employée en 2004 dans une étude pionnière pour identifier des gènes ciliaires, par la comparaison des répertoires génétiques de deux espèces ciliées, *Chlamydomonas reinhardtii* et *Homo sapiens*, et d'une espèce non ciliée, *Arabidopsis thaliana* (Li et al., 2004). Ainsi, 688 gènes présents uniquement chez les espèces ciliées ont été mis en évidence, parmi lesquels 5 gènes responsables du syndrome de Bardet-Biedl.

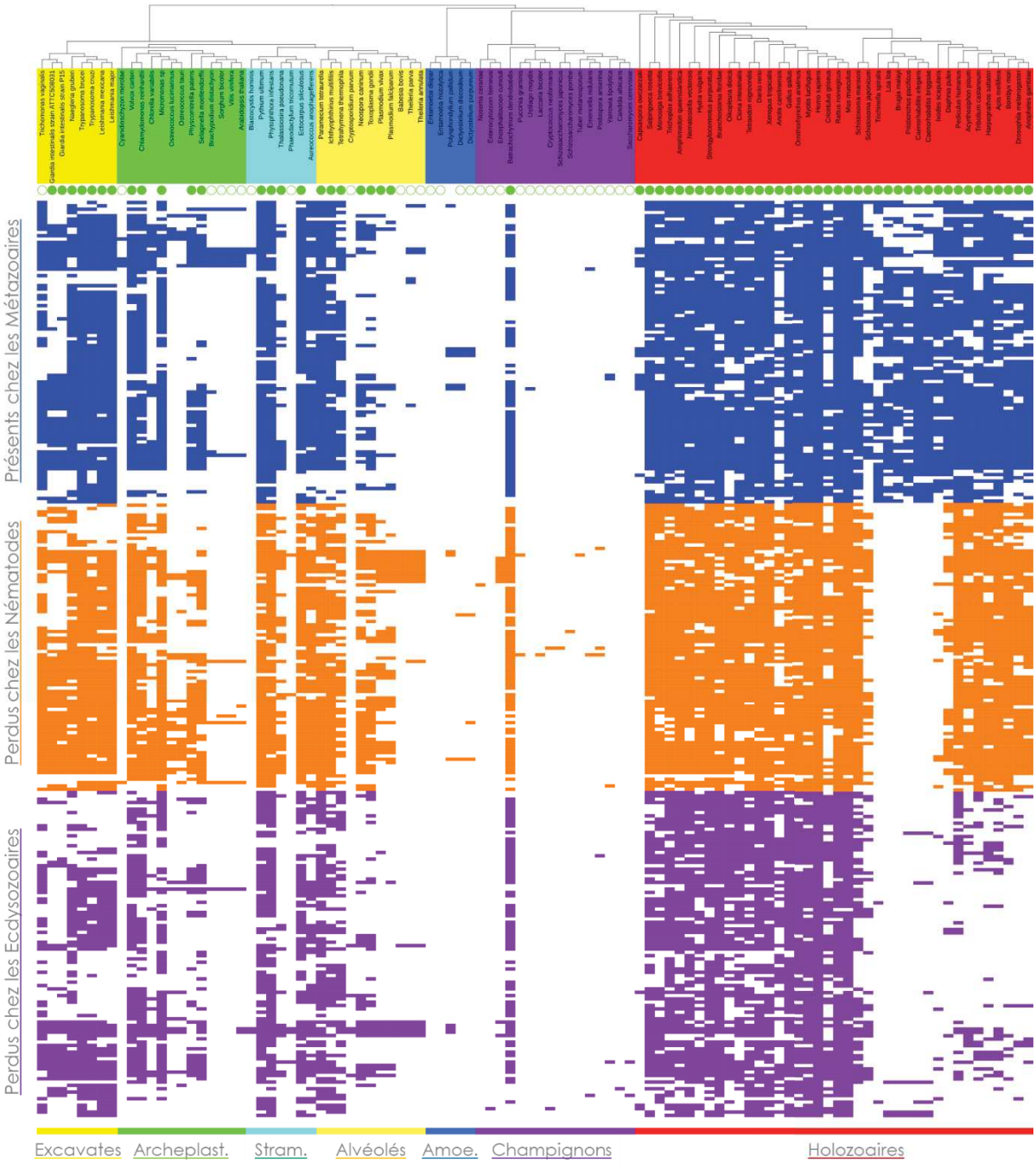

Figure 3-4: Profils phylogénétiques de gènes ciliaires classifiés en 3 modules évolutifs. Les espèces sont colorées en fonction de leur clade d'appartenance (de gauche à droite : Excavates, Archeplastides, Straménopiles, Alvéolés, Amoebozoa, Champignons et Holozoaires). La présence de cil dans l'espèce est indiquée par la présence d'un rond vert, un rond vide indique une absence de cil. La présence d'un gène dans une espèce est indiquée par un carré de couleur, chaque ligne représentant un gène. Les gènes en bleu représentent les gènes présents chez tous les métazoaires, en orange ceux perdus spécifiquement chez les nématodes, et en violet ceux perdus chez les Ecdysozoaires. Figure adaptée de (Nevers et al., 2017).

Plus récemment, le profilage phylogénétique a été utilisé pour étudier le cil, cette approche se prêtant particulièrement bien à l'exploration des gènes ciliaires, puisque, comme nous l'avons vu dans le premier chapitre, l'histoire évolutive du cil est très atypique et présente des pertes indépendantes dans plusieurs taxons. Une étude de profilage phylogénétique a permis de prédire et de valider une douzaine de gènes ciliaires en se basant sur le regroupement de gènes sous forme de *clusters* selon leurs distribution phylogénétique dans 177 espèces (Dey et al., 2015). En 2014, Li et collaborateurs ont proposé l'algorithme CLIME, capable de regrouper des gènes selon leurs profils évolutifs (Li et al., 2014). Appliqué à un groupe de 203 gènes ciliaires connus, CLIME a pu d'une part séparer les gènes selon leurs profils évolutifs pour en dégager des modules fonctionnels du cil (gènes IFT, gènes BBSome,…) et d'autre part prédire de nouveaux gènes impliqués dans le cil présentant un profil évolutif similaire aux gènes connus.

Dans une étude plus récente, le profilage phylogénétique des gènes humains dans 100 espèces eucaryotes a pu mettre en évidence 274 gènes présentant un profil « ciliaire », c'est-à-dire présents dans les espèces ciliées et absents dans les espèces non ciliées. Parmi ces gènes, 87 représentaient de nouveaux gènes candidats potentiellement impliqués dans le fonctionnement du cil ; 21 ont depuis été validés expérimentalement (Nevers et al., 2017). Au-delà de la prédiction de nouveaux gènes, cette étude a également permis une caractérisation fonctionnelle de ces gènes d'après leurs profils évolutifs. Ainsi, les gènes de profil ciliaire perdus chez les Nématodes correspondent à des gènes spécifiques du cil motile, puisque ces espèces ne sont capables de générer qu'un cil primaire, tandis que les gènes ciliaires présents chez tous les métazoaires semblent enrichis en gènes liés à la zone de transition et aux différents complexes de transport intraflagellaire (*Figure 3-4*). La génomique comparative et plus particulièrement la comparaison de gènes entre espèces est donc un outil non négligeable dans l'identification de gènes ciliaires, mais elle permet également la caractérisation fonctionnelle de ces gènes.

L'étude de la multiciliation est en revanche plus compliquée à effectuer par des approches de génomique comparative, à la fois de par la complexité du processus en lui-même, mais également par le manque d'informations actuellement disponibles. Des études de synténie ont néanmoins été réalisées et ont pu montrer que des gènes centraux de la multiciliation, CCNO, MCIDAS, miR-449 et CDC20B, étaient co-localisés chez l'Homme, les autres mammifères, ainsi que chez le xénope (Marcet et al., 2011; Stubbs et al., 2012).

En ce qui concerne la comparaison de séquences homologues, cette approche est régulièrement employée de manière ponctuelle pour étudier des protéines du cil, et nous l'avons exploitée pour étudier les gènes liés à la multiciliation (voir Chapitre 4). Nous reviendrons également sur la mise en place d'une méthode d'exploitation massive de cette approche dans le Chapitre 5.

2. Génomique médicale

Les avancées technologiques de ces dernières années en matière de NGS ont permis de révolutionner les approches de génomique, en particulier dans le domaine médical et dans la prise en charge et l'identification de pathologies. L'utilisation de technologies comme le séquençage de génomes ou d'exomes se démocratise de plus en plus en clinique dans la détermination de prédisposition génétique à la maladie, de diagnostic et de choix thérapeutique. A l'inverse, le séquençage de génomes complets peut également être employé pour mettre en lien des variations

génétiques et des pathologies dont les causes ne sont pas encore comprises. Nous discuterons ici brièvement des deux approches fréquemment employées en génomique médicale : le séquençage de génome ou d'exome complet, et les études d'association à l'échelle du génome complet.

2.1. Séquençage de génomes et d'exomes complets

Depuis le développement du séquençage de génomes complets (WGS : *Whole Genome Sequencing*) à la fin des années 1990, le cout et le temps nécessaire au séquençage d'un génome humain entier sont passées de plusieurs années et plus de 3 milliards de dollars, à quelques milliers de dollars pour un séquençage en quelques jours, grâce aux techniques de NGS. On peut classifier ces méthodes en deux grands groupes en se basant sur leurs caractéristiques spécifiques, on retrouvera ainsi les méthodes de séquençage de deuxième génération, qui produisent des lectures courtes (35 – 800 bases) et qui nécessitent une amplification d'ADN, et les méthodes de séquençage de troisième génération, capables de créer des lectures longues (plusieurs milliers de bases) à partir d'une seule molécule d'ADN, et cela en temps réel (Kumar et al., 2019). Des méthodes hybrides combinant le séquençage de deuxième et de troisième génération sont de plus en plus employées pour permettre de compenser les points faibles de chacune. Les lectures longues sont alors exploitées pour réaliser l'assemblage général du génome et permettent entre autre de déterminer les régions répétées et les variants structuraux que les lectures courtes ne permettent pas de résoudre. Ces dernières sont ensuite utilisées pour obtenir une séquence de haute qualité, puisque le taux d'erreur du séquençage de troisième génération reste relativement élevé. Pour le séquençage routinier dans le cadre médical, le séquençage d'exome complet (WES : *Whole Exome Sequencing*) a été mis au point par soucis de rapidité et d'économie. Ainsi, uniquement les fragments d'ADN codants pour des protéines, les exons, et éventuellement les sites d'épissage sont séquencés (Kim et al., 2017).

Le séquençage de génome ou d'exome entier peut être employé à la fois pour la recherche de mutations connues pour établir un diagnostic ou une susceptibilité de développer une pathologie, ainsi que pour l'identification de nouvelles mutations responsables de pathologies. Lors de la détection d'une mutation, il est possible d'utiliser la fréquence des allèles dans la population pour tenter de distinguer polymorphisme et variants pathogènes, le premier type étant beaucoup plus fréquent que le second. Les données de fréquence d'allèles peuvent typiquement être issues du projet 1000 génomes (Auton et al., 2015) ou de la *genome Aggregation Database* (*gnomAD* ; Karczewski et al., 2020).

2.2. Etude d'association de génome à grande échelle

Le séquençage de génome ou d'exome complet est particulièrement bien adapté à l'identification de mutations dans le cas de maladies rares monogéniques, mais dans beaucoup de cas de maladies communes, les causes génétiques sont multiples et nécessitent une approche de détection différente. En effet, ce type de maladies, qui inclut par exemple l'hypertension, le diabète ou les maladies cardiovasculaires, est lié à la présence de plusieurs variants génétiques retrouvés fréquemment dans la population, dont la combinaison provoque un phénotype pathologique.

Une méthode traditionnellement utilisée pour tenter d'élucider les profils génétiques de ces pathologies est l'étude d'association de génome à grande échelle (GWAS : *Genome-Wide Association*

Study), qui consiste en l'analyse des variations génétiques les plus communes dans un ensemble d'individus. La comparaison de polymorphisme d'un seul nucléotide (SNP : *Single nucleotide polymorphism*) se fait sur plusieurs centaines de milliers de positions dans deux groupes d'individus distincts : un groupe sain et un groupe présentant la pathologie étudiée, et les fréquences de ces SNPs sont alors calculées pour mettre en évidence les loci associés à un risque de développer la pathologie (Frayling, 2014; Sebastiani et al., 2009).

2.3. Application au cil et à la multiciliation

Au vu de leur pouvoir de détection, il n'est pas surprenant que des WES soient effectués pour diagnostiquer des patients présentant des phénotypes de type ciliopathie (Castro-Sánchez et al., 2017). Ils ont également été fréquemment employés pour identifier les différentes variations responsables de ces phénotypes, par exemple dans le cadre des PCD (Horani et al., 2012; Knowles et al., 2013b; Onoufriadis et al., 2014), de rétinopathies (El Shamieh et al., 2014) ou dans des cas de ciliopathies avec une atteinte rénale (Braun et al., 2016).

En ce qui concerne l'application à la multiciliation, nous avons vu dans le chapitre précédent que le WES de patients atteints de RGMC avait permis d'identifier MCIDAS, CCNO, FOXJ1 et NEK10 comme étant des causes génétiques d'aplasie ou d'hypoplasie ciliaire (Boon et al., 2014; Chivukula et al., 2020; Wallmeier et al., 2014, 2019). En revanche, les ciliopathies étant des maladies essentiellement monogéniques, le GWAS ne semble pas être une méthode appropriée pour les étudier, et il n'existe à notre connaissance aucune expérience de ce type.

3. Génomique fonctionnelle

Nous avons vu jusqu'à présent plusieurs méthodes dédiées à l'étude du génome ayant pour but la caractérisation des gènes, soit par la comparaison de phénotypes entre plusieurs espèces dans le cadre de la génomique comparative, soit par la comparaison de phénotypes sains et de phénotypes pathologiques en génomique médicale. La génomique fonctionnelle quant à elle se concentre sur l'analyse des aspects dynamiques des processus cellulaires, tels que la transcription, la régulation de l'expression des gènes, la traduction ou encore les interactions entre protéines. Cette discipline se base notamment sur un panel de technologies à haut-débit dont le but est de mesurer à la fois qualitativement et quantitativement les différents produits issus de l'expression du génome.

3.1. Transcriptomique

La transcriptomique est l'étude de l'ensemble des transcrits d'ARN dans un organisme, une population de cellules voire une seule cellule, à un moment précis dans le temps. Les analyses de transcriptomique ont commencé à voir le jour dans les années 1990, avec l'émergence d'une des techniques majeures en 1995 : les puces à ADN (Lowe et al., 2017). Cette technologie se base sur l'hybridation des transcrits à des fragments de séquence ADN fixés aux puces, appelés sondes, correspondant à des gènes connus. Les transcrits sont préalablement étiquetés à l'aide d'une molécule fluorescente pour permettre leur détection à la surface des puces ; l'intensité de la fluorescence à chaque position correspond alors au niveau d'expression du gène. La seconde technique majeure employée pour réaliser des études de transcriptomique est le séquençage ARN, ou *RNASeq*. Mise au point au milieu des années 2000, son développement a largement été influencé par les avancées technologiques de cette époque liées aux NGS. Le *RNASeq* est une approche basée

sur le séquençage des transcrits présents dans un échantillon, qui sont ensuite alignés à un génome de référence avant d'être comptés. Plus récemment, le *RNASeq* a été adapté à l'étude d'une seule cellule, ce qui a l'avantage de permettre l'analyse de tissus hétérogènes ou des analyses en cours de développement de manière beaucoup plus précise, on parle alors de *single-cell RNASeq* (*scRNASeq*).

En pratique, la transcriptomique est fréquemment utilisée pour comparer deux conditions et mettre en évidence les variations d'expression de gènes, pour les caractériser de manière dynamique et précise. L'exploitation des données issues d'expériences de transcriptomique permettent ainsi d'obtenir un certain nombre d'informations telles que : (1) l'expression d'un ou de plusieurs gènes au cours du temps (*e.g.* pendant le développement), (2) la réponse cellulaire face à un médicament, un pathogène ou une maladie, (3) l'influence d'un ou plusieurs transcrits sur la régulation et l'expression d'autres gènes ou encore (4) les différences de régulation ou d'expression de gènes entre différentes espèces, différents tissus ou entre un individu sain et un patient atteint d'une pathologie.

3.2. Protéomique et interactomique

La protéomique est la discipline haut-débit s'intéressant à l'ensemble des protéines produites par un organisme, un tissu ou un organe, à un moment donné. Elle permet d'apporter des informations complémentaires à celles obtenues par la génomique ou la transcriptomique, notamment en ce qui concerne les évènements de modifications post-traductionnelles, à l'origine d'une grande diversité protéique qui ne peut être expliquée par le nombre de gènes connus ni même par le nombre d'isoformes par gène. De nos jours, les études de protéomique sont majoritairement réalisées par des techniques de spectrométrie de masse. Le principe de base est le suivant : les protéines sont extraites de l'échantillon, digérées par des enzymes pour obtenir des fragments protéiques, séparées par chromatographie liquide, puis quantifiées et caractérisées par spectrométrie de masse (Zhang et al., 2014). La comparaison à des bases de données permet d'identifier les différentes protéines et mettre en évidence la présence d'éventuelles modifications post-traductionnelles.

Les méthodes de protéomique peuvent également être employées pour mettre en évidence des liens physiques et fonctionnels entre protéines, dans une discipline que l'on appelle interactomique. Il est ainsi possible d'identifier à la fois des interactions binaires entre protéines, mais également les différents éléments formant un complexe. Dans le cas de l'étude des complexes, l'approche reste très similaire à celle employée pour l'analyse du protéome complet d'un échantillon biologique, à la différence près que l'échantillon analysé ne contient que le complexe d'intérêt purifié. En ce qui concerne les interactions binaires, il existe deux méthodes possibles pour les identifier. La première, toujours basée sur la spectrométrie de masse, consiste en la purification d'une protéine d'intérêt préalablement étiquetée, pour recouvrer l'ensemble de ses interactants et les identifier. La seconde méthode est celle du double hybride ; elle consiste en la fusion d'un domaine d'interaction à l'ADN à une protéine d'intérêt (protéine 'appât') et la fusion d'un domaine d'activation à un ensemble de protéines cibles. Ces protéines fusionnées sont exprimées dans des cellules de levure, et leur interaction va provoquer la transcription d'un gène rapporteur. Pour permettre les analyses à plus grande échelle, un système de puces a été développé pour tester un plus grand nombre de protéines cibles à la fois (Rajagopala and Uetz, 2011; Sardiu and Washburn, 2011).

3.3. Métabolomique et autres disciplines

En plus de la génomique, transcriptomique et protéomique, il existe de nombreuses autres approches à haut-débit, et les améliorations techniques constantes permettent régulièrement la mise au point de nouvelles disciplines. Il existe ainsi le domaine de la métabolomique, dédiée à l'étude des métabolites par des méthodes de spectrométrie (Yang et al., 2019), de l'épigénomique, qui recense les modifications de l'ADN et la structure de la chromatine par séquençage ciblé (Friedman and Rando, 2015), ou encore de la lipidomique, sous-ensemble de la métabolomique, centrée sur l'analyse des lipides (Han, 2016).

3.4. Application au cil et à la multiciliation

Différentes approches de génomique fonctionnelle ont été appliquées à l'étude du cil et de la multiciliation. Des études de transcriptomiques ont par exemple comparé l'expression de gènes dans différents tissus ciliés et non ciliés chez *C. elegans* (Blacque et al., 2005) et chez la souris (McClintock et al., 2008), permettant de confirmer l'expression ciliaire de 14 et 99 gènes respectivement. Une autre approche régulièrement employée est l'analyse de l'expression des gènes au cours de la ciliogénèse ; appliquée à *Chlamydomonas reinhardtii*, elle a permis de mettre en évidence 1850 gènes surexprimés dont 4 nouveaux candidats ayant des orthologues chez l'homme (Albee et al., 2013). Il existe également la base de données CilDB, qui regroupe plusieurs expériences de génomique fonctionnelle et de génomique comparative basées sur le cil, elle n'est en revanche pas maintenue à jour (Arnaiz et al., 2014).

Dans le contexte de la multiciliation, les expériences de transcriptomiques sont souvent conçues pour identifier les rôles de certains gènes dans le processus régulatoire, en comparant les profils d'expression de cellules multiciliées en condition normale et de cellules multiciliées dans lesquelles l'expression d'un gène spécifique est inhibée. Ce type d'approche a par exemple mis en évidence la régulation de l'expression de FOXJ1 par MCIDAS et MYB (Stubbs et al., 2012; Tan et al., 2013), ou le contrôle de l'expression de FOXJ1, RFX2, RFX3, miR-34b/c et MYB par p73 (Nemajerova et al., 2016). Plus récemment, des expériences de *scRNASeq* ont été réalisées sur l'épithélium mucociliaire de l'homme et de la souris et ont permis de mieux caractériser la dynamique de différenciation des cellules du système respiratoire au cours du temps (Ruiz García et al., 2019).

Les méthodes de protéomique quant à elles, ont été appliquées plusieurs fois dans le cadre de l'identification de protéines ciliaires, et ont permis d'identifier un certain nombre de protéines potentiellement impliquées dans le fonctionnement du cil. Ainsi, Ishikawa et collaborateurs ont mis en évidence une quarantaine de gènes candidats du cil primaire, dont plusieurs ont été confirmés expérimentalement par la suite (Ishikawa et al., 2012). De même, Blackburn et collaborateurs ont identifié 8 protéines spécifiquement exprimées dans l'axonème de cellules multiciliées de l'épithélium respiratoire (Blackburn et al., 2017).

4. Approches intégratives

Avec le développement des disciplines à haut-débit et leurs améliorations constantes, l'accès aux données de type '-omiques' est de moins en moins onéreux et il devient fréquent de voir des études combiner des données issues de plusieurs types d'expériences. Nous aborderons ici les principes de base des méthodes intégratives, les limitations qu'elles rencontrent encore aujourd'hui

ainsi que la façon dont elles peuvent être employées pour aider à une meilleure compréhension des processus cellulaires.

4.1. Méthodes

On considère généralement qu'il existe deux types d'approches intégratives : les analyses multi-étapes, où l'information est intégrée de manière successive ou hiérarchique, et les analyses méta-dimensionnelles, où toutes les données sont intégrées à la fois dans le but de créer un modèle basé sur plusieurs types d'informations (*Figure 3-5*; Ritchie et al., 2015).

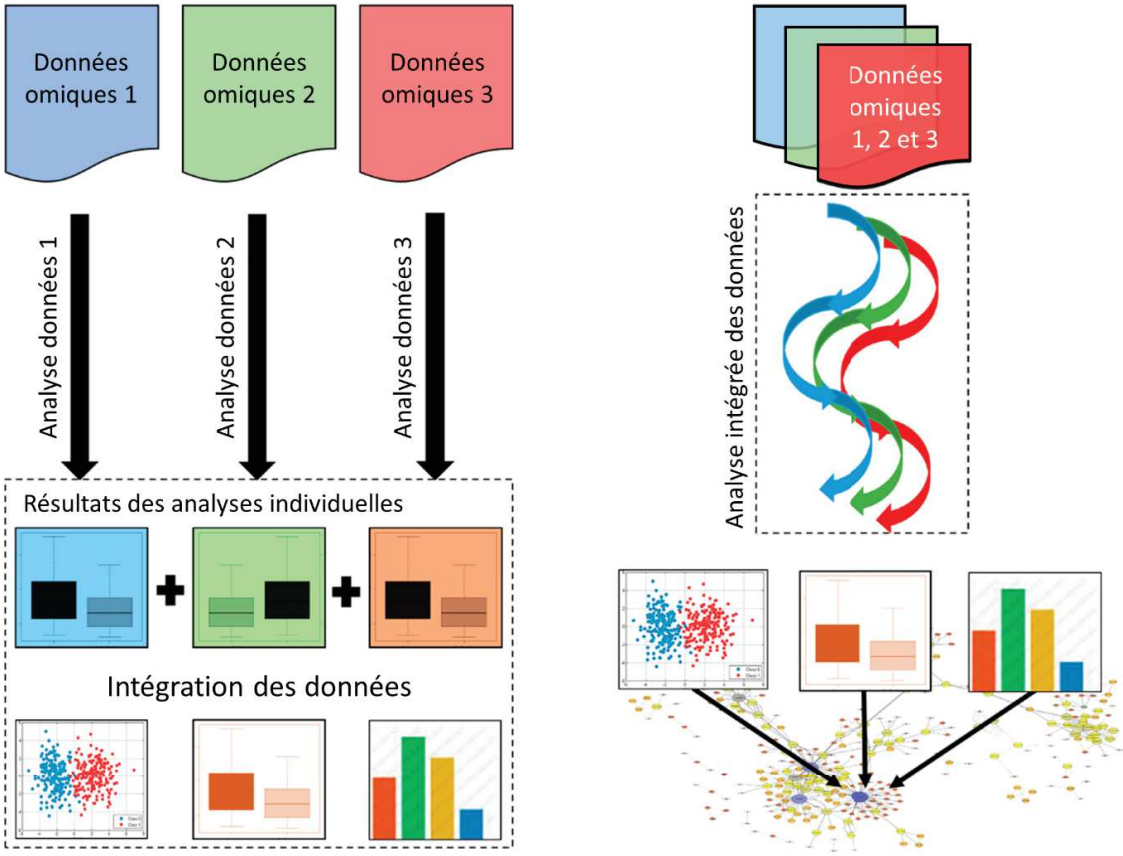

Figure 3-5: Schématisation des deux types d'approches intégratives existants. A gauche l'approche multi-étapes, où chaque type de données est analysé de façon individuelle. A droite, l'approche méta-dimensionnelle où toutes les données sont analysées de façon simultanée et intégrée. Adaptée de (Pinu et al., 2019).

4.1.1. Approches multi-étapes

Il est plus courant de voir des approches multi-étapes, leur développement s'étant fait de manière naturelle au cours des recherches tentant d'élucider les relations génotype/phénotype. Depuis maintenant plusieurs années, on trouve des études basées sur la comparaison de données issues de séquençage ADN et d'expériences de transcriptomique, en particulier dans le domaine médical et de la médecine personnalisée. De plus en plus souvent, ces analyses tendent à intégrer également des données de protéomique et de métabolomique pour obtenir des données de type moléculaire et tenter de remettre les variations génétiques dans leur contexte biologique. Classiquement, ces données sont analysées individuellement puis comparées entre elles pour mettre en évidence un lien entre les différents types de données ainsi qu'un lien avec le phénotype

d'intérêt. Il est à noter que ce type d'approche multi-étapes est souvent lié à une hypothèse de transmission de variation linéaire, selon laquelle une variation génomique sera reflétée par une variation transcriptomique, puis protéomique, puis phénotypique (Hypothèse A sur la *Figure 3-6*).

4.1.2. Approches méta-dimensionnelles

A mesure que les données générées par les différentes disciplines à haut débit prennent de l'ampleur, il devient de plus en plus intéressant de mettre au point des approches méta-dimensionnelles automatisées pouvant traiter simultanément plusieurs types de données pour en faire ressortir un modèle d'interactions biologiques. Ce type d'approche permet alors de se défaire du biais introduit par l'interprétation humaine et d'inférer des relations beaucoup plus complexes que celles vues dans les hypothèses de transmission de variation linéaires, où chaque type de molécule peut influer sur une autre (Hypothèse B sur la *Figure 3-6*). Il existe aujourd'hui de nombreux outils dédiés à l'exploitation de données multi-omiques de manière simultanée (pour une revue récente des outils actuellement disponibles, voir Subramanian et al., 2020).

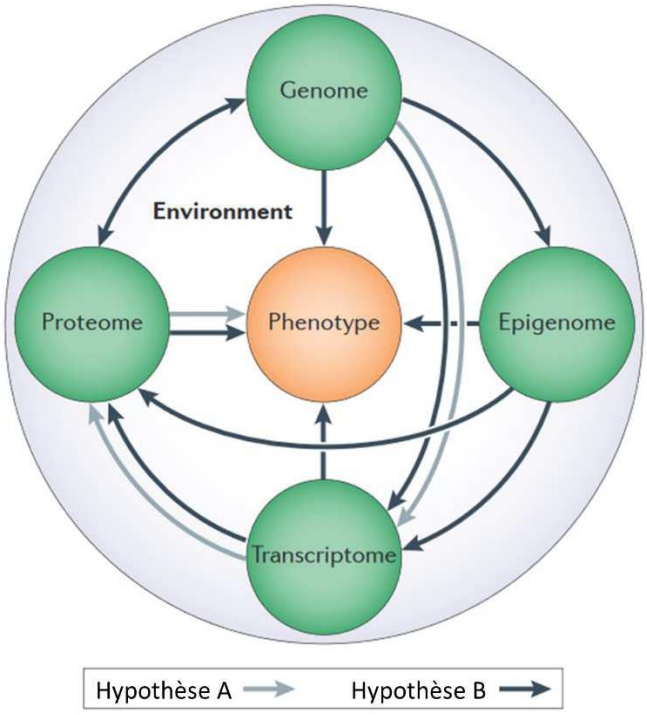

Figure 3-6: Hypothèses de transmission de variations dans des cas de phénotypes complexes. L'hypothèse A correspond à une transmission de variation linéaire, où chaque niveau moléculaire influe sur le suivant. L'hypothèse B correspond à une transmission multifactorielle où tous les niveaux moléculaires interagissent pour donner lieu au phénotype. Adaptée de (Ritchie et al., 2015).

Ces différentes méthodes sont basées sur des approches algorithmiques diverses qu'il est parfois difficile de classifier, nous pouvons néanmoins les diviser en plusieurs catégories : intégration précoce, intégration tardive, et intégration intermédiaire (*Figure 3-7* ; Rappoport and Shamir, 2018). Les approches à intégration précoce concatènent des matrices correspondant aux données de chaque 'omique', puis appliquent un algorithme de *clustering* unique à la matrice résultante. Ce type d'approche a pour inconvénient d'augmenter largement la dimensionnalité des données et de rendre

le *clustering* parfois complexe. Les approches tardives se basent sur le *clustering* individuel de chaque jeu de données avant l'intégration, ceci permettant le traitement de chaque 'omique' de façon optimale. Les *clusters* sont ensuite intégrés, au risque de perdre des informations concernant les interactions les plus faibles. Enfin, les approches dites 'intermédiaires' ont pour but de créer un modèle basé sur l'ensemble des données disponibles, et ce de différentes manières : (1) en se basant sur la similarité des échantillons, (2) en se basant sur la réduction de dimensionnalité conjointe des données, ou encore (3) en utilisant des modèles statistiques. Ainsi, les méthodes basées sur la similarité utilisent la distance entre les échantillons de chaque 'omique' pour permettre le *clustering* des données, et n'intègrent que ces valeurs de similarité. Les algorithmes de type réduction de dimensionnalité représentent dans un premier temps les données dans un espace de plus petite dimension avant de les intégrer. Les méthodes statistiques quant à elles génèrent un modèle probabiliste des données, en employant par exemple une approche de type bayésienne, pour générer les *clusters* les plus probables. Ce type d'approche a l'avantage de pouvoir inclure des connaissances biologiques existantes lors de la création du modèle statistique.

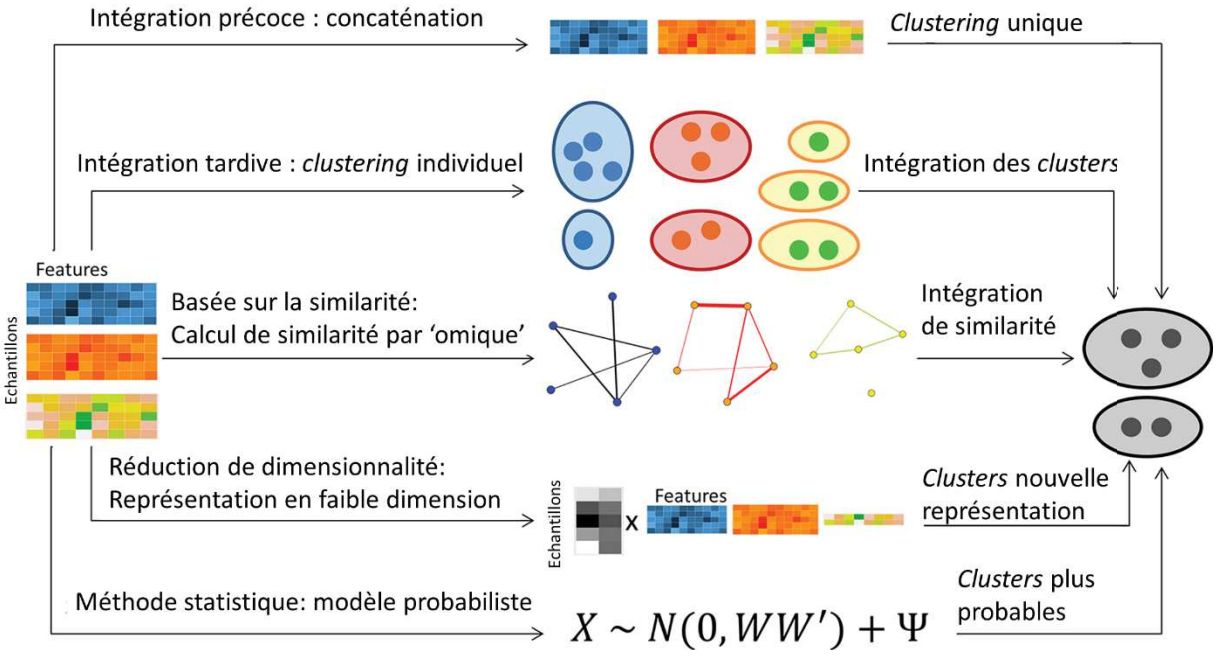

Figure 3-7: Schéma récapitulatif des différentes approches méta-dimensionnelles. Figure adaptée de (Rappoport and Shamir, 2018).

4.2. Applications et défis des approches intégratives

Les approches intégratives sont employées depuis maintenant une décennie pour tenter d'améliorer notre compréhension de la biologie moléculaire et des relations génotype/phénotype, en particulier dans le domaine médical. Historiquement, les premiers travaux multi-omiques réalisés combinaient en plusieurs étapes des données de séquençage ADN et de transcriptomique, et étaient généralement orientés vers l'étude des différents types de cancers (Curtis et al., 2012; The Cancer Genome Atlas Network, 2012). A l'heure actuelle, ce type d'approche reste très fréquemment employé pour des études médicales, par exemple pour l'étude des maladies cardiovasculaires (Leon-Mimila et al., 2019; approches diverses), l'établissement de diagnostics et de traitements adaptés dans des cas de cancers ovariens (Zheng et al., 2019; analyse méta-dimensionnelle de type

statistique), ou encore la classification précise de pathologies rénales chroniques pour améliorer les possibilités de traitement (Eddy et al., 2020; approches diverses).

Malgré les avancées dans le domaine des approches intégratives réalisées ces dernières années, il reste un certain nombre de challenges à surmonter. On retrouve naturellement les défis inhérents à chaque discipline '-omique'. Ces approches haut-débit sont en effet caractérisées par un bruit important, et parfois par un manque de reproductibilité, qui peuvent être liés à des défauts expérimentaux ou à l'analyse en elle-même des résultats. A cela s'ajoutent les problèmes liés à l'intégration de données provenant de différentes plateformes technologiques voire de différentes espèces. Cette diversité crée une hétérogénéité au niveau des identifiants des données, des formats et des méthodes de stockage, rendant l'intégration automatique d'autant plus complexe et aboutissant parfois à un faible recouvrement entre les différents sets de données. Ainsi, la plupart des outils intégratifs requièrent un prétraitement des résultats pour travailler avec un format spécifique, mais il n'existe pas à l'heure actuelle de méthode consensus permettant de choisir adéquatement les critères de traitement préalable (Subramanian et al., 2020).

Enfin, s'il existe comme on l'a vu des études multi-omiques, peu à l'heure actuelle combinent des approches de génomique comparative et de génomique fonctionnelle. En ce sens, la base de données STRING (Szklarczyk et al., 2019) peut être considérée comme pionnière parmi les ressources généralistes, puisqu'elle regroupe des informations relevant aussi bien de la génomique fonctionnelle (co-expression, interactomique…) que de la génomique comparative ou du *textmining.* Elle combine ces informations pour élaborer un score de confiance attribué aux liens fonctionnels entre protéines, afin de prédire au mieux ces dernières.

4.3. Vers une approche intégrative de la multiciliation

En ce qui concerne l'étude du cil, nous avons vu précédemment que la plupart des travaux relevaient soit de la génomique comparative soit de la génomique fonctionnelle, mais là encore, on retrouve peu de croisements. Exception notable, une étude récente a utilisé une approche intégrative de type bayésien combinant les résultats de diverses expériences de protéomique, génomique, transcriptomique et génomique comparative pour prédire de nouveaux gènes ciliaires, parmi lesquels OSCP1, dont la fonction a été validée expérimentalement (van Dam et al., 2019). Cette analyse s'appuie sur 8 jeux de données : (1) trois jeux de protéomique, (2) un jeu de génomique caractérisant la présence ou l'absence de site de liaison à un facteur de transcription spécifiques à RFX et FOXJ1 dans des régions promotrices, (3) un jeu de données comprenant des gènes co-exprimés avec des gènes ciliaires connus, (4) un jeu recensant la corrélation entre le profil de présence/absence d'un gène et le profil de présence du cil chez les eucaryotes, (5) l'étude protéomique du cil primaire de Liu et collaborateurs (Liu et al., 2007a) et (6) un jeu public d'expression de gènes au cours de la ciliogénèse (Ross et al., 2007). Ces données ont ensuite été intégrées par une approche statistique bayésienne, et pour chaque gène un score de probabilité d'être ciliaire a été calculé en se basant sur des informations préalables de gènes liés au cil. L'ensemble des résultats de cette étude a été regroupé dans le compendium CiliaCarta [http://bioinformatics.bio.uu.nl/john/syscilia/ciliacarta/].

De telles études intégratives n'existent pas à notre connaissance pour la multiciliation qui demeure un processus cellulaire complexe, essentiellement abordé par des études de génomique

fonctionnelle et largement inexploré sur le plan évolutif. L'objectif de mes travaux de thèse a été d'améliorer la compréhension des mécanismes de la multiciliation par diverses approches de génomique. Nous avons dans un premier temps employé des méthodes de génomique comparative pour réaliser une étude évolutive de la multiciliation, qui sera détaillée dans le chapitre 4 de ce manuscrit. A partir de ces résultats, nous avons mis au point une nouvelle méthode de génomique comparative basée sur la comparaison de séquences homologues, qui fait l'objet du chapitre 5. Enfin, le dernier chapitre traitera de l'étude intégrative que nous avons réalisée pour comparer des données issues d'approches fonctionnelle et évolutive dans le but de mettre en évidence de nouveaux gènes candidats potentiellement impliqués dans la multiciliation.

CONTRIBUTIONS

Chapitre 4 : Etude évolutive des protéines de la multiciliation chez les métazoaires

Nous avons vu précédemment que la multiciliation est un processus complexe encore mal compris, dont l'évolution n'est à ce jour pas précisément caractérisée. En effet, de nombreuses incertitudes persistent quant à la présence de ce processus chez une variété d'organismes, notamment chez les Invertébrés (voir *Figure 2-1*) ; seule son absence chez les Ecdysozoaires semble être véritablement établie. Il existe toujours beaucoup d'interrogations, à la fois en termes des acteurs impliqués dans la multiciliation, mais également en ce qui concerne la conservation de ce processus entre les différentes espèces voire entre les différents tissus d'une même espèce.

Dans le but d'éclaircir les mécanismes régulant la multiciliation et d'en identifier de nouveaux acteurs, nous avons appliqué plusieurs des approches de génomique comparative décrites dans le chapitre précédent. Notre objectif a été dans un premier temps d'obtenir des informations relatives à l'évolution de la multiciliation chez les Métazoaires, les données disponibles dans la littérature étant relativement parcellaires, et les études fonctionnelles étant limitées à des organismes modèles tels que la souris, le xénope ou le poisson zèbre. Nous avons donc établi un bilan évolutif plus complet, prenant en compte un plus grand nombre d'espèces représentatives des taxons majeurs des Métazoaires. Nous avons limité notre analyse à 10 familles protéiques que nous avons jugées comme centrales dans la multiciliation.

Après avoir établi la distribution phylogénétique de ces protéines dans un ensemble d'espèces choisies, nous avons analysé de manière approfondie leur conservation à travers l'évolution par l'étude d'alignements multiples. Nous avons également analysé le contexte génomique des gènes CCNO, MCIDAS et CDC20B que nous savons co-localisés dans plusieurs espèces. L'ensemble de ces résultats a finalement servi de point de départ pour une recherche de gènes de la multiciliation par application du profilage phylogénétique grâce à la ressource OrthoInspector.

1. Distribution phylogénique des protéines de la multiciliation

La première étape de notre étude des gènes de la multiciliation a consisté en l'établissement d'une distribution phylogénétique précise de ces 10 familles, pour cela nous avons procédé à une recherche de séquences protéiques homologues dans des bases de données publiques.

1.1. Choix des familles protéiques

Les familles de protéines sur lesquelles nous avons choisi de nous concentrer, sur la base de leurs rôles centraux dans la multiciliation (*Figure 4-1*), sont les suivantes : CEP63, DEUP1, MCIDAS, GEMC1, CCNO, CCDC78, E2F4, E2F5, CEP152 et CDC20B (voir Chapitre 2).

MCIDAS et GEMC1 ont été choisis pour leur rôle majeur de régulateurs de la multiciliation grâce à l'activation, par leur biais, de divers facteurs de transcription. Pour permettre cette activation, nous avons vu au cours du chapitre 2 que la présence d'E2F4 ou d'E2F5 était requise, nous avons les avons donc étudiés également.

Nous avons ensuite sélectionné les gènes impliqués dans la multiplication massive des corps basaux, étape clé de la multiciliation. Nous avons ainsi retenu CCNO, régulateur de la voie impliquant le deutérosome, et également connu pour être la cause de RGMC, ces ciliopathies spécifiques des MCCs. Nous avons ensuite choisi les acteurs principaux des deux voies d'assemblage connues : CEP63 et CEP152 pour la voie dépendante du centriole père, et DEUP1 pour la voie dépendante du deutérosome, qui interagit également avec CEP152, à l'aide de CCDC78.

Enfin, le dernier gène que nous avons choisi d'étudier est CDC20B, qui non seulement semble être impliqué dans le désengagement des corps basaux à la fin de leur amplification, mais est également le gène hôte de la famille de microARN miR-449, dont la présence est nécessaire à la détermination du devenir multicilié par l'inhibition de Notch.

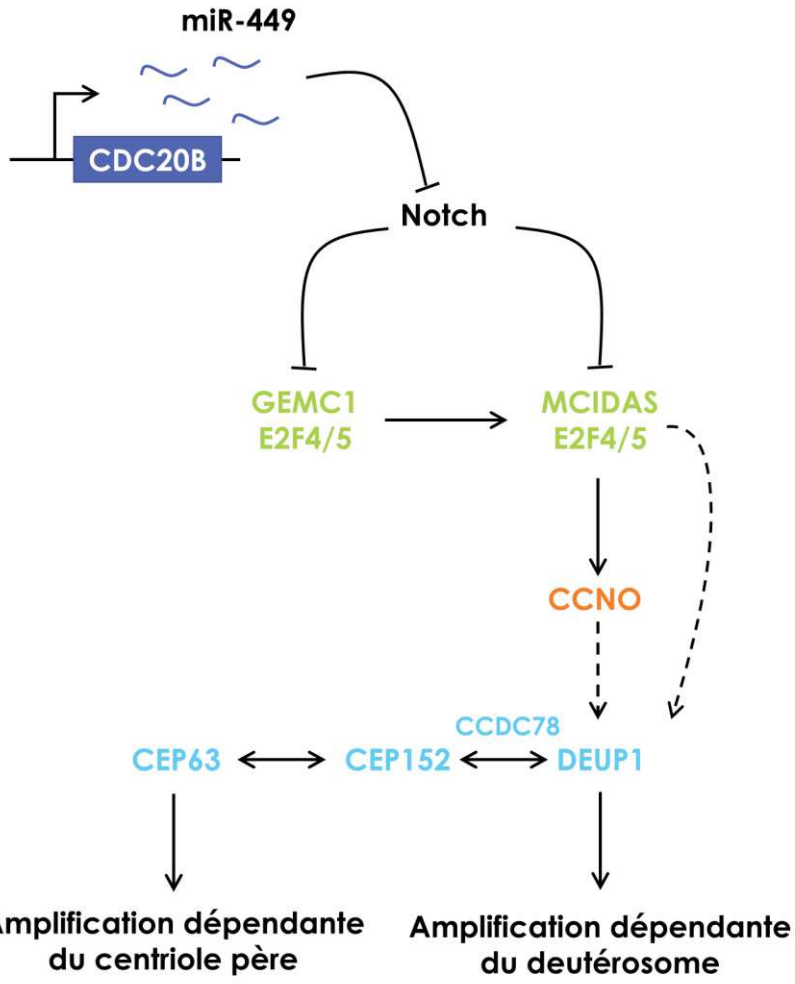

Figure 4-1: Schéma récapitulatif des interactions entre les gènes centraux sélectionnés pour l'étude évolutive.

1.2. Recherche de séquences homologues

Pour limiter les erreurs de prédiction d'homologie pouvant être liées aux erreurs d'annotation ou aux programmes de détection d'orthologie, nous avons effectué une recherche de séquences homologues manuelle, à partir des séquences protéiques humaines des 10 familles d'intérêt, en utilisant BLASTp (Altschul et al., 1997). Une première recherche a été effectuée sur la section protéique de la base de données RefSeq (O'Leary et al., 2016) du NCBI, une seconde sur la

base de données Uniprot (The Uniprot Consortium, 2019) pour obtenir des séquences additionnelles, les deux banques ne se recouvrant que partiellement. Cette approche a ensuite été complétée par des recherches tBLASTn au niveau génomique pour détecter certains gènes non prédits ou mal prédits lors de l'annotation du génome, ou pour confirmer l'absence du gène chez un organisme (voir Chapitre 8: Matériel et Méthodes). Pour chaque famille de protéines, un alignement multiple a été réalisé (voir partie 2 plus bas) à partir des séquences homologues détectées pour identifier des motifs conservés au cours de l'évolution et confirmer l'appartenance de chaque séquence à la famille étudiée et ainsi distinguer, le cas échéant, orthologues et paralogues.

1.3. Bilan évolutif

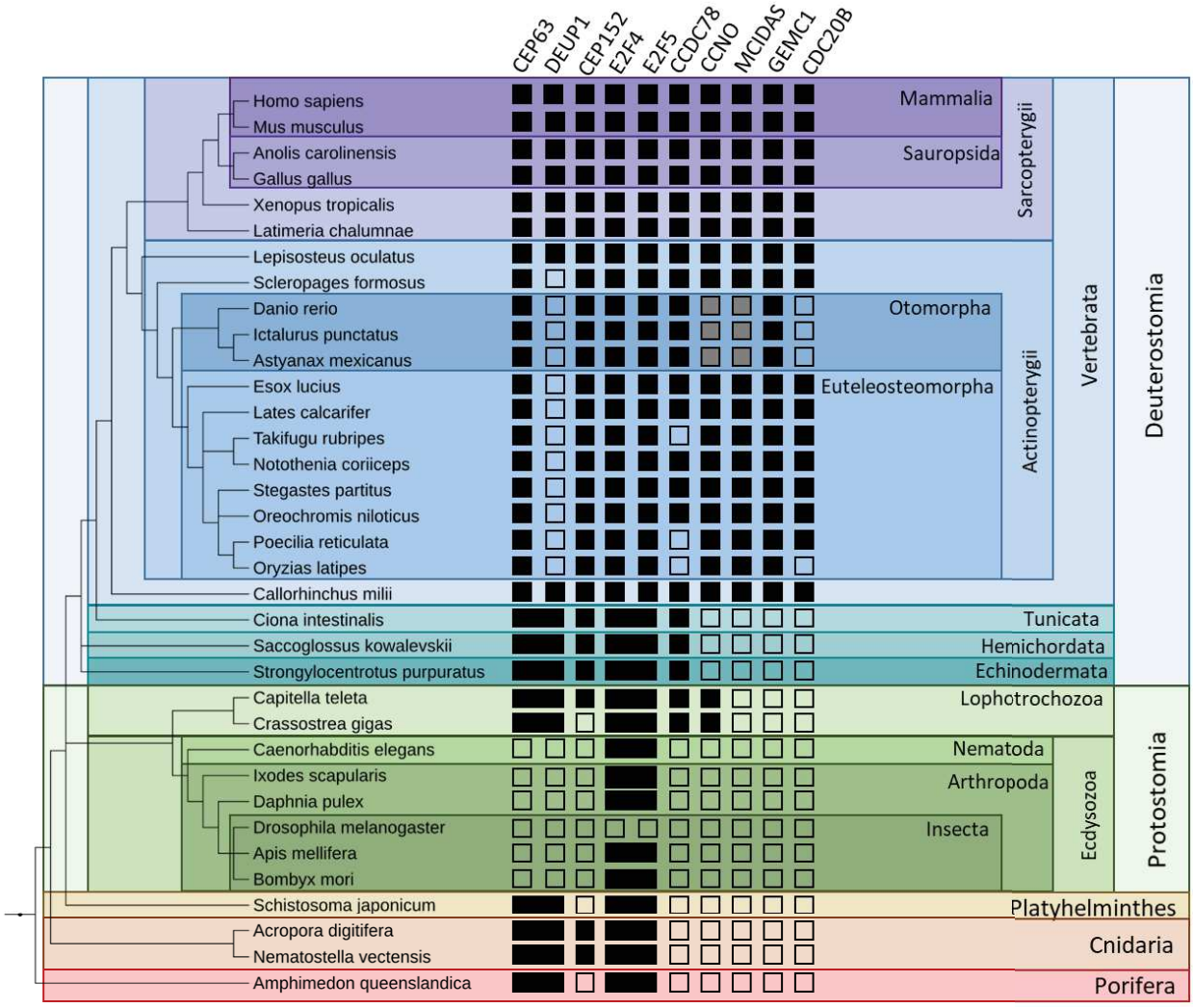

Figure 4-2: Distribution phylogénétique des gènes de la multiciliation chez les Métazoaires. Seule une partie des espèces étudiées est ici représentée. La présence d'orthologues dans un organisme est représentée par un carré noir, une absence par un carré vide. Les carrés gris dénotent des homologues présents mais dont la séquence est anormalement divergente.

Lors de l'établissement de la distribution phylogénétique des gènes de la multiciliation, nous avons pu discerner deux types de gènes, ceux à origine ancienne, et ceux apparus plus récemment au cours de l'évolution. L'ensemble des résultats de ce bilan est représenté dans la Figure 4-2, dans laquelle nous avons indiqué la présence et l'absence des protéines étudiées dans un sous-ensemble d'espèces que nous avons sélectionnées de manière à offrir une couverture optimale des principaux taxons Métazoaires, en choisissant préférablement des espèces modèles. Notre étude se base

néanmoins sur un ensemble plus vaste d'espèces, les familles protéiques étudiées contenant de 92 à 351 séquences, et les distributions que nous avons établies s'appuient sur l'ensemble des organismes étudiés. Nous avons cherché de cette manière à obtenir une distribution la plus fiable possible, les protéomes de certaines espèces actuellement disponibles dans les bases de données publiques étant de qualité discutable et parfois incomplets. Dans certains cas, nous avons également représenté la présence d'homologues divergents dont la séquence est conservée de manière atypique ; nous en reparlerons dans la partie suivante lorsque nous aborderons la conservation des séquences des familles protéiques au cours de l'évolution.

1.3.1. Familles anciennes

Trois des familles étudiées (CEP63/DEUP1, E2F4/E2F5 et CEP152) semblent être apparues au début de l'évolution des Métazoaires. Elles sont en effet présentes dans la majorité des taxons que nous avons étudiés : les Deutérostomiens (Chordés et Echinodermes), certains Protostomiens (principalement chez les Lophotrochozoaires), les Cnidaires tels que l'anémone *Nematostella vectensis*, et pour CEP63/DEUP1 et E2F4/E2F5, chez les éponges (*Amphimedon queenslandica*) et les Plathelminthes (*Schistosoma japonicum*). Compte tenu de cette origine ancestrale, les absences observées dans certains groupes relèveraient de pertes secondaires.

1.3.1.1. Famille CEP63/DEUP1

Nous avons dans un premier temps tenté d'éclaircir les évènements évolutifs à l'origine de la famille de paralogues CEP63/DEUP1 et notamment préciser le moment de leur duplication. Une précédente étude réalisée sur un nombre réduit d'espèces a relevé une absence marquée de DEUP1 chez les Actinoptérygiens (poissons à nageoires rayonnées), et émet l'hypothèse d'une duplication ayant eu lieu chez les Sarcoptérygiens (Cœlacanthes, Dipneustes et Tétrapodes) (Zhao et al., 2013). Nos résultats suggèrent une duplication plus ancienne puisque nous avons détecté deux séquences annotées comme étant des orthologues de DEUP1 chez *Lepisosteus oculatus* (appartenant aux Actinoptérygiens) et chez la chimère (*Callorhinchus milii*) appartenant à la classe des chondrichtyens (poissons cartilagineux). Nous n'avons pas pu confirmer avec certitude l'appartenance de ces séquences à la famille d'orthologues DEUP1 sur la base de l'étude de l'alignement multiple, nous avons donc construit un phylogramme pour clarifier les relations entre les séquences *(Figure 4-3*).

L'arbre phylogénétique généré par PhyML (Guindon et al., 2010) regroupe clairement les séquences CEP63/DEUP1 de Vertébrés par opposition aux séquences apparentées à CEP63 d'Invertébrés. Les séquences de Vertébrés se divisent en 2 groupes statistiquement robustes : (1) le groupe CEP63 contenant les branches des tétrapodes, du cœlacanthe, de la chimère et des Actinoptérygiens, (2) un groupe contenant l'ensemble des séquences de DEUP1, y compris les séquences de *Callorhinchus* et de *Lepisosteus*. Ces résultats semblent indiquer qu'un gène ancestral apparenté à CEP63 est apparu au début de l'évolution des Métazoaires, qu'un évènement de duplication a eu lieu chez les Vertébrés, puis que DEUP1 a été perdu chez les poissons Téléostéens.

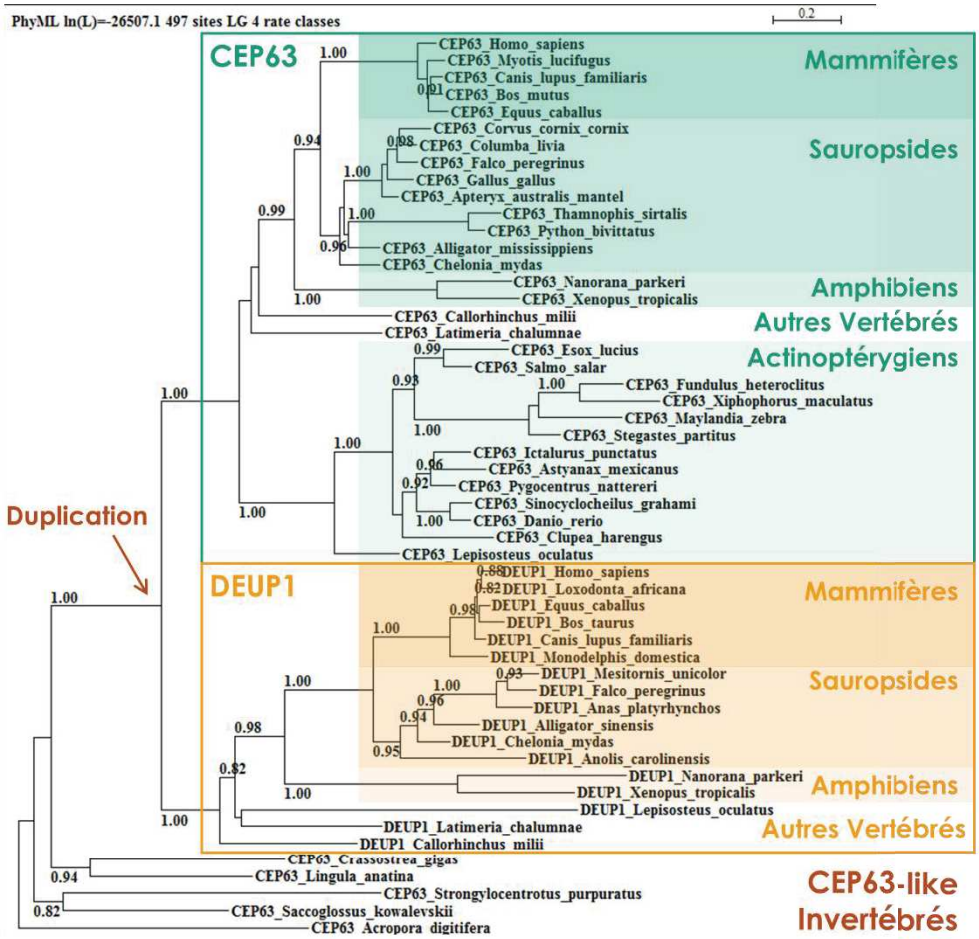

Figure 4-3: Arbre phylogénétique de la famille CEP63/DEUP1 réalisé à partir de l'alignement multiple par PhyML (maximum de vraisemblance). La robustesse des branches a été évaluée par la méthode aLRT (Approximate Likelihood Ratio Test) ; seules les valeurs de support supérieures à 0,8 sont représentées. Les relations entre les séquences suggèrent une duplication au début de l'évolution des Vertébrés.

1.3.1.2. CEP152

CEP152 présente une distribution phylogénétique globalement similaire à celle observée pour CEP63, avec une présence dans les clades multiciliés (Deuterostomiens, Mollusques et Cnidaires), et une absence marquée chez les organismes non multiciliés (Arthropodes et Nématodes). Ces similarités de distribution semblent cohérentes avec l'interaction observée entre CEP63 et CEP152 au cours de l'amplification des corps basaux médiés par le centriole père. Aucune séquence homologue de CEP152 n'a été trouvée au niveau génomique chez les Porifères ou les Plathelminthes, confirmant l'absence de ces gènes dans les bases de données actuelles. Le petit nombre et le caractère encore fragmentaire des génomes disponibles pour ces espèces ne permet cependant pas d'être catégorique quant à l'absence réelle de ce gène dans les génomes considérés. En revanche, l'absence de CEP63 et de CEP152 dans les clades non multiciliés suggère une perte directement corrélée à l'absence de multiciliation chez les Arthropodes et les Nématodes.

1.3.1.3. Famille E2F4/E2F5

Tout comme pour la famille CEP63/DEUP1, nous avons voulu préciser l'histoire évolutive de la famille E2F4/E2F5, qui semble avoir subi une duplication au cours de l'évolution des Métazoaires,

qui a par la suite donné lieu à deux paralogues proches. La génération d'un arbre phylogénétique à partir de l'alignement multiple de cette famille a permis de déterminer que cette duplication s'est produite vraisemblablement chez l'ancêtre des Vertébrés (*Figure 4-4*). En ce qui concerne la distribution phylogénétique de cette famille, sa présence dans plusieurs clades Ecdysozoaires, qui ne possèdent pas de MCC, suggère une implication de ces facteurs de transcription dans un ou des processus autre(s) que la multiciliation.

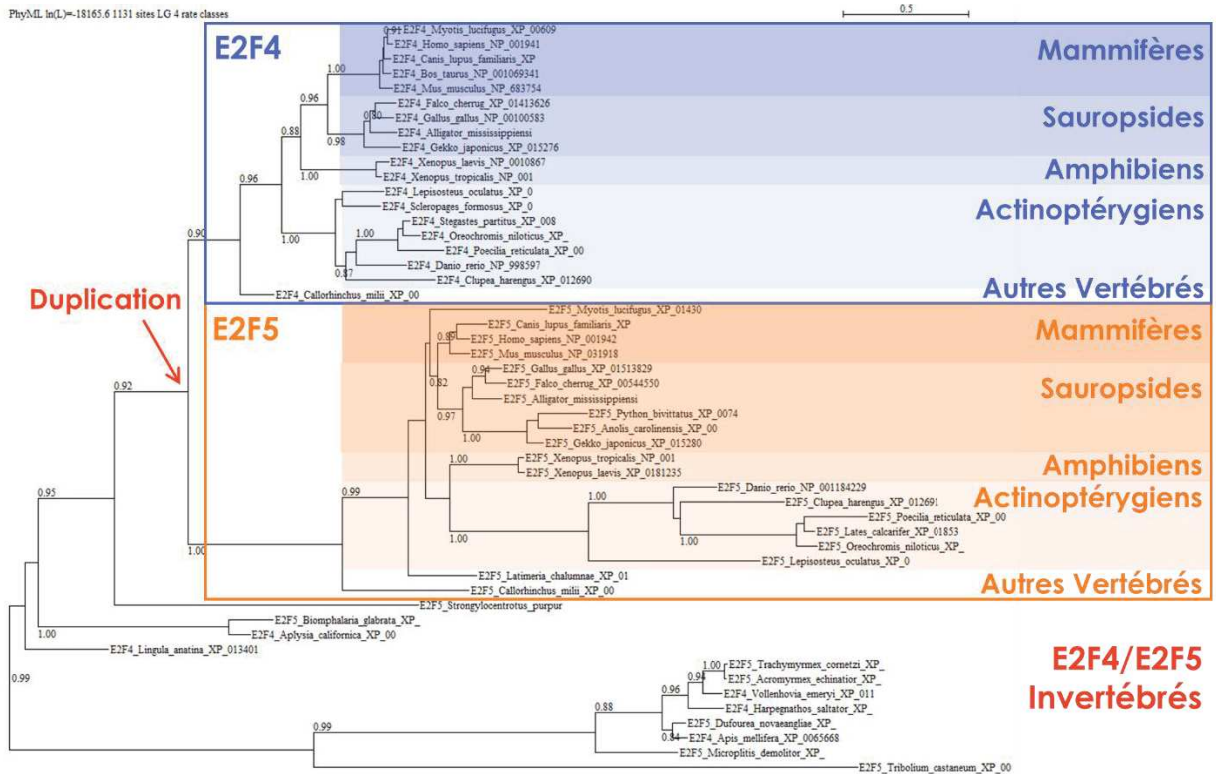

Figure 4-4: Arbre phylogénétique de la famille E2F4/E2F5 réalisé à partir de l'alignement multiple par PhyML (maximum de vraisemblance). La robustesse des branches a été évaluée par la méthode aLRT (Approximate Likelihood Ratio Test) ; seules les valeurs de support supérieures à 0,8 sont représentées. Les relations entre les séquences suggèrent une duplication au début de l'évolution des Vertébrés.

1.3.2. Familles récentes

Parmi les familles les plus récentes, il est possible de distinguer deux sous-groupes de gènes : ceux apparus chez les Vertébrés, et ceux dont la distribution atypique ne permet pas de déterminer le moment exact de leur apparition.

Outre DEUP1 dont nous avons parlé plus tôt, GEMC1, MCIDAS et CDC20B présentent une distribution limitée aux Vertébrés, suggérant une origine « récente » pour ces gènes Aucune séquence homologue de CDC20B n'a été retrouvée chez les Otomorpha, groupe de poissons Téléostéens (*Figure 4-5*) dont fait partie l'organisme modèle *Danio rerio*, confirmant les résultats obtenus lors d'études précédentes (Marcet et al., 2011). Les Otomorpha représentent un grand groupe de poissons osseux ayant émergé il y a environ 145 millions d'années, constitué de plus de 10 000 espèces majoritairement d'eaux douces, et quelques espèces marines (Betancur-R et al., 2013; Straube et al., 2018). De manière intéressante, ces espèces semblent présenter un nombre réduit de cils à la surface de leurs cellules multiciliées ; on en dénombre moins de 16 chez *Danio rerio* (Hansen and Zeiske, 1993; Kramer-Zucker et al., 2005), tandis que près de 140 sont retrouvées chez *Anguilla*

anguilla (Schulte, 1972), qui appartient également à l'infra-classe des Téléostéens. MCIDAS et son régulateur, GEMC1, sont en revanche globalement retrouvés chez l'ensemble des Vertébrés.

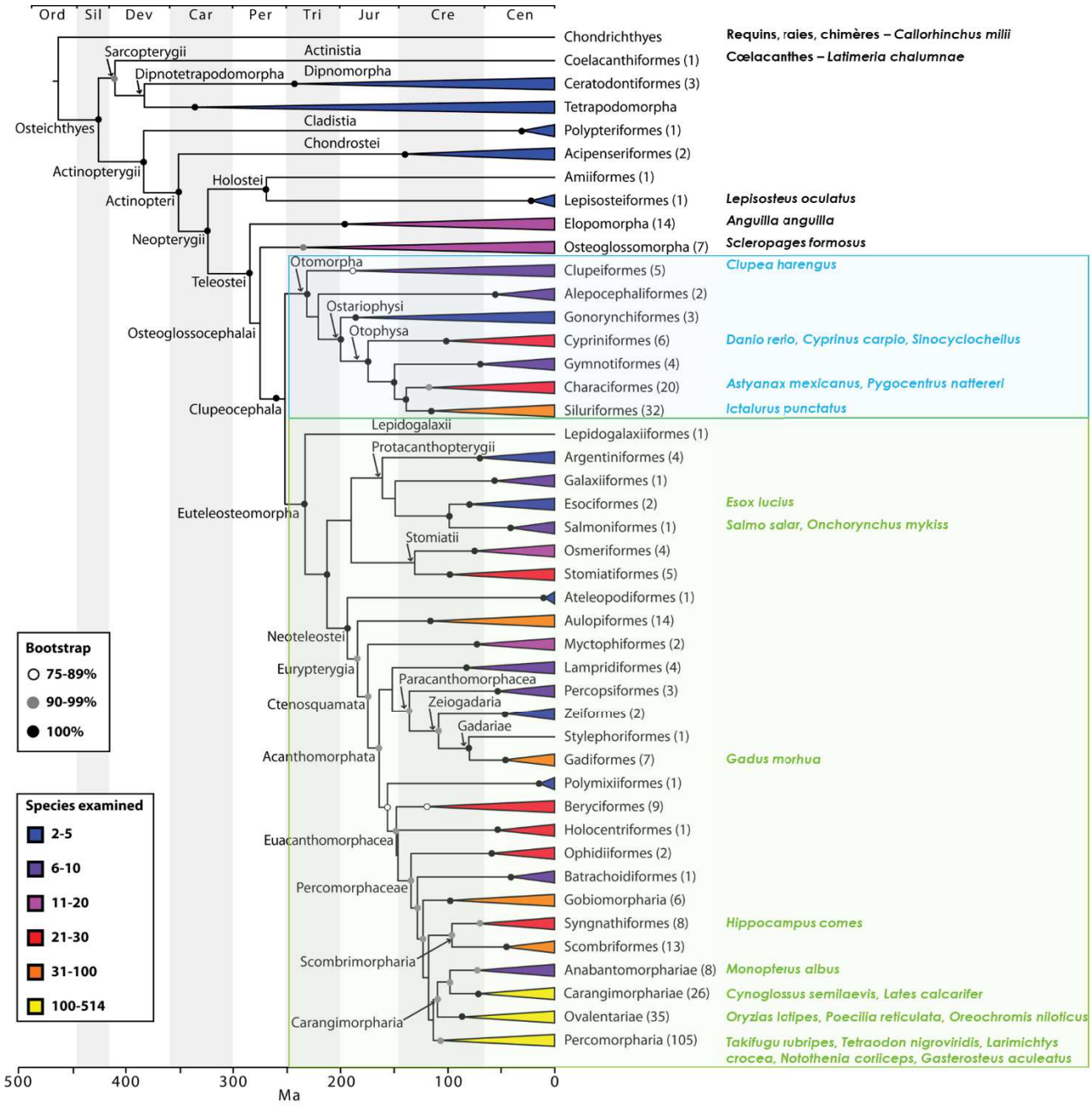

Figure 4-5: Classification majeure des poissons osseux. Les espèces les plus communément utilisées lors de nos analyses sont représentées dans leurs clades, en bleu pour les Otomorpha, en vert pour les Euteleosteomorpha, en noir pour les autres espèces. Adapté de (Betancur-R et al., 2013)

Deux familles de gènes présentent une distribution relativement atypique : CCNO et CCDC78. Des homologues de ces deux gènes sont retrouvés chez les Vertébrés et chez quelques espèces d'Annélides et de Mollusques, ainsi que chez la cione (*Ciona intestinalis*), les Echinodermes et les Hémichordés pour CCDC78. On ne retrouve en revanche pas d'homologues de CCNO chez ces derniers organismes. Le peu d'informations disponibles concernant les Spiraliens non Lophotrochozoaires (voir *Figure 2-1*) et la complexité de la famille des cyclines ne permet pas de déterminer avec certitude l'histoire évolutive de CCNO, qui semble particulièrement complexe. La

présence de CCDC78 et de CCNO chez les Deutérostomiens et les Lophotrochozoaires suggère cependant l'apparition de ces deux gènes dans un ancêtre commun des Deutérostomiens et Protostomiens et une perte secondaire chez les Ecdysozoaires.

2. Conservation des séquences orthologues au cours de l'évolution

Afin d'obtenir plus d'informations concernant l'histoire évolutive de la multiciliation et des gènes qui lui sont centraux, nous avons analysé de façon détaillée les alignements multiples réalisés préalablement.

Pour la majorité des gènes que nous avons étudiés, nous avons pu mettre en évidence des différences de séquences protéiques prévisibles entre les grands groupes taxonomiques : les Mammifères, les Sauropsides (oiseaux et reptiles), les « poissons » Actinoptérygiens et, le cas échéant, les Invertébrés. Nous avons en revanche détecté certains cas de divergence inattendue, et ce dans les séquences d'Otomorpha des alignements de MCIDAS et de CCNO.

2.1. MCIDAS

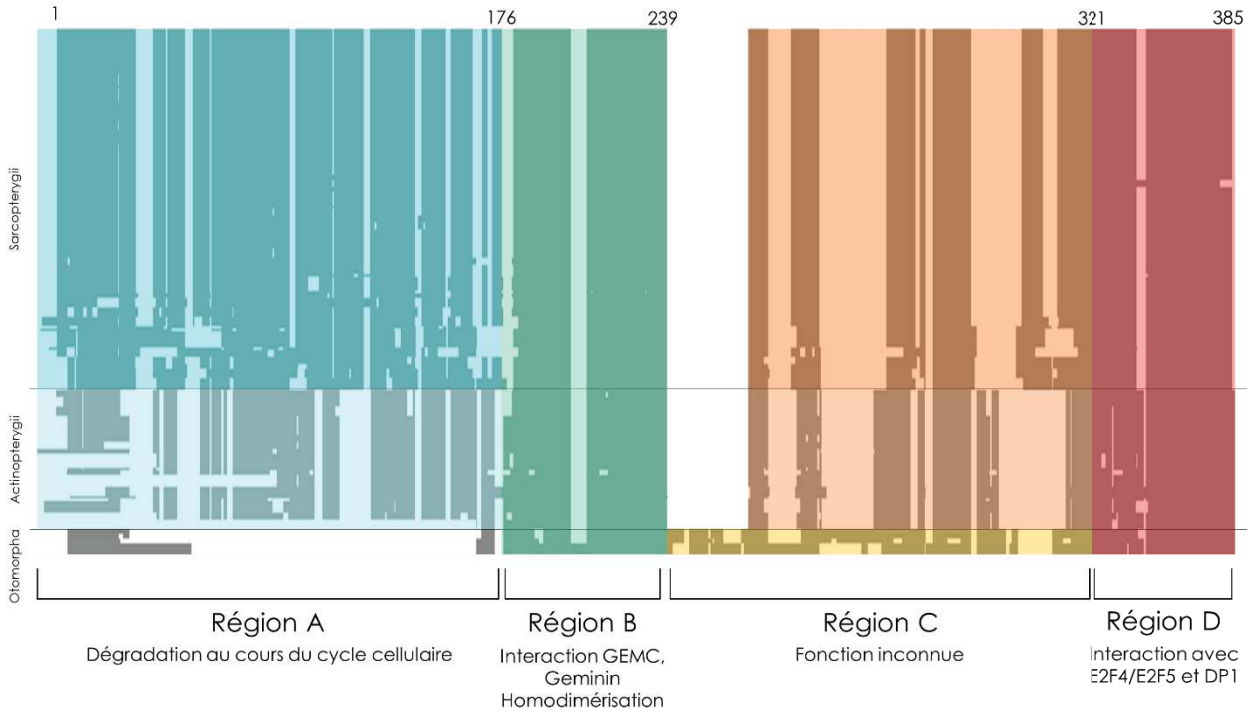

Figure 4-6: Représentation schématique de l'alignement multiple protéique de MCIDAS. Une variation d'intensité d'une même couleur signifie une légère divergence d'une région dans la séquence de certaines espèces. Un changement de couleur (Région C chez les Otomorpha) indique une divergence prononcée de la séquence.

Dans le cas de MCIDAS, il est possible de diviser la séquence protéique humaine en 4 grandes régions : (1) région A (M1-P176), nécessaire et suffisante à la dégradation de MCIDAS au cours du cycle cellulaire (Pefani et al., 2011), (2) région B (L177-D239), nécessaire à l'interaction de MCIDAS avec GEMC1 et la Geminin (impliquée dans la régulation de la réplication de l'ADN), ainsi qu'à son homodimérisation, (3) région C (K240-H320), de fonction encore inconnue, et (4) région D (G321-S385), contenant le domaine TIRT permettant l'interaction avec les facteurs de transcription E2F4/E2F5 et DP1 (*Figure 4-6*). Les régions B et D sont bien conservées chez tous les organismes

étudiés, tandis que les régions A et C varient quelque peu entre les Sarcoptérygiens et la plupart des Actinoptérygiens. Au sein de ces derniers, le groupe des Otomorpha se distingue en revanche très nettement : la région A est presque intégralement absente, suggérant une différence au niveau de la régulation de MCIDAS au cours du cycle cellulaire, et la région C est très divergente (*Figure 4-7*).

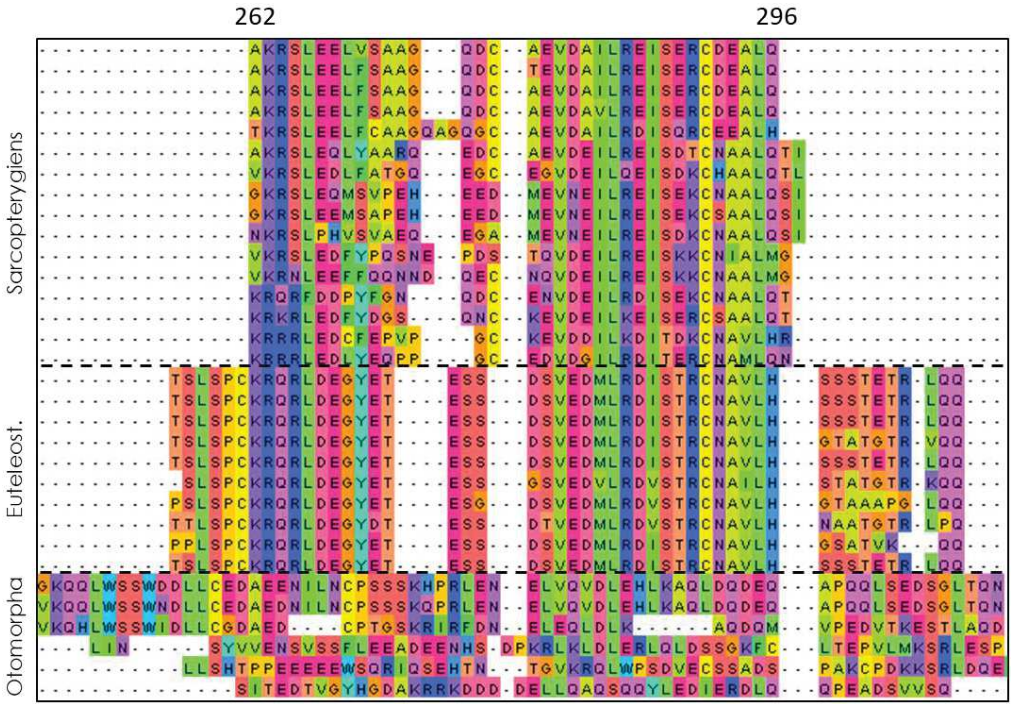

Figure 4-7: Portion de l'alignement multiple de MCIDAS contenant une partie de la région C (positions 262 à 296 sur la séquence humaine). On peut voir une divergence marquée entre les séquences d'Otomorpha et les autres séquences. Eutelost. : Euteleosteomorpha.

2.2. CCNO

La protéine CCNO quant à elle présente des différences chez les Otomorpha tout le long de la séquence (*Figure 4-8*). Sa région N-terminale semble tronquée chez ces espèces, et plusieurs motifs divergent de ceux retrouvés chez les autres Vertébrés. Il n'existe à l'heure actuelle pas d'annotations relatives aux fonctions des différentes régions de CCNO, il n'est donc pas possible d'inférer une potentielle conséquence à ces divergences, si ce n'est qu'elles pourraient avoir un impact sur la régulation de la multiciliation.

Par la comparaison de séquences homologues, nous avons ici pu mettre en évidence des cas de divergence de séquence totalement inattendue chez les espèces appartenant aux Otomorpha qui se distinguent très nettement des autres Actinoptérygiens, et ce pour CCNO et pour MCIDAS. Nous avons également vu précédemment que les Otomorpha ne possédaient pas le gène CDC20B contrairement, là encore, aux autres Actinoptérygiens. Or, les Otomorpha semblent générer une quantité de cils au niveau de leurs MCCs plus réduite que d'autres espèces de poissons. Cette multiciliation réduite pourrait être ainsi liée à l'absence de CDC20B, mais également aux divergences de séquence observées dans deux des familles les plus importantes de la multiciliogénèse, CCNO et MCIDAS.

Figure 4-8: Vue d'ensemble d'une partie de l'alignement multiple de CCNO. Les régions encadrées représentent les zones ayant le plus de divergences chez les Otomorpha. Oto. : Otomorpha ; Actino. : Actinopterygien.

3. Analyse du contexte génomique

Nous avons vu dans le chapitre précédent que des études de l'organisation génomique réalisées sur l'Homme, la souris et le xénope avaient permis de mettre en évidence une co-localisation de MCIDAS, CCNO et CDC20B chez ces organismes. Nous avons donc voulu établir si d'autres gènes de cette région sont également co-localisés dans plusieurs espèces, et si cette synténie s'étend au-delà des Tétrapodes, en particulier chez les « poissons ». Pour cela, le contexte génomique de ces trois gènes a été étudié dans un ensemble d'organismes, couvrant le clade des Vertébrés, allant de l'Homme à la chimère.

Chez la plupart des Vertébrés étudiés, CCNO, MCIDAS et CDC20B sont co-localisés, tout comme GPX8 et la famille miR-449, contenus dans le locus de CDC20B, mais deux autres gènes sont également retrouvés à proximité : GZMA et GZMK, ce qui pourrait indiquer une éventuelle participation de ces gènes à la multiciliation (*Figure 4-9*). Cette conservation s'étend au-delà des Tétrapodes et concerne la majorité des Vertébrés que nous avons étudiés, mettant ainsi en évidence l'existence d'un réel locus multicilié maintenu au cours de l'évolution. De manière intéressante, seuls les Otomorpha ne présentent pas de conservation de synténie de CCNO et MCIDAS, que l'on retrouve sur des chromosomes distincts. Nous avons précédemment mis en évidence la perte de CDC20B chez ces espèces ainsi que l'existence d'une divergence de séquence de CCNO et MCIDAS, l'ensemble de ces résultats semble pointer vers des évènements évolutifs complexes ayant eu pour conséquence une différence de phénotype au niveau des MCCs des Otomorpha.

Au-delà du bloc synténique présenté ci-dessus, nous avons également pu remarquer que les gènes situés à l'extrémité 5' de CCNO étaient conservés chez les Tétrapodes, *Latimeria chalumnae*, et *Callorhinchus milii*, suggérant qu'il pourrait s'agir là du contexte génomique retrouvé chez l'ancêtre des Vertébrés. Ce contexte génomique est en revanche tout autre chez les poissons ; tandis que chez l'Homme CCNO est proche de DHX29, MTEX et PLPP1, chez les Actinoptérygiens, CCNO est retrouvé à proximité de ZCCHC6 et de ISCA1, ce qui suppose l'existence d'un évènement de recombinaison génomique dans cette région. Curieusement, l'organisation retrouvée chez *Danio rerio* inclut des

blocs retrouvés dans les deux groupes d'organismes. En effet, CCNO est retrouvé à proximité du bloc 'poisson' : ZCCHC6 et ISCA1, alors que MCIDAS est co-localisé avec le bloc 'ancestral' contenant DHX29, MTREX et PLPP1. Il apparaît clair qu'au cours de l'évolution des Otomorpha, un certain nombre de réarrangements chromosomiques ont dû avoir lieu pour donner naissance au contexte génomique actuel, et qu'une conséquence majeure de ces évènements semble avoir été la perte de CDC20B.

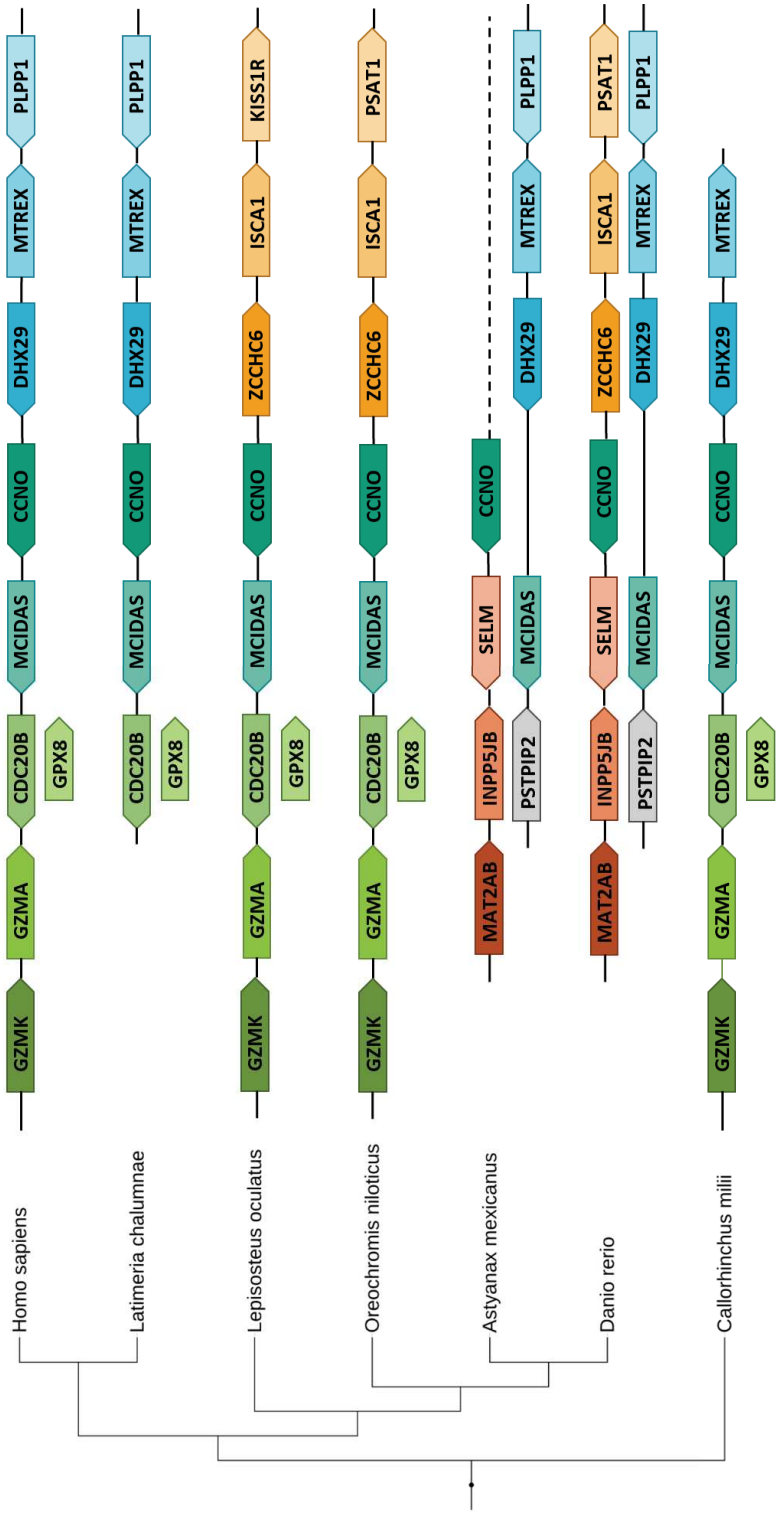

Figure 4-9: Contexte génomique du locus multicilié contenant CCNO, MCIDAS et CDC20B.

4. Recherche de nouveaux gènes par profilage phylogénétique

Sur la base des résultats obtenus précédemment, et en nous appuyant sur les informations disponibles dans la littérature, nous avons voulu établir une stratégie de recherche de nouveaux gènes candidats liés à la multiciliation, et ce par la méthode du profilage phylogénétique et l'outil OrthoInspector (Nevers et al., 2019).

4.1. OrthoInspector : une ressource pour l'analyse comparative

OrthoInspector est un programme de prédiction de relations d'orthologie basé sur les graphes, capable d'inférer des relations entre les gènes protéiques de différentes espèces (Linard et al., 2011). Pour ce faire, OrthoInspector se base sur des résultats de recherche de similarité BLAST effectuées sur des protéomes complets (ici, l'ensemble des séquences protéiques codées par un génome) pour prédire les relations d'orthologie et d'inparalogie entre les gènes d'un ensemble d'espèces choisies.

OrthoInspector est disponible sous forme de programme pour comparer les protéomes complets de l'utilisateur. Des relations d'orthologie sont aussi pré-calculées et accessibles *via* une ressource en ligne hébergeant de nombreux outils permettant la réalisation simple et rapide d'analyses de génomique comparative. La version actuelle, OrthoInspector 3.0 (Nevers et al., 2019), s'appuie sur 4 bases de données de relations d'orthologie, 3 spécifiques à chaque domaine du vivant, et une quatrième regroupant des organismes modèles des 3 domaines permettant de réaliser des analyses transversales. Ainsi, les 3 premières bases s'appuient sur les relations d'orthologie inférées entre les protéomes complets de 711 espèces eucaryotes, 3863 Bactéries et 179 Archées respectivement, tandis que la base transverse contient les relations entre les protéomes de 317 espèces dont 142 Bactéries, 144 Eucaryotes et 31 Archées.

Outre la recherche interactive des orthologues d'une protéine donnée, la ressource en ligne permet l'étude des relations génotype/phénotype de différentes manières. L'ensemble de ces outils est basé à la fois sur les relations d'orthologie prédites par OrthoInspector, ainsi que sur les annotations fonctionnelles disponibles dans le cadre de la ressource *Gene Ontology* (GO) (The Gene Ontology Consortium, 2019). Il est ainsi possible de : (1) étudier l'histoire évolutive d'un trait phénotypique pour tenter de dégager un profil de distribution des gènes qui lui sont liés, (2) établir des liens entre différents gènes d'une même espèce sur la base de leurs distributions phylogénétiques et donc de leur histoire évolutive, et enfin (3) retrouver des gènes liés à un phénotype d'intérêt dont la distribution est connue par application du profilage phylogénétique à proprement parler. Au cours de cette thèse, j'ai pu participer au développement de ce dernier outil en mettant en place la possibilité de réaliser une analyse fonctionnelle des résultats du profilage, par la réalisation d'un test d'enrichissement en termes fonctionnels GO grâce au *webservice* de la base de données Panther (Mi et al., 2019). La publication d'OrthoInspector 3.0 dont je suis co-auteur est présentée en Annexe.

Afin d'identifier de nouveaux gènes candidats impliqués dans la multiciliation, nous avons employé les différents outils proposés par la ressource en ligne.

4.2. Application des outils de profilage à la multiciliation

Pour appliquer une stratégie de recherche de gènes basée sur le profilage phylogénétique, nous avons tout d'abord cherché à établir un profil de distribution multicilié, en nous appuyant à la fois sur les résultats de l'étude évolutive, sur la bibliographie, ainsi que sur l'outil d'analyse d'histoire évolutive d'OrthoInspector.

4.2.1. Etablissement d'un profil multicilié

Nous avons vu dans le chapitre 2 que la multiciliation a été observée dans de nombreux clades Métazoaires, en revanche, l'établissement de la distribution phylogénétique des gènes centraux à la multiciliogénèse a montré que plusieurs gènes, dont MCIDAS et GEMC1, que nous savons importants pour ce processus, ne sont apparus que chez les Vertébrés. D'autres gènes sont au contraire apparus beaucoup plus tôt au cours de l'évolution des Métazoaires, il est donc complexe d'établir un profil représentatif de la multiciliation.

Pour complémenter nos résultats, nous avons employé l'outil d'analyse d'histoire évolutive d'OrthoInspector, en recherchant toutes les protéines associées au terme GO '*multi-ciliated epithelial cell differenciation*'. Cet outil regroupe ensuite, le cas échéant, les protéines dont la distribution est similaire en *clusters.* Avec le terme GO utilisé, nous avons obtenu 8 protéines (CCDC78, MCIDAS, CCNO, DEUP1, CEP63, CEP152, PLK4 et E2F4). Curieusement, 3 des protéines que nous avons jugées centrales à la multiciliation, à savoir GEMC1, E2F5 et CDC20B, ne sont pas associées à ce terme dans la base GO, bien que leur rôle au cours de ce processus ait été démontré. Le manque d'annotation d'E2F5 n'est que peu surprenant, dans la mesure où ce gène participe à de nombreux processus et que la majorité des études réalisées sur le complexe EDM (E2F4/5-DP1-MCIDAS) concernent préférentiellement E2F4. Le manque d'annotation de GEMC1 est en revanche largement inattendu puisque ce gène est bien connu comme étant majeur dans le processus et nécessaire à son bon fonctionnement ; pour autant, les seules annotations GO de GEMC1 concernent son rôle dans la réplication de l'ADN, bien que l'on retrouve également '*cilium assembly*' parmi elles. L'absence d'annotation multiciliée associée à CDC20B est aussi en grande partie inexpliquée, mais pourrait être due au manque d'études concernant le rôle de ce gène dans la multiciliation.

Au sein des 8 protéines retrouvées, aucune ne présente de distribution similaire selon l'outil de *clustering*. En effet, si la plupart de ces protéines sont bien considérées comme spécifiques des métazoaires par OrthoInspector, la distribution des protéines au sein des espèces animales est considérée comme différente. Ceci s'explique en partie par l'hétérogénéité mentionnée plus haut mais aussi par les problèmes de prédictions automatiques de relations d'orthologie liés, notamment, à la qualité insuffisante de certains protéomes.

Nous avons donc décidé de nous concentrer sur l'absence de multiciliation chez les Ecdysozoaires mais aussi sur le caractère atypique de la multiciliation chez les Otomorpha, qui, pour rappel, présentent une multiciliation incomplète, des divergences de séquence anormales au niveau de MCIDAS et CCNO, ainsi qu'une absence de CDC20B et une rupture de synténie du locus multicilié. Notre recherche s'est donc dirigée vers les protéines absentes chez les Ecdysozoaires (non-multiciliés) et les Otomorpha (multiciliation incomplète), mais présentes chez les Euteleosteomorpha (groupe de poissons osseux à multiciliation *a priori* complète).

4.2.2. Recherche par la ressource en ligne

L'outil de recherche par profil phylogénétique d'OrthoInspector permet de chercher des gènes respectant des contraintes de présence ou d'absence dans un ensemble d'espèces choisies. Lorsque l'absence des gènes est demandée dans un ensemble d'espèce, aucun orthologue ne doit être détecté chez ces organismes, tandis que si la présence est demandée dans un ou plusieurs taxons, la présence d'un orthologue dans une seule espèce est suffisante, et ce pour s'affranchir des erreurs de prédictions (gène non prédit), ou des exceptions biologiques.

Dans la base Eucaryote, nous avons cherché les protéines humaines absentes chez les Ecdysozoaires et les Otomorpha, et présentes chez au moins un représentant Euteleosteomorpha. Cette recherche a permis de trouver 424 protéines, dont l'analyse fonctionnelle a pu mettre en évidence une surreprésentation de gènes liés au système immunitaire (*Figure 4-10*). Une étude récente de transcriptomique à grande échelle a mis en évidence l'existence d'une grande diversité du système immunitaire chez les poissons osseux, à la fois en fonction de leurs habitats, mais également entre les espèces Otomorpha et les autres espèces d'Actinoptérygiens, ce qui pourrait expliquer en partie cet enrichissement inattendu (Yi et al., 2019). On retrouve également 22 protéines de fonction inconnue et, de manière intéressante MCIDAS, tandis que CDC20B n'a pas été retrouvé. Dans le cas de MCIDAS, ces résultats peuvent s'expliquer par le fait que les orthologues chez les Otomorpha n'ont pas été détectés par OrthoInspector car leurs séquences sont trop divergentes, ainsi, MCIDAS respecte le profil demandé, bien qu'il n'en soit rien en réalité. Les résultats obtenus pour CDC20B semblent s'expliquer par un manque de prédiction des séquences orthologues dans les protéomes de référence utilisés.

GO Biological Process

Name	Number of Proteins	P-value	False Discovery Rate
defense response	78	3.30E-18	5.272177723128875E-14
regulation of immune system process	87	4.38E-18	3.4931233242839175E-14
cellular metabolic process	65	5.02E-17	2.6693553728553446E-13
metabolic process	81	2.79E-15	1.1133767422634768E-11
immune response	86	7.67E-15	2.4478342803119872E-11
primary metabolic process	68	2.15E-14	5.712189688097681E-11
defense response to other organism	58	3.67E-14	8.381153403179933E-11
inflammatory response	40	4.36E-14	8.70602655127268E-11
G protein-coupled receptor signaling pathway	61	9.22E-14	1.6346485367689448E-10
regulation of immune response	61	1.15E-13	1.829367751266046E-10

Figure 4-10: Enrichissement en termes Gene Ontology des résultats de la recherche par profilage phylogénétique. Seule une partie des résultats de la catégorie « Biological Process » sont ici représentés. On note un enrichissement en plusieurs termes liés à l'immunité.

La base de données Eucaryote ne contenant que 2 espèces d'Otomorpha, nous avons cherché à voir si un profilage phylogénétique appliqué à une plus grande sélection d'espèces de ce groupe permettrait d'obtenir des résultats différents.

4.2.3. Recherche dans une base d'orthologie personnalisée

La ressource OrthoInspector 3.0 est basée sur les protéomes de référence UniProt (voir Chapitre 8: Matériel et Méthodes), parmi lesquels ne se trouvent que 2 Otomorpha. Pour compléter cette

sélection d'espèces, nous avons utilisé des protéomes issus de la base RefSeq du NCBI. Nous avons ensuite généré une base de relations d'orthologie personnalisée à l'aide de la suite programmatique OrthoInspector, ne contenant que des organismes appartenant au Métazoaires, pour un total de 169 espèces. Cette base contient 13 Euteleosteomorpha (*Esox lucius, Gasterosteus aculeatus, Larimichthys crocea, Lates calcarifer, Maylandia zebra, Nothobranchius furzeri, Oreochromis niloticus, Oryzias latipes, Poecilia formosa, Stegastes partitus, Takifugu rupribes, Tetraodon nigroviridis, Xiphophorus maculatus),* et 7 Otomorpha (*Astyanax mexicanus, Clupea harengus, Cyprinus carpio, Danio rerio, Ictalurus punctatus, Pygocentrus nattereri, Sinocyclocheilus anshuiensis*), contre 8 et 2 respectivement dans la base en ligne.

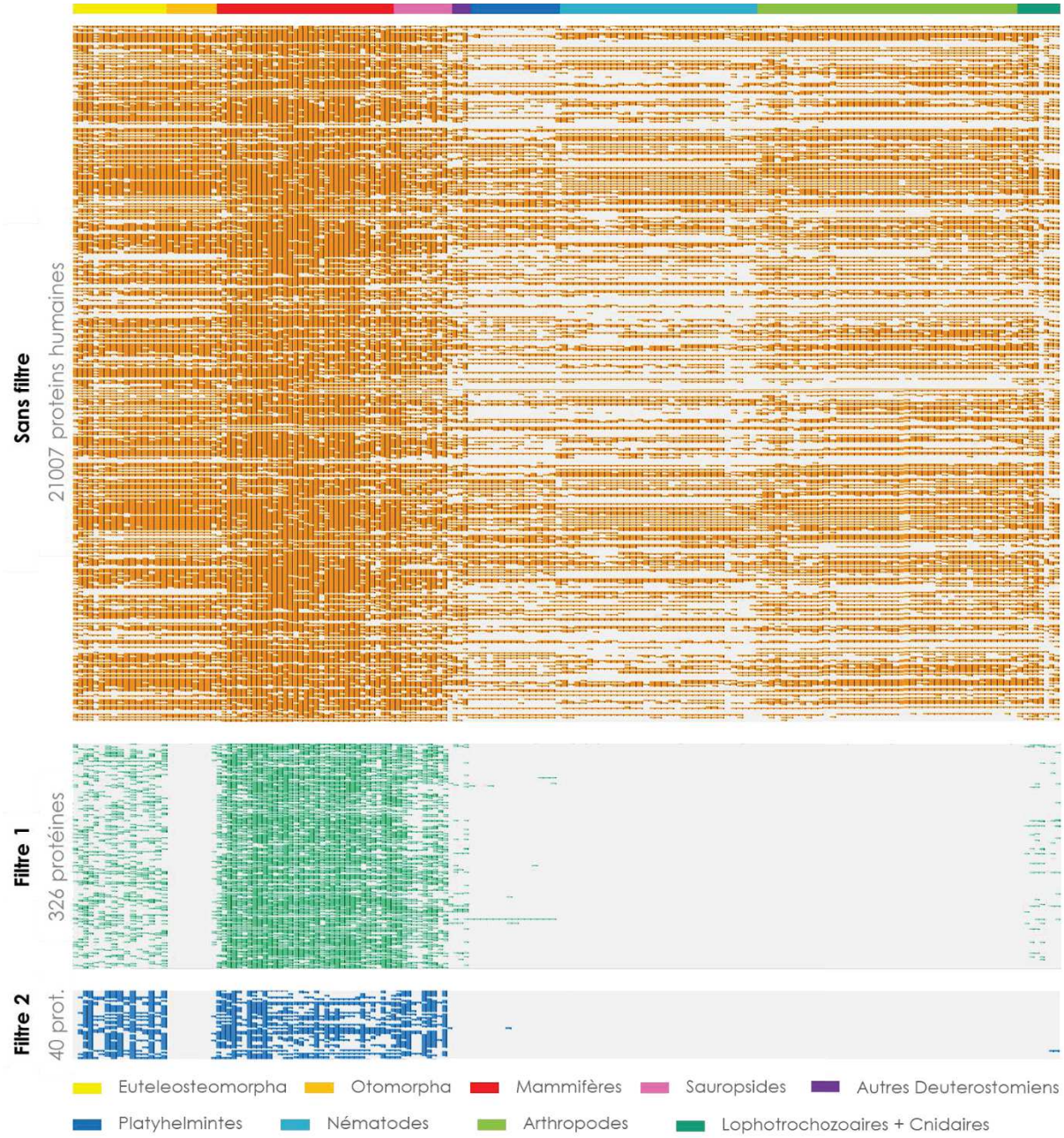

Figure 4-11: Profils phylogénétiques des protéines humaines chez 169 Métazoaires. Filtre 1: Protéines absentes chez Otomorpha, Nématodes et Arthropodes, présentes chez au moins un Euteleosteomorpha; Filtre 2: protéines absentes Otomorpha, Nématodes et Arthropodes, présentes chez au moins 50% des Euteleosteomorpha.

Afin d'effectuer une recherche de profil, nous avons utilisé le programme Phyligrane, développé au laboratoire, pour générer dans un premier temps une matrice phylogénétique contenue dans un fichier CSV, dont les lignes représentent les protéines, et les colonnes les espèces (*Figure 4-11*). Tout comme pour la recherche par la ressource en ligne OrthoInspector, nous avons choisi de chercher les protéines absentes des Ecdysozoaires et des Otomorpha, et présentes chez au moins une espèce Euteleosteomorpha (Filtre 1 dans la Figure 4-11). De cette manière, 326 protéines ont été sélectionnées, dont 178 communes à la recherche en ligne. L'enrichissement en termes GO a également fait ressortir une surreprésentation des termes liés à l'immunité, ainsi que des termes liés à la phagocytose. Parmi les 326 protéines, près de la moitié ne sont retrouvées que chez un seul représentant des Euteleosteomorpha, pouvant être lié à des erreurs de prédiction à la fois de la part du programme d'inférence d'orthologie, mais également lors de l'annotation des génomes.

Pour obtenir une sélection plus restreinte, et éventuellement plus qualitative, de protéines à étudier, nous avons choisi de rechercher des profils de protéines présentes chez au moins la moitié des représentants des Euteleosteomorpha (Filtre 2 dans la Figure 4-11). Cette recherche nous a permis de retrouver 40 protéines, parmi lesquelles 25 sont impliquées là encore dans la réponse immunitaire. On ne retrouve en revanche pas MCIDAS, des orthologues ayant été prédits dans certaines des espèces Otomorpha que nous avons rajoutées. Les 15 autres protéines sont impliquées dans divers processus biologiques, et de manière intéressante, deux n'ont pas de fonction renseignée et pourraient être étudiées de manière approfondie : ANKRD61, et C4orf50.

5. Discussion

L'étude évolutive des protéines impliquées dans la multiciliation nous a permis dans un premier temps d'apprécier entièrement la complexité du processus, déjà aperçue dans le chapitre 2. En effet, nous avions vu que la régulation de la multiciliation était largement basée sur le recyclage des mécanismes existant, en particulier ceux utilisés lors de la duplication du centrosome, et l'implication de gènes apparus très tôt dans l'évolution des Métazoaires tels que la famille E2F4/E2F5, CEP63 ou CEP152 a permis de le souligner (*Figure 4-12*). Notre étude a également permis de confirmer l'apparition relativement récente de la voie d'amplification par le deutérosome, qui semble être apparue chez les Vertébrés, bien que l'on retrouve une ambiguïté concernant l'origine de CCNO et de CCDC78 qui, eux, semblent être apparus chez l'ancêtre des Deutérostomiens et Protostomiens. De même, nous avions soulevé le point des différences potentielles de régulation entre les différentes espèces, et notre étude sur la distribution phylogénétique des gènes centraux montre clairement l'apparition de plusieurs gènes de façon tardive, chez l'ancêtre des Vertébrés. Sachant que la multiciliation a été observée dans diverses espèces au-delà des Vertébrés, la question se pose encore de savoir par quels mécanismes ces espèces sont-elles capables de générer plusieurs dizaines de cils motiles. Enfin, cette étude nous a également permis de mettre en évidence l'étendue de la conservation du 'locus multicilié' chez les Vertébrés, qui n'avait jusqu'alors été confirmée que chez les mammifères et le xénope.

Dans un second temps, les analyses de ces 10 familles de gènes à différents niveaux nous ont permis de faire ressortir de nombreuses particularités chez le groupe des Otomorpha. L'étude du contexte génomique au niveau du 'locus multicilié' contenant les gènes MCIDAS, CCNO et CDC20B ainsi que les familles miR-34/449 semble indiquer l'existence de réarrangements chromosomiques

complexes dans cette région chez les Otomorpha, provoquant ainsi une rupture de synténie qui semble avoir été accompagnée de la perte totale de CDC20B et de la famille miR-449. De plus, des divergences au niveau des séquences protéiques de MCIDAS, régulateur central, et CCNO, nécessaire à la formation des deutérosomes et donc à l'amplification des corps basaux, ont été mises en évidence par l'étude approfondie d'alignements multiples. L'ensemble de ces modifications pourraient avoir un lien avec l'apparente incomplétude du phénotype multicilié chez ces organismes, qui semblent être capables de générer un nombre limité de cils comparés à d'autres espèces de poissons osseux. Nous avons choisi pour le reste de ces travaux de nous concentrer sur ces caractéristiques propres aux Otomorpha et de limiter notre étude aux mécanismes régissant la multiciliogénèse chez les Vertébrés.

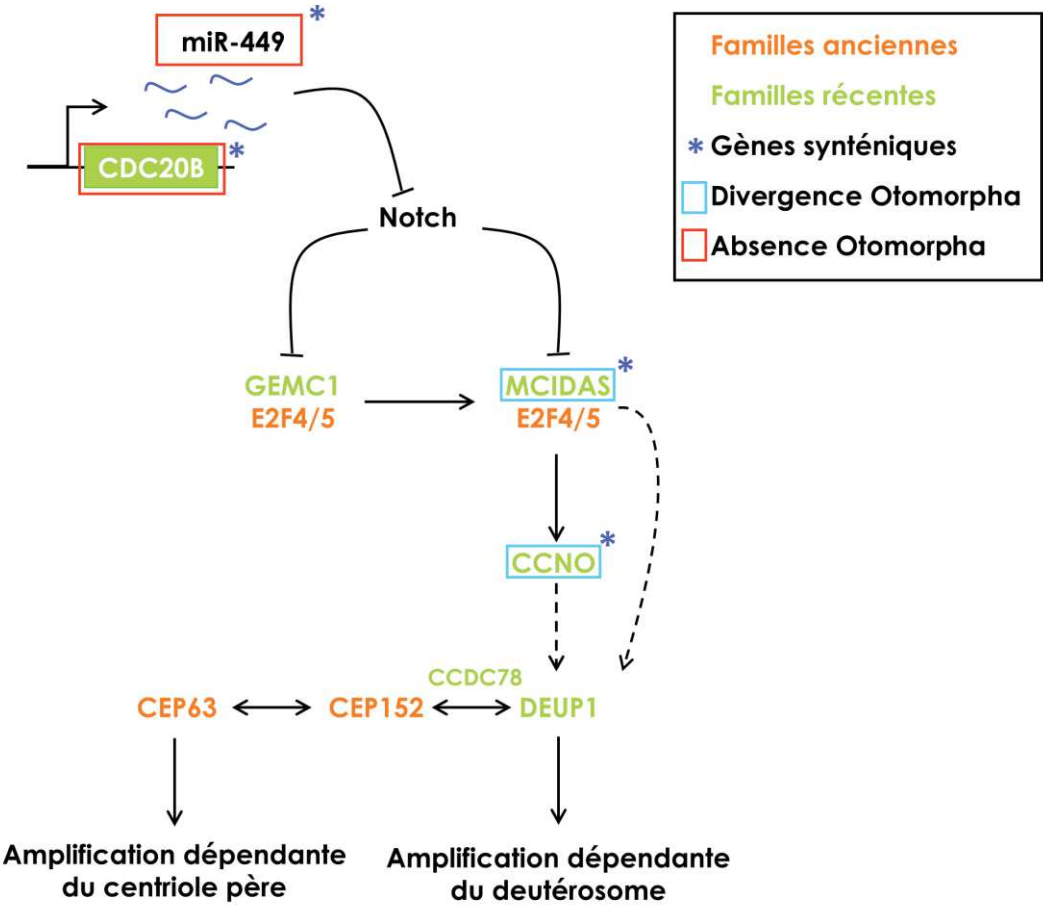

Figure 4-12: Résumé graphique des résultats de notre analyse évolutive de la multiciliation.

L'hétérogénéité de distribution des différents gènes déjà connus pour être impliqués dans la multiciliation n'a pas permis d'établir un profil consensus pour l'application de la méthode du profilage phylogénétique. En conséquence, les profils appliqués restent peu spécifiques, et il apparait clair que l'histoire évolutive complexe de ce processus n'est pas aussi bien adaptée au profilage phylogénétique qu'a pu l'être celle du cil lors des différentes études présentées dans le chapitre précédent. En effet, le profil que nous avons établi ne présente une absence que dans deux taxons, les Otomorpha et les Ecdysozoaires, et il est très fortement probable que plusieurs processus diffèrent entre les espèces que nous avons considérées. Ainsi, l'enrichissement GO majoritairement obtenu lors de nos résultats concerne des termes liés au système immunitaire. Au-delà de ce problème de spécificité, nous avons également été confrontés à un problème récurrent des bases de

données publiques, qui est la qualité des protéomes disponibles. Nous avons néanmoins pu identifier deux gènes de fonction inconnue, ANKRD61 et C4orf50, absents chez les Otomorpha, pouvant constituer deux nouveaux gènes candidats, qu'une étude expérimentale permettrait de valider.

De manière générale, la distribution taxonomique des différents gènes multiciliés n'est pas assez atypique pour générer des résultats concluants en ce qui concerne l'identification de nouveaux candidats, et la recherche de gènes absents chez les Otomorpha semble être un critère trop strict, au vu de la présence, bien qu'imparfaite, de cellules multiciliées chez ces espèces. Il nous a fallu établir une nouvelle stratégie de recherche, cette fois-ci basée sur la conservation atypique observée dans les séquences de MCIDAS et CCNO et sur l'exploitation d'un nouvel aspect potentiel des relations génotype/phénotype, qui est l'impact des variations d'une région protéique sur le phénotype. Dans cette optique, nous avons développé un nouvel outil de génomique comparative dédié à la comparaison de protéomes et au profilage phylogénétique multi-niveaux qui sera présenté dans le chapitre suivant.

Chapitre 5 : BLUR, un outil de profilage multi-niveaux

Les relations génotype/phénotype sont étudiées depuis maintenant bien longtemps, principalement sous deux angles : (1) au niveau intraspécifique, par l'étude des variants entre individus ou entre populations, qu'ils soient de type SNV (*Single Nucleotide Variant*) ou de type variants structuraux et (2) au niveau interspécifique, essentiellement par l'étude des pertes et des gains de gènes, comme nous l'avons déjà mentionné au chapitre 3 avec les comparaisons de répertoires de gènes, par approche soustractive ou par profilage phylogénétique. En revanche, il devient de plus en plus clair que ces corrélations génotype/phénotype sont beaucoup plus complexes qu'un simple gain ou perte de gène. En effet de nombreuses études soulignent la variabilité existante dans des familles de protéines, en particulier au niveau des domaines fonctionnels, connus pour évoluer au cours du temps et créer de nouvelles fonctions. Certaines de ces variations ont été liées à des divergences de phénotype et ont permis de faire ressortir la complexité certaine du Vivant et des relations entre le génotype et le phénotype. Dans cette optique, nous avons cherché à mettre au point une méthode capable de détecter des variations génotypiques à plusieurs niveaux dans le cadre de comparaison de protéomes complets.

Dans ce chapitre, nous commencerons par souligner l'intérêt de prendre en compte les variations des gènes ou des protéines à un niveau plus fin que celui de la séquence complète, et présenterons les principales approches développées à cet effet. Nous introduirons ensuite BLUR (*Blast Unexpected Ranking*), un outil que nous avons développé dans le but de réaliser des études de génomique comparative à plusieurs niveaux de granularité, avant de conclure sur les futurs développements prévus pour cette ressource.

1. Variations au niveau du domaine ou de la région

Bien que la majorité des approches d'inférence d'orthologie et de profilage phylogénétique soit basée sur les séquences complètes de gènes ou de protéines, il serait parfois pertinent de prendre en compte la nature modulaire de l'évolution des gènes, et de considérer les régions fonctionnelles comme unités évolutives. Nous diviserons ces régions fonctionnelles en trois catégories : (1) les domaines protéiques, parties fonctionnelles de séquences possédant une structure tertiaire définie et pouvant être autonome, (2) les motifs, petites parties de séquences conservées et formées d'uniquement quelques acides aminés, et (3) les régions protéiques, de taille variable et n'ayant pas obligatoirement de fonction ou de structure définie.

1.1. Evolution modulaire des protéines

De nombreuses études ont été réalisées ces dernières années pour quantifier et caractériser la nature modulaire de l'évolution des protéines, montrant de cette manière la fréquence des réarrangements architecturaux et des évènements de gain, de perte ou de duplication de domaines protéiques (Buljan and Bateman, 2009; Dohmen et al., 2020; Moore and Bornberg-Bauer, 2012; Stolzer et al., 2015). En effet, de nombreuses protéines, en particulier chez les Eucaryotes, sont composées de plusieurs domaines, combinés à la façon de briques pour créer de nouvelles fonctions (Lees et al., 2016). De manière intéressante, des architectures différentes ont été observées entre les orthologues de différentes espèces, avec une apparente complexité des structures corrélée à celle propre à chaque organisme (Koonin et al., 2004). Des divergences d'architecture ont également été

observées entre des orthologues d'espèces proches, notamment chez les Mammifères ou entre des membres du genre *Drosophila* (Forslund et al., 2011; Moore et al., 2013). Il a ainsi été montré que dans le clade *Drosophila*, environ 36% des familles de gènes présentent des réarrangements (Wu et al., 2012) et des estimations suggèrent qu'entre certaines espèces, de 10 à 50% des orthologues peuvent différer en termes d'architecture (Sonnhammer et al., 2014).

Alors que les variations de domaines entre orthologues peuvent avoir des conséquences non-négligeables sur la fonction des protéines, il en est de même pour des divergences plus subtiles, concernant des régions protéiques, des motifs ou encore des acides aminés uniques. L'importance de la conservation des acides aminés au niveau des sites catalytiques d'enzymes n'est plus à démontrer, mais une étude a permis de mettre en évidence que de simples substitutions au niveau de leur site actif créent des homologues inactifs de ces enzymes, qui, par la suite, développent de nouvelles fonctions, bien souvent régulatoire (Bartlett et al., 2003). De mêmes, plusieurs études ont montré que des variations de séquences, allant de petits motifs de quelques acides aminés à des régions plus étendues, au niveau des gènes homéotiques chez les Arthropodes étaient liées à des modifications de développement observables entre les différentes espèces (Löhr et al., 2001; Ronshaugen et al., 2002; Shiga et al., 2002).

Il est donc clair que dans l'étude des relations génotype/phénotype, la seule présence ou absence d'un gène, bien qu'informative, n'est pas toujours suffisante à expliquer un trait ou un processus, il devient alors indispensable de prendre en compte les variations observables à un niveau de granularité plus fin, qu'il soit celui du domaine, de la région ou du motif.

1.2. Approches

L'ensemble des divergences existant entre les séquences d'orthologues, qu'elles concernent les domaines et leurs réarrangements ou les variations de région ou de motif, peuvent influer les prédictions de relations d'homologie et les rendre plus complexes. Un gène issu d'une fusion ou présentant un réarrangement complexe pourra par exemple être similaire de façon significative à plusieurs familles de gènes, sans que l'on puisse déterminer si un lien d'orthologie existe. A l'inverse, un gène trop divergent au niveau d'une partie de sa séquence pourra ne pas être détecté en tant qu'orthologue ; nous l'avons vu par exemple dans le chapitre précédent, les divergences de séquences de MCIDAS observées chez les Otomorpha n'ont pas permis leur détection automatique en tant qu'orthologues de la séquence humaine.

Bien que l'importance des domaines dans l'évolution des protéines ait été établie il y a déjà plusieurs années, et que la nécessité de développer des méthodes les incluant soit reconnue, très peu d'approches actuelles les prennent en compte. En ce qui concerne la prédiction d'orthologie, on retrouve néanmoins deux exceptions. La base de données de génomes microbiens MBGD (Uchiyama et al., 2019) se base sur l'algorithme DomClust (Uchiyama, 2006), qui permet de construire des groupes d'orthologues par *clustering* hiérarchique appliqué aux domaines constituants chaque séquence. Le second outil, Domainoid (Persson et al., 2019), construit des groupes d'orthologues sur la base de leurs domaines. Pour cela, il extrait dans un premier temps les domaines définis par PFAM (El-Gebali et al., 2019) des protéomes, pour ensuite appliquer à leurs séquences l'algorithme d'inférence d'orthologie InParanoid (Sonnhammer and Östlund, 2015). Cet outil a permis de prédire des orthologues n'ayant pas été retrouvés par une approche classique au niveau de la séquence

complète, montrant bien l'intérêt de combiner ces deux types de stratégies. Plus récemment, Han et collaborateurs (Han et al., 2020) ont développé une nouvelle approche permettant l'identification de familles de gènes selon des relations d'homologie partielles entre protéines. Pour cela, ils se sont basés sur les modules protéiques et le travail réalisé par Wu et collaborateurs (Wu et al., 2012) sur l'évolution modulaire des protéines afin de permettre une meilleure compréhension des mécanismes évolutifs régissant le monde du Vivant. Rapidement, leur méthode se base sur un arbre phylogénétique et les architectures des protéines en termes de modules pour récapituler les évènements évolutifs ayant eu lieu et regrouper des familles de protéines homologues sur la base de leurs architectures.

Au-delà de la prédiction de relations d'orthologie, d'autres méthodes ont été développées pour apprécier les variations de séquences au sein de familles de gènes ou de protéines. La base de données PhyloPro2.0 permet par exemple d'explorer des profils phylogénétiques et de visualiser la conservation de domaines PFAM par une carte de chaleur, et ce chez 164 espèces eucaryotes (Cromar et al., 2016). PROBE, une ressource développée dans l'équipe, permet l'identification de domaines et régions variables au sein d'une famille de protéines par l'analyse de blocs conservés dans des alignements multiples (Kress et al., 2018). Dernièrement, FAS, une approche permettant de comparer les architectures de deux protéines et de mesurer leur similarité, a été développée [https://github.com/BIONF/FAS].

Dans l'ensemble, les outils disponibles à l'heure actuelle ne rendent pas bien compte des variations pouvant exister à un niveau plus petit que celui du gène complet ; certains tentent malgré tout d'intégrer la notion de domaine dans leurs approches. Sauf exception, ces approches sont basées sur les annotations de la base de données PFAM, ce qui ne permet pas d'analyser des séquences non annotées ou des domaines non caractérisés, ni des divergences concernant seulement une région ou un motif. Or, les exemples que nous avons vus dans le chapitre précédent concernent uniquement des régions non annotées. Sur la base de ce constat, nous avons développé BLUR (*Blast Unexpected Ranking*), une nouvelle approche de génomique comparative qui vise à détecter toute variation de conservation jugée anormale entre deux groupes d'espèces d'intérêt, qu'elle concerne la protéine entière, ou seulement une partie de la séquence, et ce au niveau des protéomes dans leur intégralité. Si cette approche a été conçue pour permettre l'étude approfondie de la multiciliation, elle peut être appliquée à de multiples questions biologiques, que ce soit chez les Eucarytes ou les Procaryotes. Un site web [http://lbgi.fr/blur/] associé à une base de données a été développé pour permettre l'accès global à cette ressource et permettre une utilisation par la communauté, ce qui a donné lieu à une publication, actuellement en cours de révision pour le journal Genome Biology and Evolution, dont le manuscrit est présenté au point suivant.

2. Publication: *Proteome-scale detection of differential conservation patterns at protein and sub-protein levels with BLUR*

Proteome-scale detection of differential conservation patterns at protein and sub-protein levels with BLUR

Audrey Defosset[1], Arnaud Kress[1], Yannis Nevers[1,2,3,4], Raymond Ripp[1], Julie D. Thompson[1], Olivier Poch[1] and Odile Lecompte*,[1]

[1]Complex Systems and Translational Bioinformatics, ICube UMR 7357, Université de Strasbourg, Strasbourg, France

[2]SIB Swiss Institute of Bioinformatics, Lausanne, Switzerland

[3]Department of Computational Biology, University of Lausanne, Lausanne, Switzerland

[4]Center for Integrative Genomics, University of Lausanne, Lausanne, Switzerland

*Corresponding author: odile.lecompte@unistra.fr

Data deposition: Supplementary data are available online.

Abstract

In the multi-omics era, comparative genomics studies based on gene repertoire comparison are increasingly used to investigate evolutionary histories of species, to study genotype-phenotype relations, species adaptation to various environments, or to predict gene function using phylogenetic profiling. However, comparisons of orthologs have highlighted the prevalence of sequence plasticity among species, showing the benefits of combining protein and sub-protein levels of analysis to allow for a more comprehensive study of genotype/phenotype correlations.

In this article, we introduce a new approach called BLUR (Blast Unexpected Ranking), capable of detecting genotype divergence or specialization between two related clades at different levels: gain/loss of proteins but also of sub-protein regions. These regions can correspond to known domains, uncharacterized regions, or even small motifs. Our method was created to allow

two types of research strategies: (1) the comparison of two groups of species with no previous knowledge, with the aim of predicting phenotype differences or specializations between close species or (2) the study of specific phenotypes by comparing species that present the phenotype of interest with species that do not. We designed a website to facilitate the use of BLUR with a possibility of in-depth analysis of the results with various tools, such as functional enrichments, protein-protein interaction networks, and multiple sequence alignments. We applied our method to the study of two different biological pathways and to the comparison of several groups of close species, all with very promising results.

BLUR is freely available at http://lbgi.fr/blur/

Keywords: comparative genomics, evolution, sequence analysis, genotype/phenotype relations

Introduction

Technological advances in recent years have given rise to an ever-increasing amount of sequencing data, providing opportunities to capitalize on the available diversity of living organisms to study the evolution of various biological processes. Data from genome sequencing have been used to establish correlations between genotype and phenotype to improve gene function prediction. Full proteomes of distinct species can be compared to identify genes that are conserved, gained or lost, and could be linked to phenotypical differences or species specificity. Comparison of genes that are present or absent in various species can not only help with understanding evolution and the adaptation of living organisms to different environments, but it is also a useful comparative genomics approach for the inference of gene function. It is assumed that genes participating in the same mechanism will generally be conserved and lost together through evolution, and that functionally linked genes often present similar phylogenetic

46 distributions (Pellegrini et al. 1999). It is thus possible to infer gene function and associate genes
47 with various processes by matching a phenotype distribution to that of a set of genes. This
48 method has been successfully applied to various processes and organelles, such as cilia (Li et al.,
49 2004; Dey et al., 2015; Nevers et al., 2017), mitochondria (Cheng & Perocchi 2015), thermophily
50 (Jim 2003) and the DOXP/MEP metabolic pathway (Cunningham et al. 2000).

51 While phylogenetic profiling is a very insightful approach to explore evolutionary
52 histories of species at the gene/protein level, it does not account for the modular nature of protein
53 evolution. Many studies have quantified and characterized protein domain evolution, showing
54 that domain gains and losses are quite common, and that sequence architectures are often
55 rearranged between taxa, participating in lineage specific adaptations (Lees et al. 2016; Moore &
56 Bornberg-Bauer 2012; Zmasek & Godzik 2011; Dohmen et al. 2020). Such sequence divergences
57 have been observed even between orthologs of closely related species, such as members of the
58 genus *Drosophila* (Forslund et al. 2011; Moore et al. 2013). It has also been shown that sequence
59 divergence on the scale of a region or a small motif can have non-negligible impact, such as in
60 homeotic genes in arthropods. Variations in sequences in several Hox orthologs have indeed been
61 linked to developmental differences between various arthropod species (Ronshaugen et al. 2002;
62 Löhr et al. 2001; Shiga et al. 2002). It is expected that such interspecific sequences divergences
63 can also be observed when dealing with proteins participating in multiple processes, such as
64 moonlighting proteins, which can exhibit two or more biological functions (Jeffery 1999). So far,
65 several hundreds of proteins have been found to be involved in more than one process, and many
66 more may exist that remain to be discovered (Mani et al. 2015).

67 Differences in sequences at various levels (motif, block or domain) between orthologs can
68 be challenging for traditional orthology inference methods, making it difficult to predict the

correct relations between divergent sequences. In terms of comparison of gene repertoires, this means that the regions variations, losses or gains that may be observed in certain species will not only make it hard to predict the true orthologous relations, but also to properly annotate their function through co-occurrence methods. Consequently, while it is important to consider gain and loss of complete genes, it is also crucial to take into account the domain composition and sequence divergences between orthologs to gain better insight into the complex relations between phenotype and genotype, and potentially predict specializations and phenotype divergences between closely related species.

Some attempts have been made to extend the classical gene-level phylogenetic profiling approach, either to fixed-length protein segments (Kim and Subramaniam, 2005) or to conserved domains (Pagel et al. 2004; Persson et al. 2019) found in databases such as PFAM (El-Gebali et al. 2019) or SMART (Letunic & Bork 2018), in order to infer domain interactions and help identify physical and functional relationships between proteins. The PhyloPro2.0 phylogenetic profile database allows the visualization of PFAM domain conservation through heatmaps generated for 164 eukaryotes and can display up to 1000 genes at a time (Cromar et al. 2016). Some resources have also been designed that facilitate the identification of variable domains in protein families, such as PROBE, that allows users to find conserved blocks in a multiple sequence alignment (Kress et al. 2018), or TreeDom, a web tool designed to graphically represent domain architecture evolution in multi-domain proteins (Haider et al. 2016). Other software tools, such as DoMosaics (Moore et al., 2014) or DomArch (Vera-Parra et al.,2016), have been developed to work in conjunction with available domain annotation services, and enable the comparison, analysis and visualization of the evolution of domain architectures.

Generally, these tools are limited to the study of individual genes or gene families, and are not adapted to the study of complete proteomes. The programs also mostly focus on well characterized functional domains such as PFAM, which prevents the analysis of uncharacterized domains or of regions without domain annotations, and do not allow the detection of subtle sequence divergence, which has been shown to alter domain function entirely, even when the change affects only one amino acid (Anderson et al. 2016). The obvious need for a high-throughput method that would allow for the search of lineage-specific conservation patterns, at both the gene and sub-gene levels, in a complete proteome led us to develop a novel approach based on BLAST homology searches (Camacho et al. 2009), that is capable of detecting genotype divergence or specialization between two related lineages in a wide selection of organisms.

Here, we present the BLAST Unexpected Ranking (BLUR) method, a rapid, proteome-scale approach to analyze the protein conservation of two related taxa in order to detect atypical patterns. BLUR is designed to facilitate the study and understanding of genotype/phenotype relations by providing information both on the gain/loss of complete proteins and on the specific divergences of sub-protein regions, ranging from small motifs to complete domains. It can be used both as an exploratory tool to compare two groups of interest with no previous knowledge, or to study specific phenotypes and identify proteins linked to them. To facilitate the exploitation of results, a website was developed that includes a variety of resources for in-depth analyses, including functional annotation, interaction networks or multiple sequence alignment visualization. Finally, we demonstrate the usefulness of our method, by applying it to different use cases, notably the detection of cilia-related proteins in Eukaryotes, and sulfur oxidation related proteins in Bacteria, as well as by using it to compare various groups of species in different life domains.

Material and Methods

Definition of differential conservation

We define differential conservation as the unexpected divergence that can be observed between taxonomic groups in an otherwise well-conserved protein family, which can correspond to a diverging or missing region of variable size in the sequences of specific species. This can be due to varying evolutionary pressures between clades, resulting in a higher rate of sequence evolution leading to variations along the protein sequence or in the complete or partial gain/loss of one or several proteins. Complete protein gain/loss can be detected either by searching for homologous sequences through BLAST searches, or by predicting orthologous relations with dedicated programs such as OrthoInspector 3.0 (Nevers et al. 2019). In the case of partial protein gain/loss or sequence divergence, relative conservation between groups in a protein family can be inconsistent with what is expected based on the species tree. The proposed approach is based on the analysis of the respective conservation of two groups of closely related species compared to a more distant query species used as a reference. For instance, we can estimate the relative conservation of two groups of Teleost fish (e.g. Otomorpha and Euteleosteomorpha) to *Homo sapiens*. The two groups of Teleost fish are expected to have a similar conservation when compared to human. If one group of Teleost fish is significantly closer to human than the other in a given protein family, it may reflect a case of differential conservation.

For the two chosen related groups of species, a comparison is done to establish a baseline behavior of conservation in the whole proteome, which can then be used to highlight cases where the conservation is atypical. Relative conservation and taxonomic proximities compared to a query species can be assessed through BLAST homology searches. By using a more distant

reference species, we ensure that in the case of a well-conserved protein, the two selected taxa of interest should be indistinguishable from one another in a BLAST result, while in proteins presenting an atypical conservation pattern, there should be a clear separation between the two groups (Figure 1).

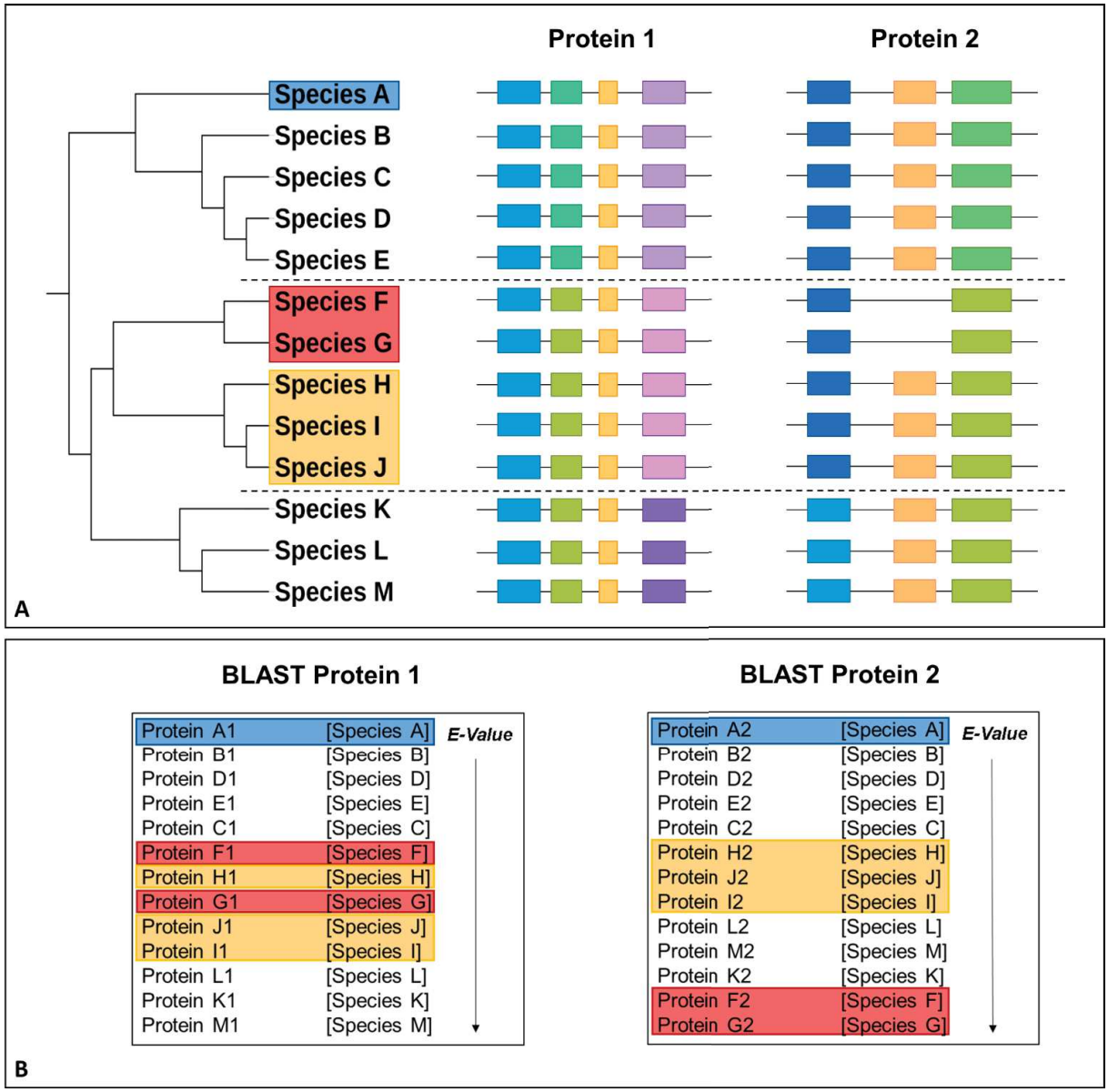

Fig 1: *Schematic representation of the proposed approach. (**A**) The relative conservations for two proteins (1 and 2) in 13 different species. Colored blocks represent conserved sequence regions (blocks). A variation of hue between two blocks of the same color indicates a small divergence in sequence. Protein 1 shows expected taxonomic variations. For protein 2, the orange block is missing in species F and G. (**B**) The BLAST results for proteins 1 and 2 using Species A as the query. In the Protein 1 BLAST, the species F, G and H, I, J are ranked together, since their respective sequences are similar. In the Protein 2 BLAST, Species H, I and J are ranked similarly to Protein 1, whereas species F and G are ranked further down due to the missing orange block.*

147 **Relative conservation at the protein family level**

148 BLAST homology searches are computed, for a complete query proteome (used as
149 reference species) and each BLAST result is then processed individually, with the first hit of each
150 species from both selected groups extracted, under the hypothesis that the sequences are
151 homologous to the query sequence. Alternatively, hits corresponding to orthologs predicted by
152 OrthoInspector 3.0 (Nevers et al. 2019) in species of both groups can be considered.

153 For each homolog or ortholog detected in the two groups, BLUR retrieves various
154 statistics of the BLAST hits (e.g., E-value, rank of hit in the BLAST, start and end position of the
155 pairwise alignment, etc.) and compares the average behavior of both groups for each protein
156 family (Figure 2). To avoid any bias caused by badly predicted sequences, hits where ranks are
157 detected as group outliers by Tukey's fences statistical method using a 1,5 interquartile range are
158 not taken into consideration for the calculations (Tukey 1977). Comparisons are only executed
159 for proteins where, for each of the groups, hits were found for at least 50% (33% for orthologous
160 sequences) of the species that are available in the BLUR database (see below).

161 For each protein family, the relative conservation of the two groups of species is evaluated
162 according to three parameters:

163 • The ratio between the mean (in log space) of the E-values of both groups
164 • The difference between the mean distances to the query of each group, where the distance
165 is defined as:

166 $$1 - \frac{(end_q - start_q) - mismatches - gap\_opens}{length_q}$$

167 With end_q, $start_q$, mismatches, gap_opens and $length_q$ being the position on the query
168 where the aligned hit ends, the position on the query where the aligned hit starts, the

number of mismatches in the alignment, the number of gaps opened, and the length of the query sequence, respectively.

- The percentage of hits of one group ranked higher in the BLAST than the other groups' best-ranked hit.

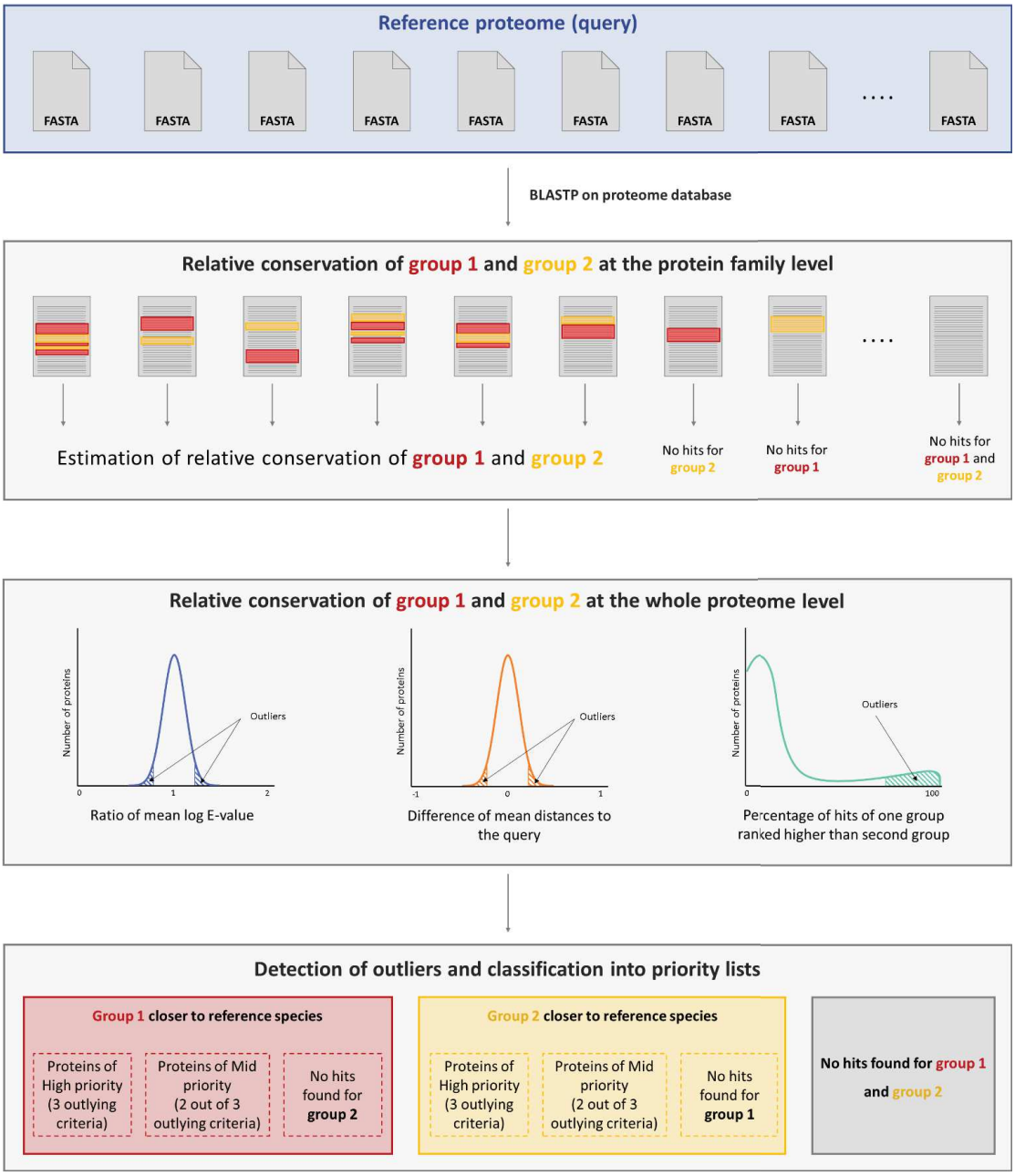

Fig 2: *Schematic representation of the BLUR protocol. A reference proteome is compared to a proteome database with BLASTP, and the results are stored in a database (not shown here). For each user-selected groups 1 and 2, BLUR establishes the relative conservation of both groups for each protein using three criteria: ratio of mean E-value in log space, difference of mean distance*

to the query, and ranking of one group compared to the other. The relative conservation is then analyzed on the whole proteome level, and outliers are detected using Tukey's fences method, and classified into priority lists.

Detection of outliers at the proteome level

The distributions of these parameters for the complete proteome are then analyzed using Tukey's fences method with a 1,5 interquartile range. Protein sequences with outlying values compared to the standard conservational behavior in the whole proteome are classified into two categories: "High priority", if all three criteria are detected as outliers, and "Mid priority", if only two out of the three criteria are detected as outliers. Proteins with no hits in one or both groups of species are classified in a third category. These three categories can then be analyzed in depth using various tools (see below).

BLUR databases

BLAST searches have been pre-calculated with default parameters and E-value threshold of 10^{-3} for 27 different query species (15 Eukaryotes, 8 Bacteria and 4 Archaea) in protein databases of the corresponding life domain (e.g., eukaryote queries on a database containing only eukaryotic proteins, etc.), using BLAST+ 2.5.0 (Camacho et al. 2009) with an E-value threshold of 1.0e-3 and a maximum of 5000 hits (Table 1). Reference species were selected to offer a broad coverage of the tree of life and allow users to study any specific groups of organisms. The Eukaryota, Bacteria and Archaea databases comprise 734, 3863 and 179 complete proteomes respectively, from the Uniprot reference proteomes (Bateman et al., 2017) and the RefSeq database (O'Leary et al., 2016). The proteomes included in the database were selected based on several criteria of quality such as low proportion of small proteins (< 100 amino anids) or proteins that do not start with a methionine. The BLUR relational database contains information (e.g., associated gene name, description, sequence length), for all the proteins available in the various proteomes used as queries for the BLAST searches (Table 1). It also stores conservation

features pertaining to the first homologous or orthologous hit of each species (e.g., percent identity to the query, length of the BLAST pairwise alignment, E-value, taxonomic id of the associated species, etc.) for all BLAST searches. Orthologous relations were predicted with OrthoInspector 3.0 and used to select relevant hits when populating the database with the results of the BLAST searches. The NCBI taxonomy (Federhen 2012) was used both in the BLAST searches, and in the database to enable an easy manipulation of the data and retrieval of target hits.

Table 1: *Query species available in BLUR for each of the three life domains, with the number of proteins in the proteome used. The last column indicates in which life group the query species can be used, as well as the number of species in the group.*

Domain	Query species (Taxonomy ID)	Number of proteins	Life Group (number of species)
Eukaryota	*Homo sapiens* (9606)	21044	Opisthokonta (557) / Metazoa (169)
	Mus musculus (10090)	22298	
	Xenopus tropicalis (8364)	24125	Metazoa (169)
	Drosophila melanogaster (7227)	13780	
	Caenorhabditis elegans (6239)	19990	
	Saccharomyces cerevisiae (559292)	6049	Fungi (384)
	Schizosaccharomyces pombe (284812)	5142	
	Cryptococcus neoformans (214684)	6601	
	Arabidopsis thaliana (3702)	27619	Viridiplantae (73)
	Chlamydomonas reinhardtii (3055)	14266	
	Cyanidioschyzon merolae (280699)	4995	Eukaryota (734)
	Plasmodium falciparum (36329)	5340	
	Dictyostelium discoideum (44689)	12731	
	Leishmania major (5664)	8031	
	Ectocarpus siliculosus (2880)	15903	
Bacteria	*Thermotoga maritima* (243274)	1852	Bacteria (3846)
	Bacillus subtilis (224308)	4260	
	Streptomyces coelicolor (100226)	8038	
	Treponema pallidum (243276)	1027	
	Chlamydia trachomatis (272561)	895	
	Escherichia coli (83333)	4347	
	Bacteroides thetaiotaomicron (226186)	4782	
	Aquifex aeolicus (224324)	1553	
Archaea	*Nanoarchaeum equitans* (228908)	536	Archaea (179)
	Pyrococcus abyssi (272844)	1788	
	Sulfolobus solfataricus (273057)	2938	
	Candidatus Thorarchaeota archaeon SMTZ1-45 (1706444)	3208	

Web implementation

To make BLUR user-friendly, a web interface was developed using the Symfony PHP web application framework (https://symfony.com/), with the Twig template engine (https://twig.symfony.com/). The website offers the opportunity to perform both global and individual analyses of the results, as well as the possibility to export the results in a CSV file. For the various lists of results, protein interaction networks can be generated using data from the STRING database when available (Szklarczyk et al. 2019), containing only direct interactions between proteins of the lists with a score greater than 0.7, and Gene Ontology (GO) enrichments can be computed using the Panther API (Mi et al. 2019). Individual analyses provide information about each protein detected by BLUR, with GO annotations, protein domain annotations provided by the InterPro webservice (Mitchell et al. 2019) and links to external resources such as UniProt and OrthoInspector. We also provide a multiple sequence alignment pre-computed using DBClustal (Thompson et al. 2000) containing up to 2000 homologous sequences and a visual representation of the BLAST result. The generated networks, GO enrichments and the pre-computed multiple sequence alignments can be exported from the website, as SIF, text and TFA files respectively.

Results

To address the need for a method capable of detecting both complete protein gain/loss and block-level divergences in a group of species, we developed a new approach based on BLAST homology search results designed to highlight atypical conservation patterns between orthologs or homologs. To facilitate both the use of BLUR and the analysis of the results, we developed a web interface that includes a variety of tools.

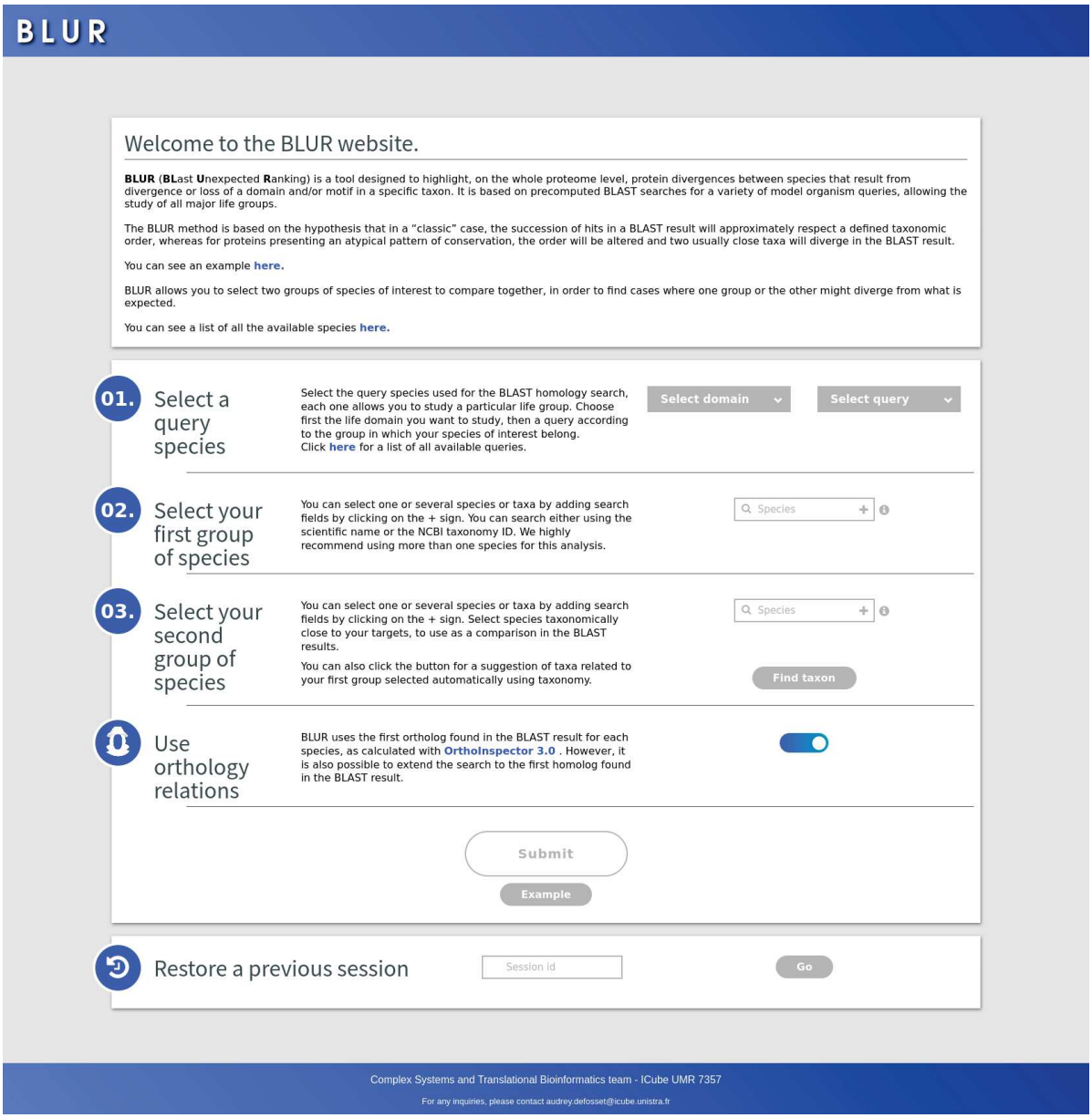

Fig. 3: *Home page of the BLUR website with the different steps necessary to run BLUR. Step 1 allows the user to select one of the three life domains (Eukaryota, Bacteria, Archaea), then the query species used for the BLAST search, as well as the life group to study. Step 2 allows the user to select the first group of interest, which can either be a taxon, several species or several taxa, but must be in the life group selected in Step 1. Finally, Step 3 consists in the selection of the second group to be compared, which can either be chosen by the user, or automatically using taxonomy. The last step is the selection of the type of relations to use for the BLAST computation: orthology (default) or homology. The user can also restore a previous session using a session ID provided on the result page.*

BLUR webserver

The home page of the website (http://lbgi.fr/blur/) shows the three steps necessary to run a BLUR analysis (Figure 3). The first step is the selection of the life group in which the species of

interest belong, and the query species (reference) to use for the BLAST searches using a drop down menu. In order to represent a large taxonomic diversity, 27 species spanning the three life domains are available as queries. The reference species should be chosen to be distant enough from both groups of interest so that in most cases they appear undistinguishable in a BLAST search. In other words, the two groups must share a more common ancestor than the one they share with the reference species. The second and third steps are the selection of the two groups of species to be compared. For each group, the user can choose several species, a single taxon, or several taxa, using a search bar containing an autocomplete feature. Only species belonging to the selected life group can be chosen. To help in the selection of groups, BLUR can automatically determine a set of possible second groups containing at least three species, according to the taxonomy of the user-defined first group. In this case, BLUR will propose taxa sharing a common ancestor with the first group and containing at least 3 species present in the database. If more than one taxa is selected, BLUR first retrieves their common ancestor, and looks for (1) other children taxa of the common ancestor containing at least 3 species and (2) sister taxa to the common ancestor with at least 3 species present in the database. Lastly, the user has the possibility of choosing whether to use only orthologs computed with OrthoInspector, or extend the search to homologs found in the BLAST search.

The results obtained from the BLUR software are presented on a Results page in three sections. The first section contains a list of proteins where the second group is closer to the query species than the first group. The second section contains a list of proteins where the first group is closer to the query species than the second group. The third section contains a list of proteins where no hits were found in the BLAST for either groups. The first two sections are divided into three sub-categories: absence of homolog/ortholog, High priority and Mid priority proteins. The

two latter correspond to differentially conserved proteins fulfilling respectively three or two BLUR criteria of differential conservation.

For each of the three blocks, and each sub-category within these blocks, interaction networks and GO term enrichments can be generated. Selecting an individual protein in any of the lists will open a protein page containing diverse information. Firstly, a header provides general data on the protein such as the associated gene name, the protein description, the length of the protein, links to external resources, GO terms and InterPro domains associated with the protein. Secondly, the user can access BLUR specific data: a representation of the BLAST output with the hits of both groups highlighted for easier analysis and a multiple sequence alignment. This alignment contains a subset of sequences of both groups of species, the query species as well as sequences of a few organisms related to the query. It is also possible to display a more complete multiple sequence alignment, containing up to 2000 homologous sequences, and in this case, species of interest will be highlighted.

The BLUR approach has been tested on different groups of species, demonstrating the advantages of combining sub-protein level and protein level information, in order to highlight lineage specialization and obtain a comprehensive view of genotype/phenotype correlations. Two examples of studies performed on the BLUR website are presented below: prediction of cilia-related proteins in Eukaryotes, and prediction of proteins involved in sulfur oxidation in Bacteria.

Use case: cilia-related proteins in Fungi

Cilia are small microtubule-based organelles present in the Last Eukaryotic Common Ancestor that exhibit an unusual evolutionary history with various independent losses in the eukaryotic lineage, which makes them a good candidate for comparative genomics studies. Most

Fungi are devoid of cilia, with a few known exceptions, namely Chytridiomycota, Blastocladiales, and *Rozella* (Adl et al. 2012). We used our method to identify cilia-related proteins with the assumption that in ciliated Fungi, proteins linked to cilia should be more similar to their metazoan homologs than to their homologs found in non-ciliated fungal species.

We chose Opisthokonta as the life group of interest, with *Homo sapiens* as the query proteome. We used Chytridiomycota, Blastocladiales and *Rozella* taxa as the first group (with a total of 6 species), and Dikarya (350 non-ciliated species) as the second group, using ortholog sequences.

For the category corresponding to our hypothesis, where ciliated Fungi proteins are closer to Human than Dikarya, 1081 proteins were absent in Dikarya, 18 were classified as High priority, and 81 as Mid priority. A manual analysis of the multiple sequence alignments showed the presence of divergent regions in most proteins, with 12 false positives found in the Mid priority list, due to either an insufficient number of sequences, or the presence of low quality sequences. As an example, the multiple alignment of RFX1 is provided as Supplementary Data, showing the presence of only 3 badly predicted sequences of ciliated Fungi. A GO enrichment analysis of the 1180 proteins showed that they were significantly enriched in terms related to cilia, such as 'cilium' (P-value: $2.30*10^{-75}$) or 'intraciliary transport' (P-value: $2.05*10^{-22}$). To further assess the quality of our results, we compared the 1180 proteins to a negative set of 971 proteins from pathways unlikely to be related to cilia constructed in a previous study (Nevers et al. 2017). Only 22 proteins of this negative set were included in the 1180 proteins, with 2 in the High priority list, and 2 in the Mid priority.

Of the 1180 proteins detected, 526 presented a high confidence interaction with at least one other. The interaction networks generated showed one main network of 400 proteins

consisting of several highly linked clusters, including ones enriched in intraciliary transport, centriole elongation and basal body docking, and cell proliferation regulation (Figure 4). Among the 400 proteins present in the main network, 362 are absent in Dikarya, including 76 related to cilia previously detected using a phylogenetic profiling method (Nevers et al. 2017). The other 38 proteins present in the network come from both the High priority list (orange nodes in Figure 4) and the Mid priority list (green nodes in Figure 4). Thus, these proteins are present in Dikarya, but exhibit a probable differential conservation. 10 of them are already annotated as related to cilia, while the other 28 represent potential new cilia-related candidates. Many clusters contain proteins detected by differential conservation both at the protein and sub-protein levels and illustrate the relevance of our approach.

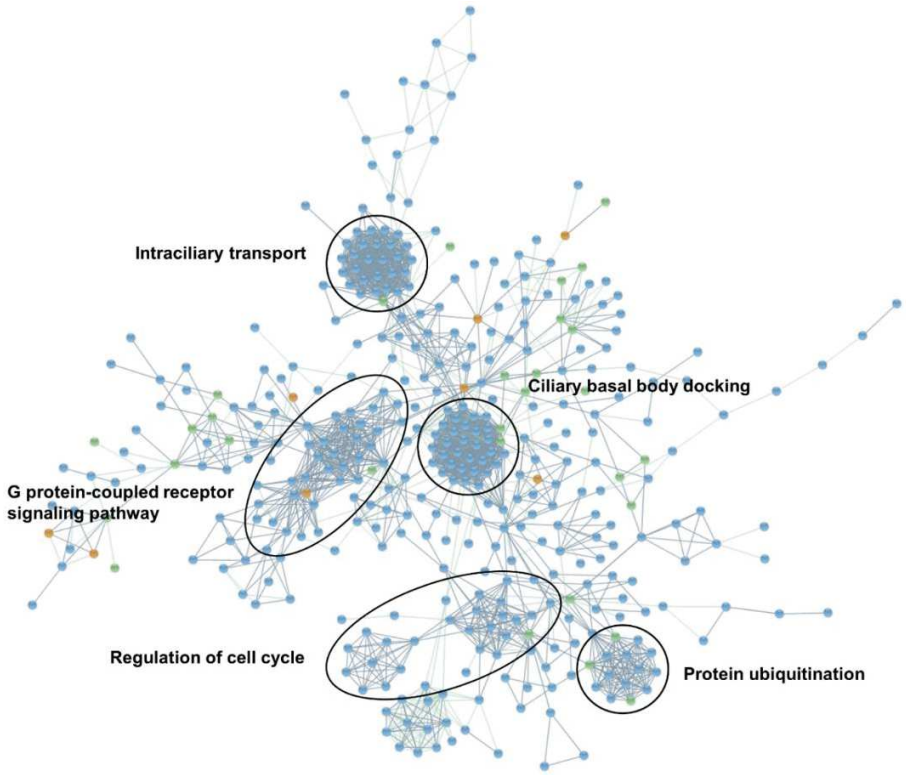

Fig. 4*: Main interaction network of proteins absent in Dikarya (blue nodes), and proteins predicted to have differential conservation with High priority (orange nodes) or Mid priority (green nodes). The network contains highly linked clusters of proteins that are both absent and divergent in Dikarya, and that are enriched in GO terms corresponding to ciliary components, thus validating the proposed method*

Among the 99 proteins in the High and Mid priority lists (including the 38 proteins found in the interaction network), 17 had annotations linked to cilia, centrosome, centriole or microtubule, of which at least 14 presented a clear differential conservation confirmed by visual inspection of the multiple alignment. A particularly striking example is ARMC4, a ciliary protein involved in left/right symmetry and axonemal outer dynein arm assembly, with homologs found in most eukaryotic clades, including Metazoa and Fungi. A multiple sequence alignment of the ARMC4 family showed a clear distinction between the sequences of ciliated versus non-ciliated Fungi, with a higher similarity between vertebrate sequences and ciliated Fungi sequences (Figure 5). In particular, Vertebrates and ciliated Fungi proteins present a long N-terminal region that could constitute a yet undiscovered functional domain, while non-ciliated Fungi proteins have a much shorter sequence.

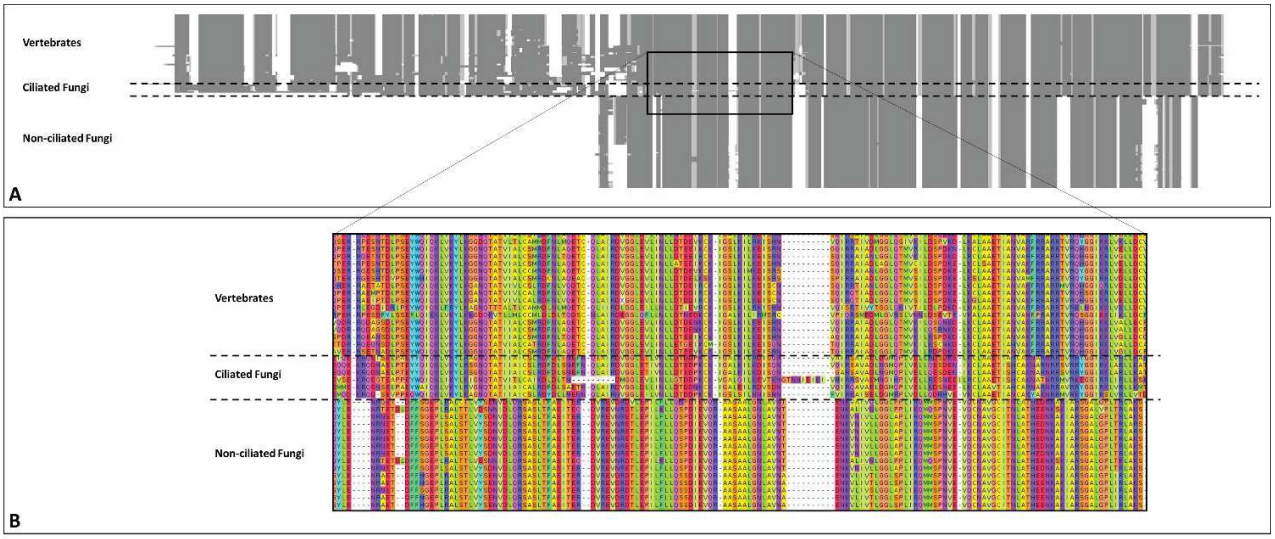

***Fig. 5**: Multiple sequence alignment of ARMC4. (**A**) Overview of the multiple sequence alignment of ARMC4. Vertebrates (ciliated species) and ciliated Fungi sequences are similar with a long N-terminal domain that is absent in non-ciliated Fungi. (**B**) Zoom on a portion of the alignment where differential conservation can be observed. Ciliated fungi are very similar to Vertebrates while other, non-ciliated Fungi are more divergent.*

Use case: Sulfur oxidation in Bacteria

In certain ecosystems, hydrogen sulfide is more abundant than oxygen, allowing certain microorganisms to use sulfur as a means to produce energy. Sulfur oxidation is performed almost

exclusively by Archaea and Bacteria, with a few eukaryotic exceptions. Here, we used BLUR to predict proteins related to sulfur oxidation in Bacteria, using the known sulfur-oxidizing Bacteria *Aquifex aeolicus* as a query proteome. We selected two close groups of Gammaproteobacteria for comparison, with one group able to oxidize sulfur (Chromatiales) and the other not (Enterobacterales). Our hypothesis is that most proteins from Chromatiales are highly similar to their orthologs in Enterobacterales and more divergent compared to *Aquifex* orthologs. In contrast, proteins involved in sulfur oxidation should be highly similar between Chromatiales and *Aquifex*, and very different from the orthologs (if any) found in Enterobacterales.

Using BLUR, we detected 223 proteins in the category where Chromatiales are closer to *Aquifex* than Enterobacterales, with 186 absent in Enterobacterales, 16 classified as High priority, and 21 as Mid priority. As for the previous example, a manual analysis of the multiple sequence alignments showed divergence in most cases, with 6 false positives in the Mid priority list. A GO enrichment analysis of these 223 proteins was not useful due to the lack of GO annotations for the majority of *Aquifex* proteins. However, the interaction networks showed the presence of several clusters (Figure 6). To investigate further the functions associated with these clusters, we used ortholog annotations provided by OrthoInspector (Nevers et al. 2019). We identified the Sox protein cluster (Figure 6), essential for sulfur oxidation that includes proteins absent from the Enterobacterales group (SoxAX, SoxF, SoxW, SoxX, SoxY, SoxZ) and also the High priority SoxB protein, well conserved in Chromatiales but highly divergent in Enterobacterales. The dimethyl sulfoxide (DMSO) reductase associated with the Sox cluster was also detected with DmsA, DmsB1 and DmsC protein subunits classified as High priority, Mid priority and absent in Enterobacteriales respectively.

We also identified a large iron-sulfur protein cluster (Figure 6), containing the proteins from the *hdr* gene cluster (dsrE2A, dsrE3B, dsrE3C, hdrA, hdrB1, hdrB2, hdrC1, hdrC2), known to be involved in sulfur oxidation (Boughanemi et al. 2016; Quatrini et al. 2009), which were found to be absent in Enterobacterales. Other proteins with no known interactions were found to have a clear distinction between Chromatiales and Enterobacterales sequences, such as Peroxiredoxin, which was verified using a multiple sequence alignment.

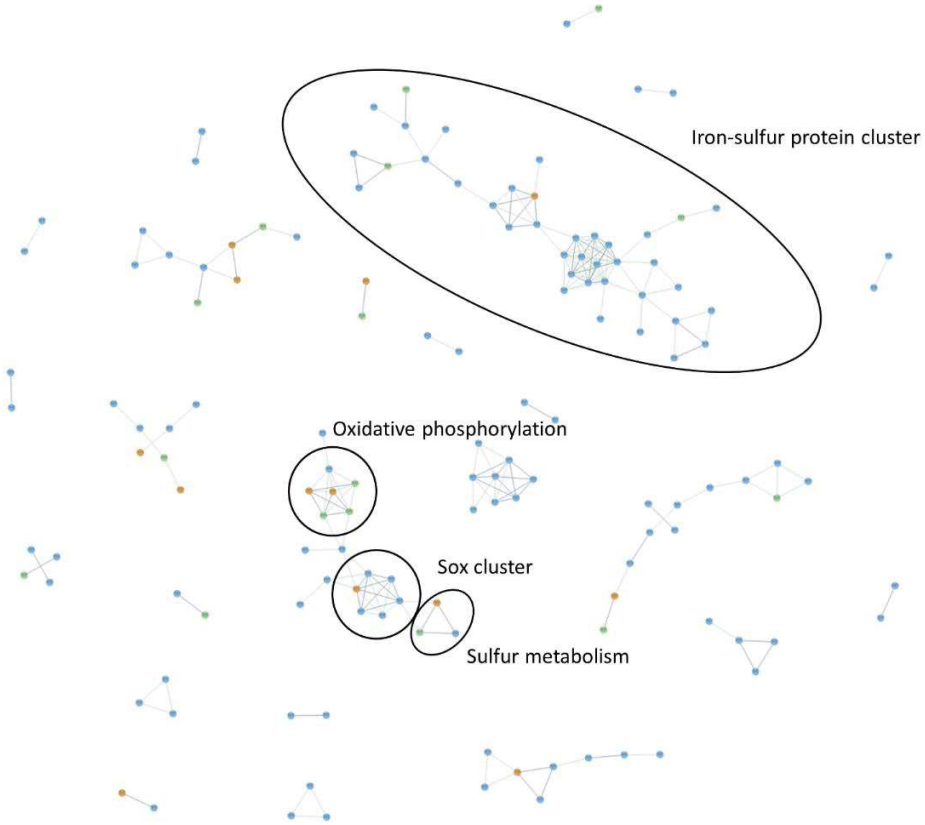

***Fig. 6**: Interaction networks of proteins absent in Enterobacterales (blue nodes), of High priority (orange nodes) and of mid priority (green nodes). Several clusters contained over ten proteins with high confidence links between them, including a cluster containing the main Sox proteins, and a cluster corresponding to the iron-sulfur proteins found in the hdr cluster.*

Examples of proteome comparisons without prior knowledge

We have previously shown that BLUR can be used to study phenotypes of interest by comparing species that present a specific character and species devoid of that character. More

generally, BLUR can be used to compare two groups of species without focusing on any specific process. Table 2 shows the results obtained when performing various searches on the BLUR webserver, using different query species in different life domains.

Table 2: *Examples of application of BLUR using various query species and groups of interest*

Query species	Comparison	Protein lists	GO enrichment	Network	Network enrichment
Homo sapiens	Basidiomycota over Ascomycota	469 absent in Ascomycota, 32 High priority, 112 Mid priority	RNA processing (P-value: 2.12E-10) Protein modification process (P-value: 3.17E-9) RNA splicing (P-value: 3.04E-8)	Main network of 208 proteins: 140 absent, 14 High priority, 54 Mid priority	Several clusters: mRNA splicing ; ribosome biogenesis; regulation of signal transduction
Mus musculus	Lophotrochozoa over Ecdysozoa	775 Absent in Ecdysosoa, 23 High priority, 105 Mid priority	Nervous system process (P-value: 1.34E-12) Sterol metabolic process (P-value: 5.62E-7) Cilium assembly (P-value: 1.37E-6)	224 Proteins with a least one interaction: 177 Absent, 10 High, 37 Mid priority	Several small networks: steroid biosynthetic process; regulation of apoptotic process; cilium assembly; cell cycle
Chlamydomonas reinhardtii	Liliopsida over Eudicotyledons	107 Absent in Eudicotyledons, 18 High priority, 81 Mid priority	Photosynthesis (P-value: 2.25E-10) Oxidation-reduction process (P-value: 1.41E-9)	44 Proteins with at least one interaction: 15 absent, 7 High priority, 22 Mid priority	Photosynthesis
Escherichia coli	Betaproteobacteria over Alphaproteobacteria	252 Absent in Alphaproteobacteria, 5 High priority, 28 Mid priority	Pilus organization (P-value: 5.31E-16) Submerged biofilm formation (P-value: 2.69E-6)	Main network of 91 proteins: 77 absent, 2 High priority, 12 Mid priority	Several clusters: cell motility; pilus organization; asexual reproduction
Bacillus subtilis	Selenomonadales over Veillonellales	635 Absent in Veillonellales, 23 High priority, 34 Mid priority	Locomotion (P-value: 7.65E-15) Chemotaxis (P-value: 1.63E-7)	Main network of 401 proteins: 364 absent, 18 High priority, 19 Mid priority	Several clusters: spore germination; locomotion; antibiotic metabolic process

In all examples, BLUR detected proteins that were absent, and proteins that showed divergences of both high and mid priority, with significant functional enrichments in all lists. Most of the networks generated showed highly linked clusters of proteins that are both absent and divergent, with GO enrichment in specific biological functions. These functional links between families showing loss/gain of a complete gene and differential conservation at the sub-gene level highlights the added value of our approach compared to an analysis based on the sole presence/absence of genes.

Discussion

BLUR represents an online resource capable of rapidly detecting differential conservation from BLAST search results at the whole proteome level, in any of the 4776 species available in the pre-calculated database. Our original approach addresses the problems generated by variable evolutionary rates between taxa, by using a reference species to perform relative comparisons and establishing an average conservation behavior over a whole proteome. These comparisons can be performed among orthologs or homologs; while using orthologs allow for a more restricted search and limit the false positives that could be attributed to the detection of close paralogs, it can also create false negatives due to the problems of orthologs inference caused by highly diverging sequences, which could be detected by using homologs.

We provide an accessible and easy to navigate website, with a substantial amount of complementary information that allows for more in-depth analysis. We have shown that our method is not limited to any specific biological process or life domain, by identifying cilia-related proteins in Eukaryotes, as well as proteins related to sulfur oxidation in Bacteria. Both examples demonstrate the usefulness of an approach combining complete protein loss/gain and sub-protein

406 variation by presenting results containing clusters of strongly interacting proteins that were both
407 completely lost and only partially divergent in some regions.

408 It is difficult to estimate the sensitivity and specificity of our approach as there are
409 currently no suitable benchmarks for differential conservation detection. However, manual
410 inspection of multiple alignments of proteins detected by our approach showed that in both use-
411 cases, most of the proteins from the High priority list exhibited a more or less pronounced
412 differential conservation, with false positives in the Mid priority lists. This manual analysis
413 showed that the precision and the quality of the results are mostly dependent on the number of
414 species in each group, and more importantly on the quality of the sequences available. In some
415 cases, one group did not contain enough reliable sequences to properly assess the conservation
416 between the two groups.

417 The quality of the BLUR results are clearly dependent on the parameters chosen (number
418 of species in each group, distance between the query and the groups, complexity of the
419 phenotypic differences between groups), and are entirely correlated with the quality of the
420 sequences available in the database, leading to a small proportion of false positives. However, we
421 have shown that our method is effective in the detection of proteins related to a given phenotype
422 and to generate relevant new candidates that can be analyzed easily and rapidly with the various
423 tools available on the website. Future developments will include the addition of new reference
424 genomes to extend the comparison possibilities, as well as an extension of the sequence databases
425 with more species to analyze and compare.

426 # Funding

This work was supported by the IdEx Unistra in the framework of the “Investments for the future” program of the French government and Institute funds from the Centre National de la Recherche Scientifique and the Université de Strasbourg.

Acknowledgments

We thank the Bio-statistics, Informatics and Complex System platform (BICS) and BISTRO bioinformatics platforms for informatics support and the European Grid Infrastructure for cloud computing facilities.

References

Adl SM et al. 2012. The Revised Classification of Eukaryotes. Journal of Eukaryotic Microbiology. 59:429–514. doi: 10.1111/j.1550-7408.2012.00644.x.

Anderson DP et al. 2016. Evolution of an ancient protein function involved in organized multicellularity in animals. eLife. 5. doi: 10.7554/eLife.10147.

Bateman A et al. 2017. UniProt: the universal protein knowledgebase. Nucleic Acids Res. 45:D158–D169. doi: 10.1093/nar/gkw1099.

Boughanemi S et al. 2016. Microbial oxidative sulfur metabolism: biochemical evidence of the membrane-bound heterodisulfide reductase-like complex of the bacterium Aquifex aeolicus. FEMS Microbiol Lett. 363. doi: 10.1093/femsle/fnw156.

Camacho C et al. 2009. BLAST+: architecture and applications. BMC Bioinformatics. 10:421. doi: 10.1186/1471-2105-10-421.

Cheng Y, Perocchi F. 2015. Prediction of Mitochondrial Protein Function by Comparative Physiology and Phylogenetic Profiling. In: Mitochondrial Medicine. Weissig, V & Edeas, M, editors. Vol. 1264 Springer New York: New York, NY pp. 321–329. doi: 10.1007/978-1-4939-2257-4_28.

Cromar GL et al. 2016. PhyloPro2.0: a database for the dynamic exploration of phylogenetically conserved proteins and their domain architectures across the Eukarya. Database. 2016:baw013. doi: 10.1093/database/baw013.

Cunningham FX, Lafond TP, Gantt E. 2000. Evidence of a Role for LytB in the Nonmevalonate Pathway of Isoprenoid Biosynthesis. Journal of Bacteriology. 182:5841–5848. doi: 10.1128/JB.182.20.5841-5848.2000.

Dey G, Jaimovich A, Collins SR, Seki A, Meyer T. 2015. Systematic discovery of human gene function and principles of modular organization through phylogenetic profiling. Cell Rep. 10:993–1006. doi: 10.1016/j.celrep.2015.01.025.

Dohmen E, Klasberg S, Bornberg-Bauer E, Perrey S, Kemena C. 2020. The modular nature of protein evolution: domain rearrangement rates across eukaryotic life. BMC Evolutionary Biology. 20. doi: 10.1186/s12862-020-1591-0.

El-Gebali S et al. 2019. The Pfam protein families database in 2019. Nucleic Acids Res. 47:D427–D432. doi: 10.1093/nar/gky995.

Federhen S. 2012. The NCBI Taxonomy database. Nucleic Acids Res. 40:D136–D143. doi: 10.1093/nar/gkr1178.

Forslund K, Pekkari I, Sonnhammer EL. 2011. Domain architecture conservation in orthologs. BMC Bioinformatics. 12:326. doi: 10.1186/1471-2105-12-326.

Haider C, Kavic M, Sonnhammer ELL. 2016. TreeDom: a graphical web tool for analysing domain architecture evolution. Bioinformatics. 32:2384–2385. doi: 10.1093/bioinformatics/btw140.

Jeffery CJ. 1999. Moonlighting proteins. Trends in Biochemical Sciences. 24:8–11. doi: 10.1016/S0968-0004(98)01335-8.

Jim K. 2003. A Cross-Genomic Approach for Systematic Mapping of Phenotypic Traits to Genes. Genome Research. 14:109–115. doi: 10.1101/gr.1586704.

Kim Y, Subramaniam S. 2005. Locally defined protein phylogenetic profiles reveal previously missed protein interactions and functional relationships. Proteins: Structure, Function, and Bioinformatics. 62:1115–1124. doi: 10.1002/prot.20830.

Kress A, Lecompte O, Poch O, Thompson JD. 2018. PROBE: analysis and visualization of protein block-level evolution. Bioinformatics. doi: 10.1093/bioinformatics/bty367.

Lees JG, Dawson NL, Sillitoe I, Orengo CA. 2016. Functional innovation from changes in protein domains and their combinations. Current Opinion in Structural Biology. 38:44–52. doi: 10.1016/j.sbi.2016.05.016.

Letunic I, Bork P. 2018. 20 years of the SMART protein domain annotation resource. Nucleic Acids Res. 46:D493–D496. doi: 10.1093/nar/gkx922.

Li JB et al. 2004. Comparative Genomics Identifies a Flagellar and Basal Body Proteome that Includes the BBS5 Human Disease Gene. Cell. 117:541–552. doi: 10.1016/S0092-8674(04)00450-7.

Löhr U, Yussa M, Pick L. 2001. Drosophila fushi tarazu: a gene on the border of homeotic function. Current Biology. 11:1403–1412. doi: 10.1016/S0960-9822(01)00443-2.

Mani M et al. 2015. MoonProt: a database for proteins that are known to moonlight. Nucleic Acids Research. 43:D277–D282. doi: 10.1093/nar/gku954.

Mi H, Muruganujan A, Ebert D, Huang X, Thomas PD. 2019. PANTHER version 14: more genomes, a new PANTHER GO-slim and improvements in enrichment analysis tools. Nucleic Acids Research. 47:D419–D426. doi: 10.1093/nar/gky1038.

Mitchell AL et al. 2019. InterPro in 2019: improving coverage, classification and access to protein sequence annotations. Nucleic Acids Research. 47:D351–D360. doi: 10.1093/nar/gky1100.

Moore AD, Bornberg-Bauer E. 2012. The Dynamics and Evolutionary Potential of Domain Loss and Emergence. Mol Biol Evol. 29:787–796. doi: 10.1093/molbev/msr250.

Moore AD, Grath S, Schüler A, Huylmans AK, Bornberg-Bauer E. 2013. Quantification and functional analysis of modular protein evolution in a dense phylogenetic tree. Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics. 1834:898–907. doi: 10.1016/j.bbapap.2013.01.007.

Moore AD, Held A, Terrapon N, Weiner J, Bornberg-Bauer E. 2014. DoMosaics: software for domain arrangement visualization and domain-centric analysis of proteins. Bioinformatics. 30:282–283. doi: 10.1093/bioinformatics/btt640.

Nevers Y et al. 2017. Insights into Ciliary Genes and Evolution from Multi-Level Phylogenetic Profiling. Mol Biol Evol. 34:2016–2034. doi: 10.1093/molbev/msx146.

Nevers Y et al. 2019. OrthoInspector 3.0: open portal for comparative genomics. Nucleic Acids Res. 47:D411–D418. doi: 10.1093/nar/gky1068.

O'Leary NA et al. 2016. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res. 44:D733–D745. doi: 10.1093/nar/gkv1189.

Pagel P, Wong P, Frishman D. 2004. A Domain Interaction Map Based on Phylogenetic Profiling. Journal of Molecular Biology. 344:1331–1346. doi: 10.1016/j.jmb.2004.10.019.

Pellegrini M, Marcotte EM, Thompson MJ, Eisenberg D, Yeates TO. 1999. Assigning protein functions by comparative genome analysis: Protein phylogenetic profiles. Proc Natl Acad Sci U S A. 96:4285–4288.

Persson E, Kaduk M, Forslund SK, Sonnhammer ELL. 2019. Domainoid: domain-oriented orthology inference. BMC Bioinformatics. 20:523. doi: 10.1186/s12859-019-3137-2.

Quatrini R et al. 2009. Extending the models for iron and sulfur oxidation in the extreme Acidophile Acidithiobacillus ferrooxidans. BMC Genomics. 10:394. doi: 10.1186/1471-2164-10-394.

Ronshaugen M, McGinnis N, McGinnis W. 2002. Hox protein mutation and macroevolution of the insect body plan. Nature. 415:914–917. doi: 10.1038/nature716.

Shiga Y, Yasumoto R, Yamagata H, Hayashi S. 2002. Evolving role of Antennapedia protein in arthropod limb patterning. Development. 129:3555–3561.

Szklarczyk D et al. 2019. STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. Nucleic Acids Research. 47:D607–D613. doi: 10.1093/nar/gky1131.

Thompson JD, Plewniak F, Thierry J-C, Poch O. 2000. DbClustal: rapid and reliable global multiple alignments of protein sequences detected by database searches. Nucleic Acids Res. 28:2919–2926.

Tukey JW. 1977. *Exploratory Data Analysis*. Addison-Wesley Publishing Company Reading, Mass.

Vera-Parra N, Gutiérrez-Ramirez M, López-Sarmiento D. 2016. Automatic construction and graph-making of functional domain architectures. Advances in Natural and Applied Sciences. 10:99–105.

Zmasek CM, Godzik A. 2011. Strong functional patterns in the evolution of eukaryotic genomes revealed by the reconstruction of ancestral protein domain repertoires. Genome Biol. 12:R4. doi: 10.1186/gb-2011-12-1-r4.

3. Discussion

3.1. Approche novatrice de génomique comparative

3.1.1. Un concept innovant

Nous avons développé BLUR dans le but d'avoir à disposition une ressource de génomique comparative novatrice combinant plusieurs niveaux de profilage phylogénétique, pouvant identifier des divergences à la fois au niveau des protéines complètes mais également au niveau d'une partie de la séquence, et ce pour permettre l'étude approfondie des relations génotype/phénotype. Alors que les autres approches connues se concentrent majoritairement sur les domaines protéiques annotés, BLUR est capable de détecter des divergences à des niveaux de granularité allant de la région élargie au motif protéique, voire à l'acide aminé dans le cas de protéines fortement conservées (*Figure 5-1*).

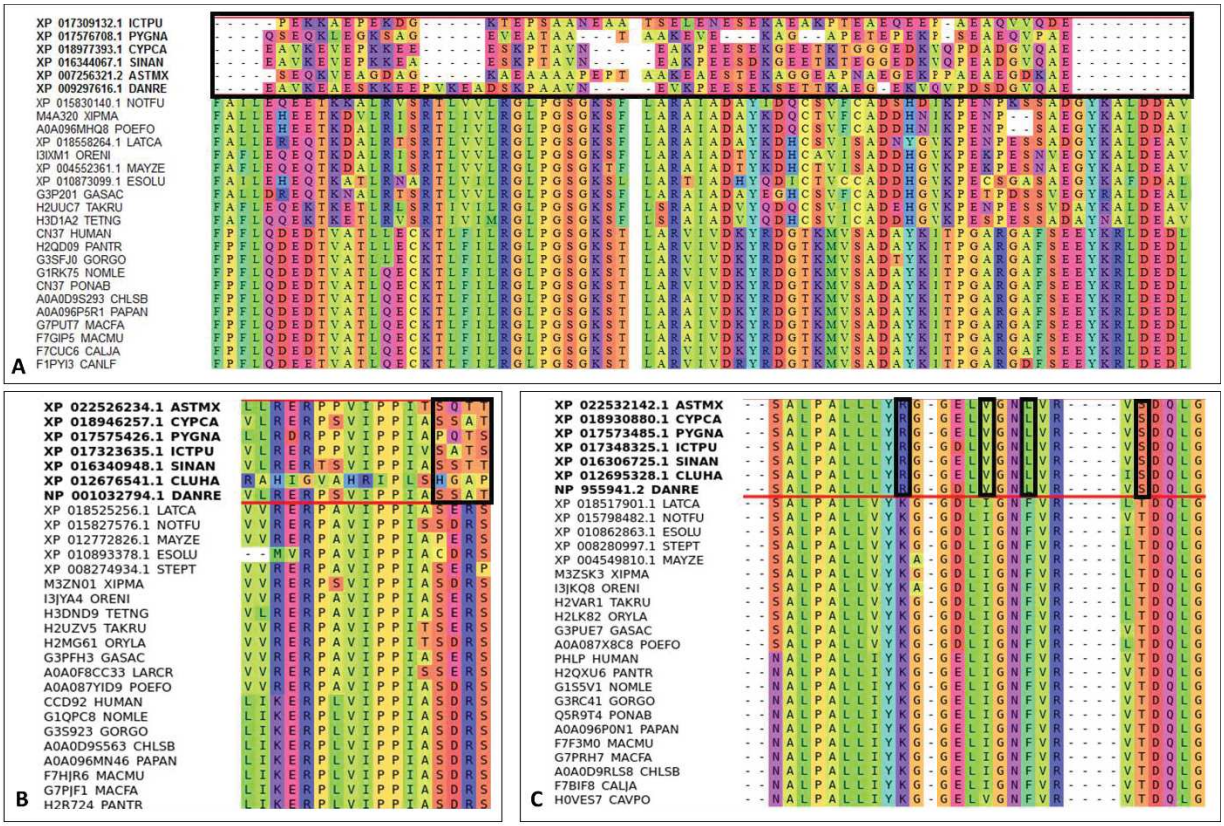

Figure 5-1: Exemples de conservation différentielle détectée par BLUR. La comparaison est faite entre deux groupes de séquences que l'on voit ici séparées par un trait rouge. (A) Alignement multiple de CNP. (B) Alignement multiple de CCDC92. (C) Alignement multiple de PDCL.

Un autre point particulier de la méthode que nous avons mise au point est l'utilisation de l'espèce de référence, de cette manière, il est possible de s'affranchir des variations de vitesses d'évolution entre les différentes familles de protéines. En effet, pour chaque famille protéique, BLUR compare la conservation relative des deux groupes d'espèces par rapport à la référence. C'est cette conservation relative qui est ensuite utilisée par BLUR pour établir le comportement moyen des deux groupes d'espèces sur l'ensemble du protéome et ainsi détecter les divergences atypiques, quelles que soient les vitesses d'évolution des familles considérées.

3.1.2. Une large couverture du Vivant

Dans un souci d'exhaustivité, d'interopérabilité et de maintien des structures, la base de données associée à la ressource BLUR est en partie basée sur celles de la ressource OrthoInspector 3.0, auxquelles ont été ajoutés les protéomes de certaines espèces de poissons issus de la banque publique RefSeq en vue d'une application à la multiciliation. Elle couvre ainsi un total de 4776 espèces appartenant aux taxons majeurs du Vivant, comprenant 734 Eucaryotes, 3863 Bactéries, et 179 Archées. Nous avons cherché de cette manière à permettre de poser de nombreuses questions biologiques par la comparaison d'espèces variées comme le montrent les exemples présentés dans la publication.

3.1.3. Un site web intégratif

Le site web associé à BLUR a été conçu de manière à faciliter d'une part l'utilisation et l'exécution d'une recherche, mais également les analyses en aval par la mise en place et l'intégration de nombreux outils de génomique comparative. La ressource BLUR inclut ainsi différents niveaux d'information permettant d'avoir une vue d'ensemble : un niveau évolutif, de par les relations d'orthologie ou d'homologie et l'analyse de conservation de séquences, un niveau structural grâce aux informations sur les domaines tirées de la base de données InterPro et un niveau fonctionnel, d'une part *via* les annotations et les enrichissements GO, et d'autre part grâce aux réseaux d'interactions générés à partir des données issues de la base STRING.

3.2. Les protéomes, une question de qualité et de quantité

Les performances de BLUR sont en majeure partie liées aux protéomes employés pour construire la base de données, à la fois en termes de qualité des séquences, mais également en termes du nombre et de la diversité des espèces choisies.

3.2.1. Qualité des protéomes

Les prédictions de conservation atypique réalisées par BLUR sont basées sur les séquences protéiques et dépendent donc intégralement de la qualité de celles-ci. Ainsi, un protéome incomplet ou une erreur de prédiction de séquence, aussi minime soit-elle, peut avoir un impact négatif sur les résultats calculés par BLUR, malgré les protocoles mis en place pour réduire au mieux ce type de biais. Malheureusement, de nombreuses séquences semblent mal prédites et sont à l'origine de faux résultats positifs lors de l'exécution d'une instance de BLUR (*Figure 5-2*). Plusieurs études sur les génomes eucaryotes ont montré que la structure exon/intron complète n'était correctement prédite que pour 50 à 60% des gènes (Brent, 2008; Guigó et al., 2006; Harrow et al., 2009). Il est donc important dans le cadre de cette ressource de ne considérer que les protéomes de qualité suffisante, ce qui peut parfois se révéler complexe. En effet, les protéines de certaines espèces, malgré leur statut de représentant au sein de leur taxon, présentent systématiquement des erreurs de prédiction et rendent difficiles les analyses automatiques. Pour tenter de remédier à cette problématique récurrente, UniProt propose depuis la fin de l'année 2019 une caractérisation de ses protéomes de référence par le score BUSCO (Waterhouse et al., 2018) ainsi qu'un indicateur prévisionnel de complétude basé sur l'algorithme CPD (*Complete Proteome Detector*), développé par UniProt. De cette manière le choix des protéome sera facilité mais l'ensemble des problèmes ne seront pas

résolus. En effet ces deux indicateurs se basent sur la complétude des protéomes en termes de nombre de protéines, en comparant notamment aux protéomes d'organismes proches pour évaluer si le protéome semble complet ou non, mais aucun d'eux ne prend en compte la qualité ou la complétude des séquences protéiques en elles-mêmes.

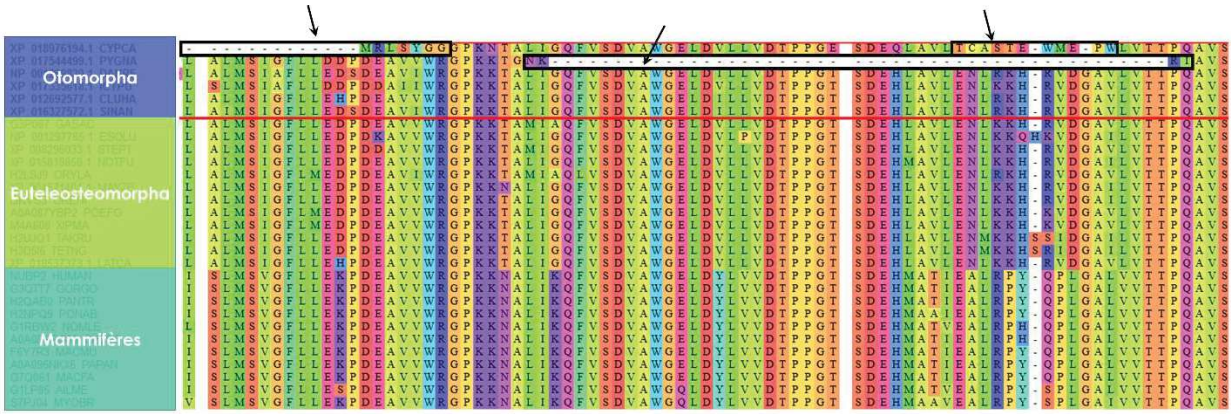

Figure 5-2: Section de l'alignement multiple de la protéine NUBP2. La comparaison a ici été réalisée entre les Otomorpha et les Euteleosteomorpha. Les flèches pointent des régions manquantes suite à des erreurs de prédictions des gènes correspondant. Ces erreurs sont à l'origine d'un faux résultat positif, la conservation de la famille étant détectée comme atypique par BLUR.

3.2.2. Quantité et diversité des protéomes

La précision des questions pouvant être posées par BLUR dépend entièrement du nombre de protéomes disponibles dans un taxon d'intérêt. Malgré la diversité des organismes présents dans la base de données BLUR, un grand nombre de taxons ne sont représentés que par une ou deux espèces, voire aucune, rendant impossibles les analyses comparatives sur ces groupes. Par exemple, l'ordre des Squamates (reptiles à écailles comprenant les serpents et les lézards) n'est représenté que par 2 espèces (3 protéomes de références sont disponibles dans UniProt), alors que plus de 10 000 espèces ont été à ce jour identifiées (Uetz et al., 2020). Ces problèmes sont en grande partie inhérents aux intérêts de la communauté scientifiques, plus prononcés pour certains groupes taxonomiques que pour d'autres. Pour pallier ces inégalités, des projets voient le jour, comme par exemple le *Vertebrate Genome Project*, débuté en février 2017, sur l'initiative du *Genome 10K*, dont le but est de séquencer, à terme, les génomes d'environ 70 000 espèces de Vertébrés et d'en faire des assemblages de référence [https://vertebrategenomesproject.org/].

3.3. Futurs développements

Les améliorations envisageables pour la ressource BLUR incluent l'optimisation des modules analytiques et visuels actuellement en place sur le site ainsi que l'ajout de nouveaux outils de génomique pour d'une part faciliter davantage les analyses en aval, et d'autre part enrichir les informations d'ores et déjà disponibles. Il serait notamment intéressant d'implémenter un autre outil d'enrichissement GO, le *webservice* mis à disposition par PANTHER étant limité à un nombre réduit d'espèces. Au vu des résultats de l'analyse évolutive réalisée sur la multiciliation (voir Chapitre 6), il serait également utile d'avoir des informations relatives à la position chromosomique des gènes codant pour les protéines obtenues en résultat, pour mettre en évidence une éventuelle co-localisation de plusieurs cibles.

Nous avons vu au point précédent dans quelle mesure la diversité et la qualité des protéomes étaient des critères importants pour le fonctionnement optimal de BLUR. Dans cette optique, un des développements futurs et indispensable de la ressource sera d'une part la mise à jour des protéomes actuellement présents dans la base, mais également une extension de cette dernière avec de nouveaux protéomes jugés de bonne qualité sur la base d'indicateurs objectifs. L'augmentation du nombre et de la diversité des protéomes et des espèces de référence permettra d'étendre le panel de questions biologiques pouvant être étudiées avec BLUR. Pour cela, il serait intéressant de mettre en place un protocole de mise à jour automatique pour récupérer, de manière périodique et régulière, les nouveaux ajouts et les mises à jours éventuelles de la banque UniProt.

Il est bien évident que malgré une mise à jour régulière et une extension de la base, il ne sera jamais possible de couvrir l'intégralité des espèces. Dans cette optique, le dernier développement en prévision pour BLUR est la mise à disposition d'une version indépendante du programme pour permettre aux utilisateurs d'appliquer le protocole de détection de conservation différentielle à des jeux de données qui leur sont propres ou qui diffèrent de la version standard proposée en ligne. De cette manière, il sera possible non seulement de poser des questions biologiques précises, mais également de compléter les résultats obtenus par BLUR par d'autres sources que celles proposées sur la ressource en ligne. C'est notamment ce que nous avons réalisé lors de notre recherche de gènes impliqués dans la multiciliation ; d'une part nous avons ajouté des protéomes d'organismes d'intérêts propres à notre question biologique (Otomorpha et Euteleosteomorpha), et d'autre part nous avons combiné ces résultats à un ensemble de résultats issus d'expériences de transcriptomique afin de prioriser les nouveaux gènes candidats que nous avons détectés. L'ensemble de ces travaux sera détaillé dans le chapitre suivant.

Chapitre 6 : Analyse intégrative de la multiciliation

L'étude évolutive de la multiciliation que nous avons réalisée plus tôt (voir Chapitre 4) a donné lieu à l'identification de quatre gènes candidats : ANKRD61, C4orf50, GZMA et GZMK. Elle nous a également permis de révéler des particularités chez les Otomorpha, sur lesquelles nous avons basé l'ensemble de nos travaux. En effet, ces poissons osseux, à multiciliation incomplète, présentent d'une part une absence de CDC20B et de la famille des miR-449, et d'autre part une divergence inattendue dans les séquences protéiques de MCIDAS et CCNO. Ces résultats ont mené au développement d'un nouvel outil de génomique comparative multi-niveaux, BLUR, conçu pour détecter des cas de pertes et divergences atypiques, dans notre cas chez les Otomorpha.

Pour identifier de nouveaux gènes candidats de la multiciliation, nous avons combiné des données de génomique comparative et de génomique fonctionnelle, à la manière d'une approche intégrative multi-étapes, dans le but de déterminer les cibles identifiées par le plus d'approches possibles. Dans ce chapitre, nous discuterons des différentes étapes de notre approche intégrative ; dans un premier temps nous avons appliqué BLUR à la multiciliation, puis, dans un second temps nous avons combiné les résultats issus de plusieurs expériences de transcriptomique dédiées à l'étude de la multiciliation. Enfin, nous avons intégré l'ensemble de ces résultats pour identifier les candidats potentiels à la multiciliation les plus pertinents.

1. Etude de la multiciliation par un profilage multi-niveaux

La première étape dans notre analyse multidisciplinaire de la multiciliation a consisté en l'exploitation du pan évolutif de la multiciliation par une approche massive, automatique et multi-niveaux, et ce par la recherche de conservations différentielles chez les Otomorpha grâce à l'outil BLUR.

1.1. Comparaison de deux groupes de poissons : Otomorpha et Euteleosteomorpha

1.1.1. Protocole de comparaison

Pour mette en évidence des cas de conservation atypique chez les Otomorpha, chez qui une multiciliation incomplète a été observée, nous avons choisi de les comparer à un groupe de poissons osseux proches, les Euteleosteomorpha, qui, à notre connaissance, présentent une multiciliation complète (*Figure 6-1*). Ces deux taxons sont des cohortes appartenant aux Clupeocephala et issus d'une séparation ayant eu lieu il y a environ 250 millions d'années (voir *Figure 4-5*). Les Otomorpha comprennent l'espèce modèle *Danio rerio*, ainsi qu'*Astyanax mexicanus*, tandis que les Euteleosteomorpha comportent notamment *Takifugu rubripes*, *Tetraodon nigroviridis* ou encore *Oreochromis niloticus*. Notre objectif étant d'identifier des gènes humains impliqués dans la multiciliation, *Homo sapiens* a été utilisé comme espèce de référence (organisme requête). Le protéome humain présente en outre l'avantage d'être particulièrement bien annoté. Dans BLUR, comme nous l'avons vu dans le chapitre précédent, il est possible d'utiliser les homologues détectés par BLAST ou les orthologues prédits par OrthoInspector. Des tests réalisés en utilisant les relations d'homologie ont faire ressortir un nombre important de faux positifs dus aux problèmes de qualité

de séquences dans les protéomes de poissons osseux et aux nombreux paralogues existant dans ces génomes. Nous avons donc choisi de nous appuyer sur les orthologues prédits par OrthoInspector.

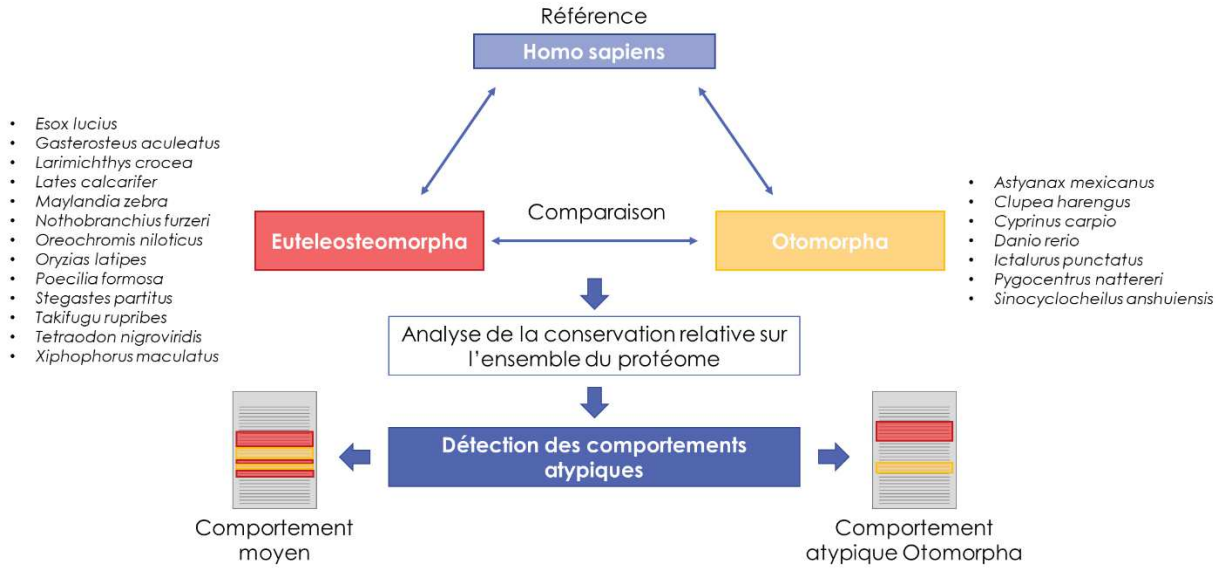

Figure 6-1: Schéma du protocole de recherche BLUR appliqué à la multiciliation.

1.1.2. Analyse globale des résultats

Les résultats de la comparaison des protéomes d'Euteleosteomorpha et d'Otomorpha par BLUR sont présentés dans le Tableau 6-1. Sur les 21 044 protéines du protéome humain utilisé, 14 853 présentent une conservation « classique », c'est-à-dire que les orthologues détectés dans les deux groupes de poissons sont conservés de façon similaire par rapport à la séquence humaine. 3 652 protéines humaines n'ont pas d'orthologues chez les groupes de poissons tandis que 1 178 protéines humaines sont plus conservées chez les Otomorpha que chez les Euteleosteomorpha. Enfin, 1 361 familles de protéines correspondant au groupe d'intérêt dans le cadre de l'étude de la multiciliation, présentent une divergence atypique chez les Otomorpha : 634 cas où aucune séquence orthologue d'Otomorpha n'a été retrouvée dans le résultat de BLAST, 104 cas de « Haute priorité » où une divergence prononcée a été détectée, et 623 cas de « Priorité moyenne » (*Tableau 6-1*).

Tableau 6-1: Tableau récapitulatif des résultats de BLUR lors de la comparaison des Otomorpha et des Euteleosteomorpha.

***Euteleosteomorpha over Otomorpha* (1361)**	Pas d'orthologue Otomorpha	**634**
	Priorité haute	**104**
	Moyenne priorité	**623**
***Otomorpha over Euteleosteomorpha* (1178)**	Pas d'orthologue Euteleosteomorpha	**479**
	Priorité haute	**98**
	Moyenne priorité	**601**
No hits found	Aucun orthologue	**3652**

Nous avons confronté les résultats obtenus par BLUR aux résultats obtenus lors de notre étude évolutive des gènes de la multiciliation présentée dans le chapitre 4. Cette étude nous avait permis de mettre en évidence d'une part l'absence de CDC20B chez les Otomorpha, et d'autre part une conservation différentielle dans les séquences protéiques de ces mêmes espèces dans deux familles : MCIDAS et CCNO. Tout comme ce fut le cas dans l'application du profilage phylogénétique par OrthoInspector, aucun orthologue de CDC20B n'est prédit chez les poissons osseux. En ce qui concerne la conservation différentielle, on retrouve MCIDAS dans la catégorie des protéines de haute priorité. Ces deux résultats confirment la capacité de BLUR à détecter des pertes mais aussi des divergences atypiques. CCNO n'est en revanche pas retrouvé dans les résultats de BLUR. Une analyse approfondie des résultats de BLAST montre que la région N-terminale, tronquée chez les Otomorpha (voir *Figure 4-8*), n'a pas été prise en compte dans l'alignement local réalisé par BLAST. En effet, cette région semble largement différente entre les Euteleosteomorpha et l'humain, en conséquence BLUR n'est pas en mesure de détecter une conservation atypique sur la base du reste de l'alignement majoritairement composé d'une région centrale conservée. Enfin, l'application du profilage phylogénétique à la multiciliation nous avait permis de faire ressortir deux candidats absents chez les Otomorpha, à savoir ANKRD61 et C4orf50, que nous retrouvons également avec BLUR.

L'analyse globale des trois types de résultats (absence d'orthologue chez les Otomorpha, conservation atypique avec priorité haute et moyenne) montre un léger enrichissement en termes GO liés à l'immunité, comme ce fut le cas lors du profilage phylogénétique réalisé avec OrthoInspector. La génération d'un réseau d'interactions protéiques à partir de l'intégralité des résultats révèle l'existence d'un réseau majoritaire comprenant 343 protéines qui, de manière intéressante, sont réparties dans les trois types de listes, soulignant les liens entre pertes de gènes et conservations différentielles. Bien que le réseau en lui-même soit également légèrement enrichi en protéines liées au système immunitaire (*'regulation of immune response', P-value* : 6.19E-05 ; *'regulation of interleukin-10 production', P-value :* 7.03E-05), on note l'existence de certains *clusters* à fonctions plus spécifiques, dont un contenant des protéines impliquées dans l'arrimage des corps basaux à la membrane (*Figure 6-2*).

Nous avons ensuite analysé chaque liste de résultats individuellement. Tout comme ce fut le cas pour la recherche de profil phylogénétique avec OrthoInspector, la liste de protéines absentes chez les Otomorpha présente un enrichissement en termes GO associés à l'immunité. Les résultats obtenus précédemment étant jugés plus précis de par la prise en compte des Ecdysozoaires, nous avons décidé de ne pas analyser de manière approfondie cette liste, à l'exception des protéines identifiées par d'autres approches par la suite. Le test de surreprésentation réalisé sur la liste de protéines de priorité haute a également montré un enrichissement en certains termes liés à l'immunité, tels que '*opsonization*' (*P-value* : 1.34E-07) ou '*regulation of complement activation*' (*P-value* : 2.65E-06), bien qu'ils ne concernent que 7 protéines parmi les 104 détectées. Enfin, le test d'enrichissement réalisé sur la liste de priorité moyenne n'a fait ressortir que très peu de termes surreprésentés, le majoritaire étant '*ribosome biogenesis*' (*P-value* : 1.93E-06). Comme nous l'avions observé dans le chapitre 4, les différences génotypiques observées entre Otomorpha et Euteleosteomorpha semblent reliées à plusieurs processus, ce qui nous a conduit à prioriser les listes de gènes obtenus.

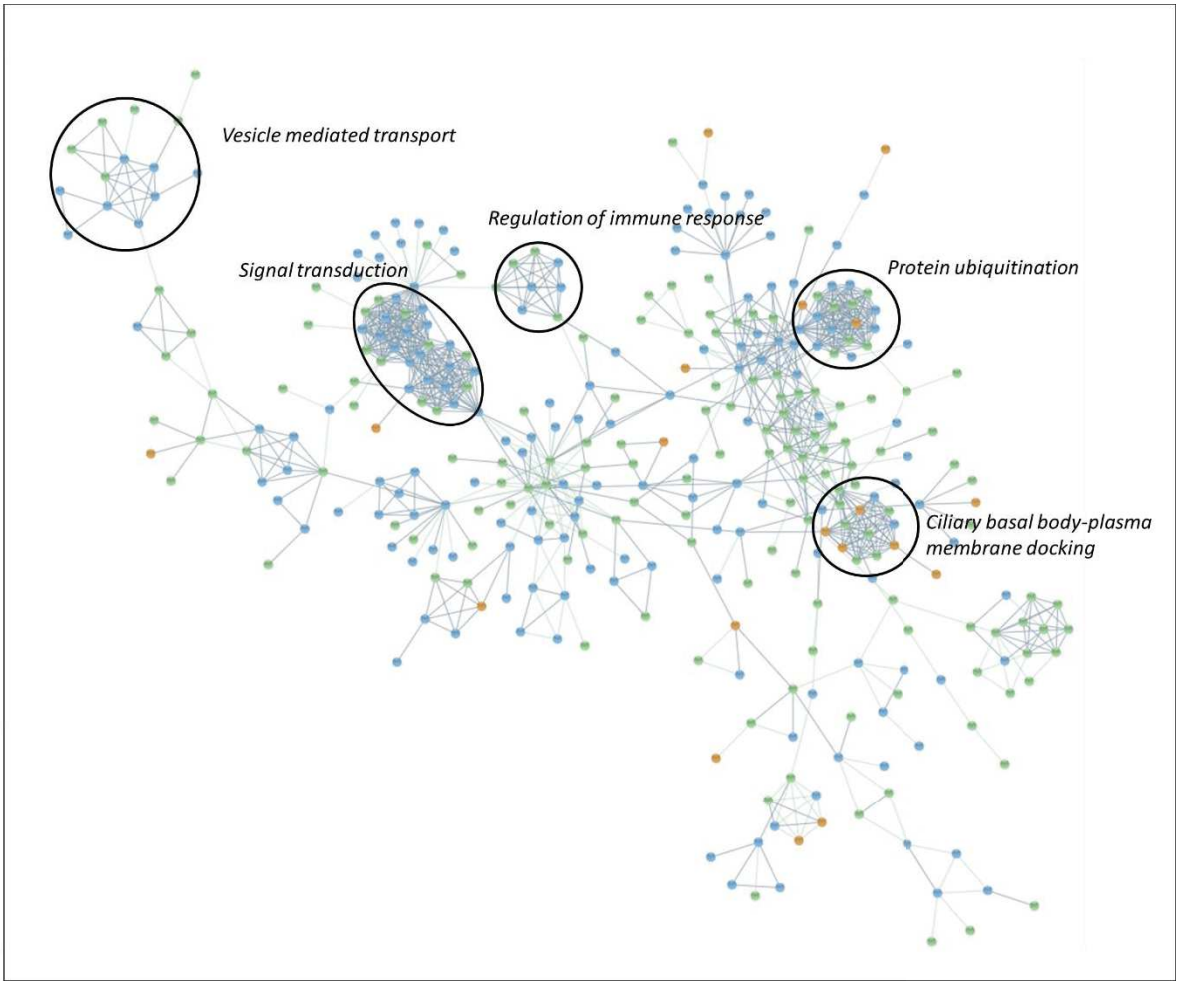

Figure 6-2: Réseau d'interaction majeur issu des trois listes de résultats de la comparaison par BLUR des Otomorpha et des Euteleosteomorpha. Les nœuds bleus représentent les protéines absentes chez les Otomorpha, en orange celles de « Haute priorité » et en vert celles de « Priorité moyenne ».

1.2. Priorisation des résultats

Afin de réaliser des analyses plus approfondies et confirmer la présence de conservation différentielle, nous avons tout d'abord priorisé nos résultats selon plusieurs critères : (1) protéines avec une annotation liée au cil ou au centrosome, (2) protéines de haute priorité, et (3) protéines de priorité moyenne interagissant avec une protéine de haute priorité selon la base de données STRING.

1.2.1. Protéines à annotation ciliaire ou centrosomale

Sur l'ensemble des 1 361 protéines détectées par BLUR, 39 possèdent une annotation de localisation cellulaire au niveau du centrosome, et 38 au niveau du cil, 8 étant communes aux deux. 28 de ces protéines sont absentes chez les Otomorpha, 10 sont de haute priorité, et 31 de moyenne priorité. Nous avons réalisé l'analyse manuelle des alignements multiples des protéines de haute et moyenne priorités pour confirmer ou infirmer la présence de conservations différentielles. Sur les 41 protéines détectées par BLUR comme étant divergentes, seules 18 présentent une réelle conservation atypique (*Tableau 6-2*), les autres étant des faux positifs majoritairement dus à des erreurs de prédictions de séquences dans les protéomes de poissons.

Tableau 6-2: Protéines prédites par BLUR présentant une conservation différentielle et une localisation ciliaire ou centrosomale.

Identifiant UniProt	Nom de gène	Priorité	Termes GO
AKAP9_HUMAN	AKAP9	Haute	Centrosome
CCD92_HUMAN	CCDC92	Haute	
CEP72_HUMAN	CEP72	Haute	
CNTLN_HUMAN	CNTLN	Moyenne	
HAUS1_HUMAN	HAUS1	Haute	
HAUS3_HUMAN	HAUS3	Moyenne	
PLAG1_HUMAN	PLAG1	Haute	
RFIP3_HUMAN	RAB11FIP3	Moyenne	
CE164_HUMAN	CEP164	Haute	Centrosome; cilium
CP110_HUMAN	CCP110	Haute	
MKKS_HUMAN	MKKS (BBS6)	Moyenne	
RIPL2_HUMAN	RILPL2	Moyenne	
EVC_HUMAN	EVC	Moyenne	Cilium
ICK_HUMAN	CILK1	Moyenne	
LBN_HUMAN	EVC2	Moyenne	
PHLP_HUMAN	PDCL	Haute	
TBCC_HUMAN	TBCC	Moyenne	
TILB_HUMAN	LRRC6	Moyenne	

Pour ces 18 séquences, il existe bien des divergences de séquences nettes comme dans EVC (*Figure 6-3*) et EVC2, deux gènes que l'on retrouve co-localisés chez l'Homme, la souris, le xénope, chez *Oreochromis niloticus*, un poisson du groupe Euteleosteomorpha, ainsi que chez *Astyanax mexicanus*, un Otomorpha. Ces deux gènes sont notamment connus pour être impliqués dans la ciliopathie Ellis-Van Creveld. La détection de ces 18 protéines ciliaires ou centrosomales par notre méthode valide l'approche que nous avons employée, mais met également en évidence un problème récurrent de l'étude de la multiciliation, qui est de pouvoir la séparer de la ciliation proprement dite, ces deux processus étant très largement interconnectés.

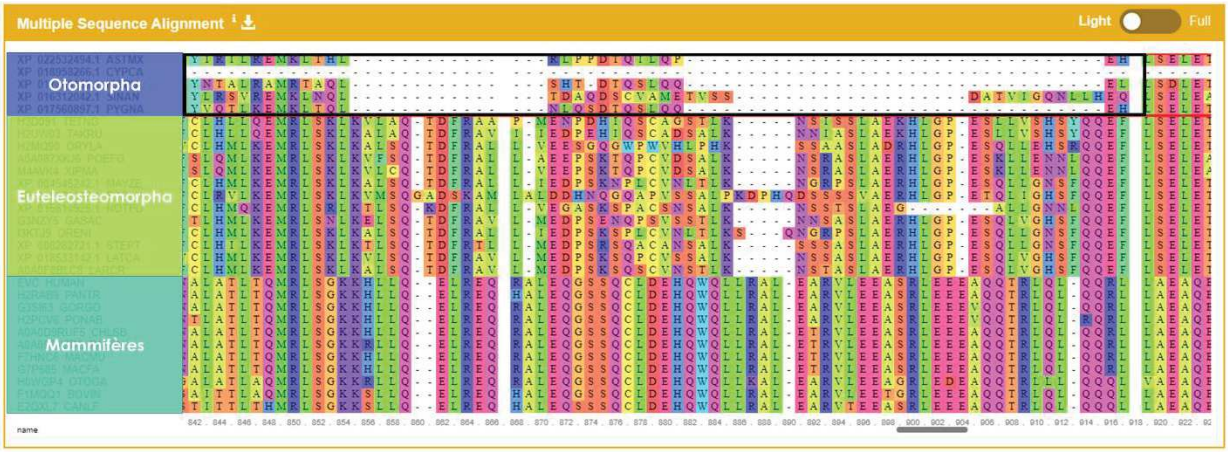

Figure 6-3: Section de l'alignement multiple d'EVC issu du site de BLUR. On note l'absence presque intégrale de la région présentée chez les Otomorpha (encadrée en noir).

1.2.2. Protéines de haute priorité

Après avoir étudié les protéines connues pour être liées au cil et au centrosome, nous nous sommes concentrés sur les protéines de haute priorité, au nombre de 104. L'analyse manuelle des alignements multiples a révélé 54 protéines présentant une divergence plus ou moins prononcée dans les séquences d'Otomorpha. Parmi elles, on retrouve notamment MSX1, une protéine de la famille Homeobox, impliquée dans le développement. De manière intéressante, MSX1 est retrouvé à proximité d'EVC et EVC2 chez l'Homme, la souris, le xénope, *Oreochromis niloticus*, *Astyanax mexicanus,* ainsi que chez la chimère *Callorhinchus milii* (*Figure 6-4*). Cela suggère que cette organisation existait chez l'ancêtre des Vertébrés, et que cette co-localisation peut être importante pour le fonctionnement de ces gènes.

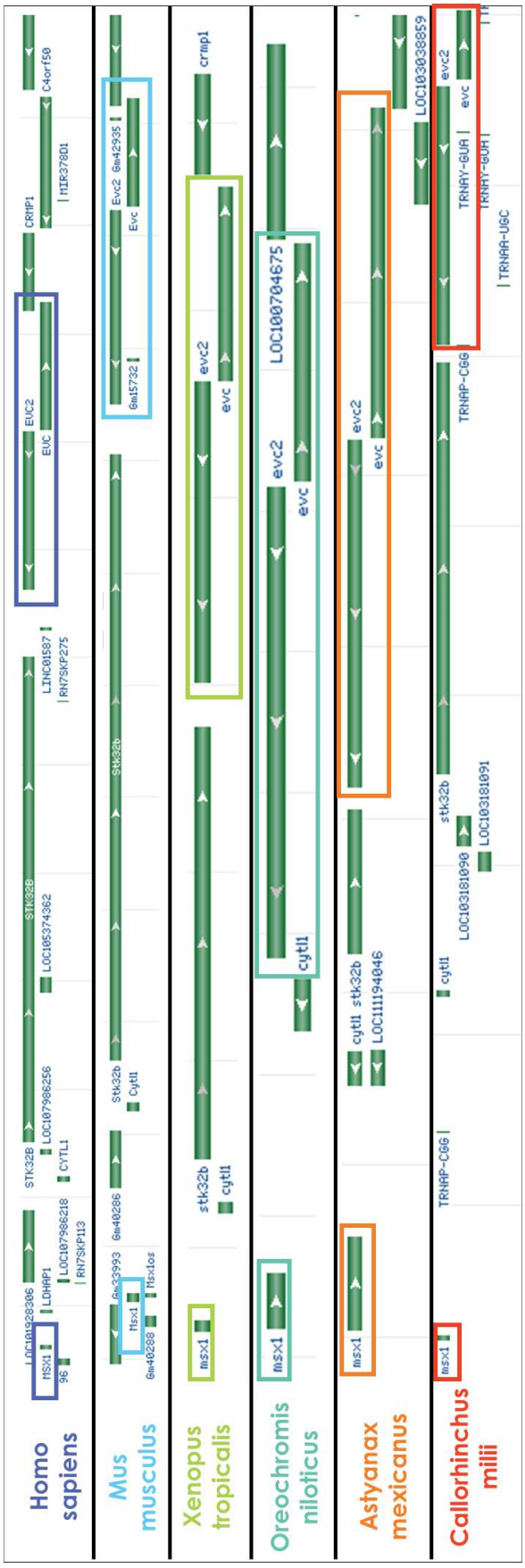

Figure 6-4: Contexte génomique de MSX1, EVC et EVC2. Ces gènes sont co-localisés chez l'Homme, la souris, le xénope, les poissons osseux et la chimère.

Parmi les gènes de haute priorité présentant une conservation différentielle, 5 n'ont pas ou peu d'annotations fonctionnelles GO et représentent des candidats particulièrement intéressants : CCDC126, c18orf25 (*Figure 6-5*), FAM181A, MANSC4 et TM6SF1. Des recherches approfondies sur diverses bases publiques nous ont permis de mettre en évidence une co-localisation de C18orf25 et HAUS1, un autre gène dont la protéine appartient à la liste de haute priorité (*Tableau 6-2*). Cette co-localisation est retrouvée chez de nombreux Vertébrés, parmi lesquels l'Homme, la souris, le xénope et *Danio rerio*.

Figure 6-5: Section de l'alignement multiple de C18ORF25 issu de BLUR.

1.2.3. Protéines de moyenne priorité

Enfin, la dernière étape de notre étude évolutive de la multiciliation a été d'analyser manuellement 27 des 623 protéines de priorité moyenne ayant une interaction avec une protéine de haute priorité sur la base des relations prédites par la base de données STRING. De cette manière, nous avons pu mette en évidence 6 protéines additionnelles pour lesquelles il existe une conservation différentielle chez les Otomorpha : ASB6, CACNB2, CDK11A, CHUCH1, HAUS3 et NPPC.

1.2.4. Bilan

L'analyse des résultats de BLUR nous a permis de mettre en évidence un certain nombre de faux résultats positifs, en grande partie dus aux séquences protéiques de poissons osseux mal prédites. Il est à noter que la proportion de faux positifs est moindre dans la liste de protéines de haute priorité (48%) par rapport aux protéines de moyenne priorité que nous avons inspectées manuellement (78%). Au total, 60 protéines parmi les 157 dont nous avons analysé manuellement les alignements ont été retenues comme présentant une conservation atypique chez les Otomorpha. Nombre de ces résultats disposent d'éléments de validation supplémentaires, tels que des annotations ciliaires, une conservation de synténie chez un ensemble d'espèces ou encore des interactions fonctionnelles entre elles. La liste des 60 protéines est disponible en annexe.

2. Analyse fonctionnelle

La seconde partie de notre analyse intégrative de la multiciliation a consisté en une étude fonctionnelle des gènes impliqués dans ce processus, et ce par la comparaison de différents jeux de données issus d'expériences de transcriptomique.

2.1. Jeux de données de transcriptomique

Nous avons choisi un ensemble de 8 jeux de données disponibles sur la plateforme GEO du NCBI, sélectionnés pour leur spécificité vis-à-vis de la multiciliation, en excluant les expériences de *scRNASeq* ou celles ne comportant pas de réplicas (*Tableau 6-3*). Les jeux de données sélectionnés sont issus d'expériences réalisées chez le xénope ou chez la souris, et consistent en l'inactivation d'un gène impliqué dans la cascade de signalisation de la multiciliation, afin de visualiser l'effet de ce gène par comparaison de valeurs d'expression. L'ensemble des 8 jeux de données contient les résultats de 10 expériences, que l'on peut globalement séparer en deux groupes selon le type de gènes étudiés : les expériences sur les gènes précoces (STK11, MCIDAS, E2F4 et p73), et les expériences sur les gènes plutôt tardifs (Myb, Foxj1 et Foxn4). Outre des comparaisons d'individus *wild type* à des individus pour lesquels un gène a été inactivé (GSE75715 : *knockout* p73 ; GSE73331 : *knockout* E2F4), ces expériences obéissent à différents protocoles dont nous expliquons rapidement le principe ci-dessous.

GSE32452. Stubbs et collaborateurs comparent l'expression de gènes entre des embryons de *Xenopus laevis* dans lesquels le domaine intracellulaire de Notch a été injecté, et des embryons dans lesquels ont été injectés le domaine intracellulaire de Notch et une forme de Multicilin (codée par MCIDAS) inductible par des glucocorticoides (Stubbs et al., 2012).

GSE59309. Dans cette expérience, les auteurs comparent l'expression de gènes dans des progéniteurs épithéliaux d'embryons de xénope dont la différenciation a été induite par la Multicilin, en présence ou non d'une forme d'E2F4 tronquée à laquelle il manque les 140 derniers acides aminés de la région C-terminale (Ma et al., 2014).

GSE89271. Ce jeu de données comprend les résultats de trois expériences visant à évaluer le rôle de FOXN4 dans la multiciliation et dans quelle mesure celui-ci est similaire à FOXJ1. Ainsi, les auteurs comparent l'expression des gènes entre des embryons de contrôle dans lesquels ont été injectés de la Multicilin inductible par glucocorticoides, et (1) des embryons injectés avec un morpholino de FOXN4 ciblant le site du début de traduction et de la Multicilin inductible, (2) des embryons injectés avec la protéine Cas9, un ARN guide ciblant FOXN4 et de la Multicilin inductible, ou (3) des embryons injectés avec la protéine Cas9, un ARN guide ciblant FOXJ1 et de la Multicilin inductible (Campbell et al., 2016).

GSE76342. Quigley et Kintner ont ici tenté d'identifier les gènes essentiels de la multiciliation par la comparaison de trois paires de conditions différentes, et ce dans des progéniteurs épithéliaux de xénope à trois temps donnés (3h, 6h et 9h). Ils comparent ainsi : (1) le blocage de Notch et l'injection du domaine intracellulaire de Notch, (2) l'injection du seul domaine intracellulaire de Notch et l'injection de ce domaine combinée à de la Multicilin, et (3) le blocage de Notch et le blocage de Notch avec injection d'une forme inactive de Multicilin (Quigley and Kintner, 2017).

GSE60365. Dans cette expérience, les auteurs comparent dans des cellules trachéales de souris, l'expression de gènes en présence d'un *shRNA* (*short hairpin RNA*) contrôle, et d'un shRNA destiné à bloquer l'expression de MYB (Pan et al., 2014).

GSE116690. Pour caractériser le rôle de STK11 dans la multiciliation, Chu et collaborateurs ont comparé l'expression des gènes dans les cellules de poumons de souris *wild type* à celle observée chez des souris dans lesquelles STK11 a été spécifiquement supprimé dans les cellules progénitrices du poumon par le système Cre-lox (Chu et al., 2019).

Tableau 6-3: Jeux de données d'expériences de transcriptomique utilisés pour l'analyse fonctionnelle de la multiciliation.

Accession	Expérience	Espèce	Technique	Publication
GSE32452	*Notch intracellular domain* (ICD) *vs* ICD + Multicilin	*Xenopus laevis*	Puce à ADN	(Stubbs et al., 2012)
GSE59309	Multicilin inductible *vs* Multicilin inductible + E2F4 tronqué	*Xenopus laevis*	*RNASeq*	(Ma et al., 2014)
GSE89271	Multicilin inductible *vs* Multicilin inductible + morpholino Foxn4	*Xenopus laevis*	*RNASeq*	(Campbell et al., 2016)
	Multicilin inductible *vs* Multicilin inductible + mutant Foxn4 CRISPR/Cas9	*Xenopus laevis*	*RNASeq*	
	Multicilin inductible *vs* Multicilin inductible + mutant Foxj1 CRISPR/Cas9	*Xenopus laevis*	*RNASeq*	
GSE76342	Notch- *vs* ICD; ICD *vs* ICD + Multicilin; Notch- *vs* Notch- + Multicilin-	*Xenopus laevis*	*RNASeq*	(Quigley and Kintner, 2017)
GSE60365	*Non-targeted* shRNA *vs* Myb shRNA	*Mus musculus*	Puce à ADN	(Pan et al., 2014)
GSE75715	*Wild Type vs* p73 *knockout*	*Mus musculus*	*RNASeq*	(Nemajerova et al., 2016)
GSE73331	*Wild Type vs* E2F4 *knockout*	*Mus musculus*	Puce à ADN	(Mori et al., 2017)
GSE116690	Stk11+ *vs* Stk11-	*Mus musculus*	*RNASeq*	(Chu et al., 2019)

2.2. Traitement et formatage

Les données étant issues de plateformes différentes (*RNASeq*, puces à ADN), d'espèces différentes, et d'années différentes, il a été nécessaire d'adapter les différents traitements à ces données hétérogènes de manière à obtenir des résultats comparables.

Nous avons récupéré les données préalablement traitées par les auteurs des expériences lorsqu'elles étaient disponibles ; dans le cas contraire, nous avons analysé les données brutes avec l'outil GEO2R du NCBI si cela était possible (voir Chapitre 8: Matériel et Méthodes). Enfin, dans le cas où uniquement les données brutes étaient disponibles sans possibilité d'analyse par GEO2R, nous

avons suivi les méthodes d'analyse employées par les auteurs de la publication originale. Pour l'ensemble des jeux de données, nous avons sélectionné les gènes surexprimés dans la condition multiciliée en utilisant les seuils de 1 pour la valeur de logFC (valeur logarithmique du *Fold Change*) et de 0.05 pour la *P-value*.

Nous l'avons mentionné dans le chapitre 3, une des difficultés lors de la comparaison de plusieurs jeux de données provenant de sources différentes est de pouvoir faire la correspondance entre des expériences réalisées sur différentes plateformes. Dans notre cas, nous avons été confrontés à des données issues d'expériences sur le xénope, ainsi que sur la souris, datant de 2012 à aujourd'hui, contenant des identifiants parfois obsolètes. Nous avons choisi de travailler avec les noms de gènes humains pour faciliter la correspondance entre les différentes espèces, bien qu'un certain nombre de séquences ne soit pas associé à des gènes dans les fichiers originaux déposés sur la banque GEO. Pour obtenir les informations les plus à jour possible, nous avons utilisé des ressources spécifiques à chacun des organismes pour retrouver les gènes correspondant ; pour la souris nous avons employé la ressource *Mouse Genome Informatics* [http://www.informatics.jax.org/], et pour le xénope, la ressource Xenbase (Karimi et al., 2018). Enfin, pour mettre à jour les annotations pouvant être obsolètes, nous avons eu recours à l'outil Retrieve/ID Mapping d'UniProt. Malgré l'ensemble de ces recherches, près de 150 identifiants restent sans gènes correspondants ; il s'agit en grande partie de séquences de type long ARN non codant (*lncRNA*).

2.3. Comparaison par *clustering*

2.3.1. *Protocole de clustering*

Nous avons compilé les résultats d'analyse d'expression différentielle des 10 expériences dans une matrice binaire à deux dimensions, avec en lignes les 4151 gènes différentiellement exprimés dans au moins une expérience, en colonne les expériences. La valeur d'une cellule est de 1 dans le cas d'une expression différentielle du gène correspondant dans l'expérience considérée. Dans le but de faciliter l'interprétation de l'ensemble de ces résultats, nous avons réalisé un *clustering* des gènes selon leur expression différentielle dans les 8 jeux de données (*Figure 6-6*). Pour cela, nous avons utilisé l'indice de Jaccard pour mesurer les distances entre profils, et l'algorithme de Ward (Ward, 1963) pour réaliser le *clustering* (voir Chapitre 8 : Matériel et Méthodes). Parmi les 28 *clusters* obtenus, 10 correspondent à des gènes différentiellement exprimés uniquement dans une expérience, pour un total de 2279 gènes (54,9% de l'ensemble des gènes). Ces gènes ont été écartés de nos analyses ultérieures. L'ensemble des gènes surexprimés dans au moins deux expériences sont disponibles en annexe, regroupés par clusters similaires.

2.3.2. *Analyse des clusters*

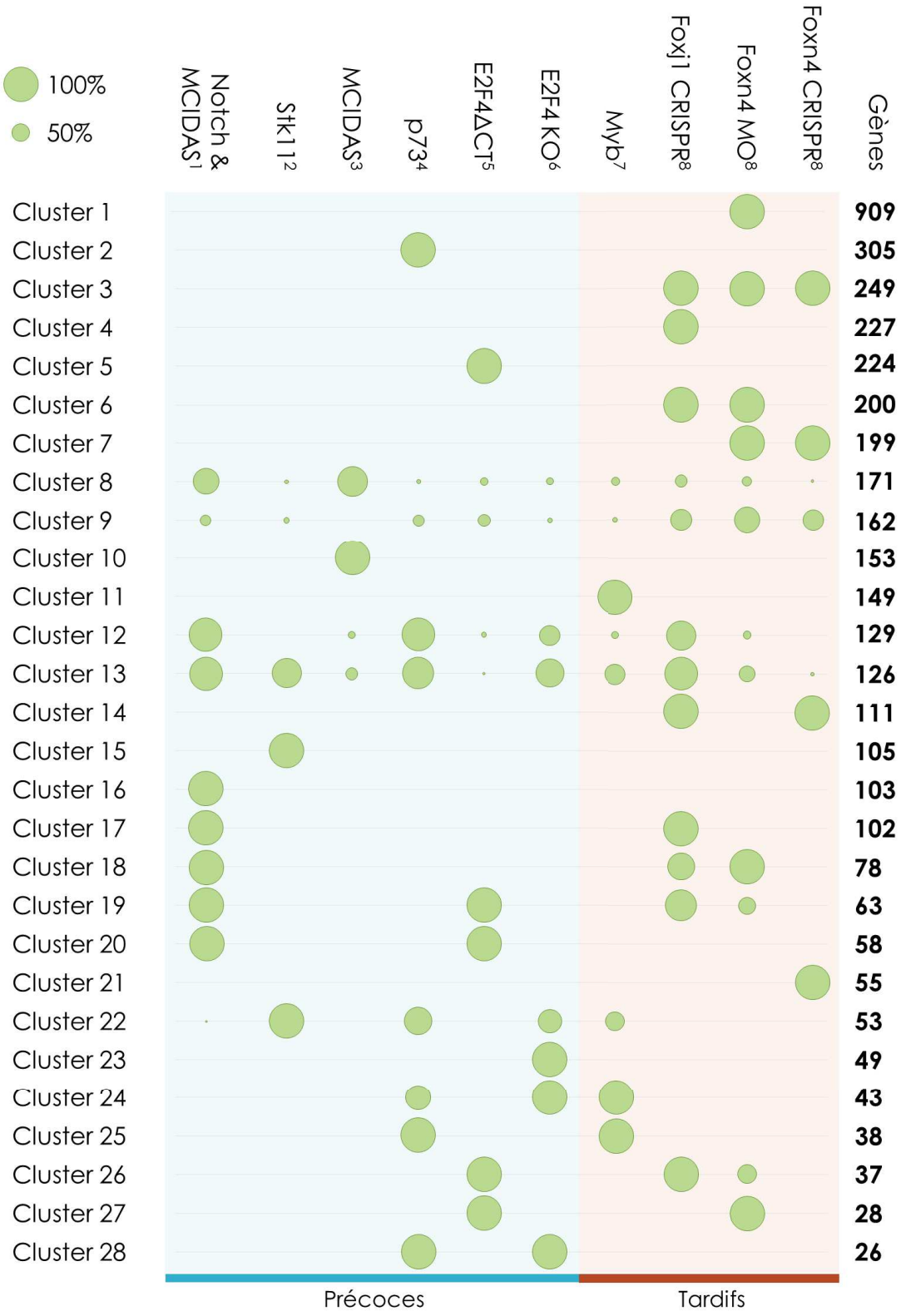

Figure 6-6: Représentation schématique du clustering des résultats d'expériences de génomique fonctionnelle. La taille des cercles est proportionnelle au pourcentage de gènes du cluster détectés dans l'expérience considérée.[1]GSE76342 ; [2]GSE116690 ; [3]GSE32452 ; [4]GSE75715 ; [5]GSE59309 ; [6]GSE73331 ; [7]GSE60365 ; [8]GSE89271.

Clusters des cibles tardives. Il est possible de regrouper les *clusters* restants par similarité de profil d'expression ; on retrouve par exemple les gènes cibles de FOXJ1 et FOXN4, à expression plutôt tardive, dans 4 clusters distincts (3, 6, 7 et 14). Nous avions mentionné précédemment que FOXJ1 et FOXN4 étaient des facteurs de transcription aux rôles parfois redondants, impliqués de manière plutôt tardive dans le processus de multiciliation, notamment dans l'arrimage des corps basaux et dans la ciliogénèse. L'enrichissement en termes GO de ces 758 gènes cibles révèle une surreprésentation de gènes impliqués dans l'établissement de la localisation de différents composés biologiques (*P-value* : 1.11^{E}-15), et beaucoup d'entre eux se trouvent localisés au niveau du système endomembranaire (*P-value* : 3.76^{E}-20). Cet enrichissement peut s'expliquer d'une part par les nombreux rôles joués par les protéines de la famille FOX au cours du développement, mais également par leur participation dans la ciliogénèse, au cours de laquelle de nombreux mécanismes de transport vésiculaire sont réquisitionnés (Long and Huang, 2020).

Clusters des cibles centrales. Les clusters 17, 18, 19, 26 et 27 regroupent les gènes influencés par le complexe MCIDAS/E2F4 et au moins l'un des deux gènes Foxj1 ou Foxn4 (*Figure 6-7*). Pour rappel, MCIDAS se lie nécessairement à E2F4 pour fonctionner, notamment pour activer Foxj1 et Foxn4, qui, eux, présentent des fonctions en grande partie redondantes. L'ensemble de ces 308 gènes est enrichi en gènes liés au cil ('*cilium assembly*', *P-value* : 1.71E-21) ; on retrouve notamment parmi eux quelques gènes centrosomaux ou d'autres impliqués dans le transport intraflagellaire. De manière intéressante, 44 gènes appartenant à ces *clusters* ne présentent aucune annotation GO chez l'Homme, et constituent de nouvelles cibles prometteuses à analyser dans l'étude du cil et de la multiciliation.

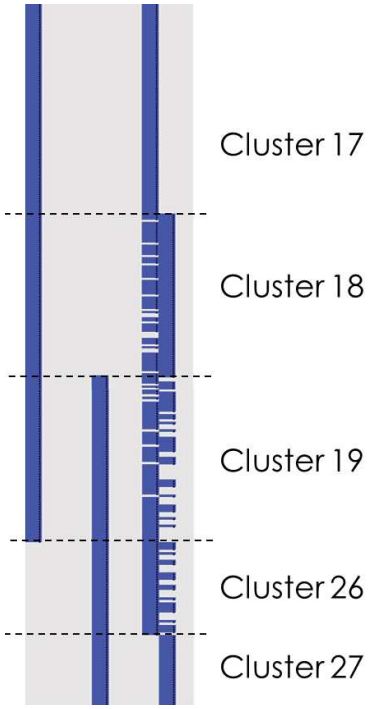

Figure 6-7: Exemple de clusters regroupés par profil d'expression similaire. Les lignes correspondent aux gènes, les colonnes aux différentes expériences analysées.

Clusters murins. Les *clusters* 22, 24, 25 et 28 regroupent 159 gènes influencés par une combinaison des gènes stk11, p73, E2F4 et Myb. Il s'agit de *clusters* ne contenant que des gènes issus

d'expériences réalisées sur la souris, il se peut donc que ces gènes fassent partie d'une régulation particulière, spécifique à la souris, ou qu'il n'existe pas d'orthologues de ces gènes chez le xénope, expliquant leur ségrégation ; on note la présence de 26 gènes dont l'identifiant ne correspond à aucune protéine de l'Homme disponible dans la base de données *Gene Ontology*. On retrouve néanmoins un léger enrichissement en termes liés au cil, notamment '*axoneme assembly*' (*P-value* : 8.84E-07) et '*cilium movement*' (*P-value* : 8.84E-07).

Cluster 20. Le *cluster* 20 contient des gènes dont l'expression est uniquement modifiée lors de variations dans l'expression de MCIDAS et E2F4, et dont les fonctions semblent liées à la multiplication des corps basaux. On retrouve en effet des enrichissements dans les termes GO suivants : '*positive regulation of centriole elongation*' (*P-value* : 1.12E-4), '*centriole replication*' (*P-value* : 4.46E-09), '*ciliary basal body-plasma membrane docking*' (*P-value* : 1.31E-08).

Clusters ciliés/multiciliés. Enfin, les 4 derniers *clusters* (8, 9, 12, 13) contiennent les gènes qui semblent être influés par le plus de facteurs liés à la multiciliation. On retrouve ainsi la majorité des gènes impliqués dans la multiciliation au sein de ces *clusters* : MCIDAS, DEUP1, CEP152, PLK4, MYB, CCP110, CDC20B, FOXJ1 et FOXN4 sont retrouvés dans le *cluster* 8, CCNO et CCDC78 dans le *cluster* 13, et RFX2 et RFX3 dans le *cluster* 12. Le test de surreprésentation fonctionnelle de ces 588 gènes a mis en évidence un enrichissement en termes ciliaires tels que '*cilium assembly*' (*P-value* : 2.38E-77), '*cilium movement*' (*P-value* : 3.41E-38), ou encore '*intraciliary transport*' (*P-value* : 2.96E-17). On retrouve également l'annotation '*de novo centriole assembly involved in multi-ciliated epithelial cell differentiation*' (*P-value* : 6.36E-05). De manière intéressante, plus de 60 gènes de ces 4 clusters ne possèdent aucune annotation GO pour leurs orthologues humains.

2.3.3. Bilan

Nous avons réalisé ici la première analyse intégrative de données de génomique fonctionnelles appliquées à la multiciliation, avec un croisement de données d'expression issues de 10 expériences. Au total, sur les 4151 gènes détectés dans ces différentes expériences de transcriptomique, 1872 ont été détectés dans au moins 2 jeux de données et ont été le point focal de notre analyse. L'intégration de plusieurs jeux de données indépendants permet en effet de renforcer la fiabilité des résultats, ce qui est démontré par la cohérence des enrichissements fonctionnels obtenus. La réalisation d'un *clustering* a permis de mettre en évidence des groupes de gènes fortement liés au cil et à la multiciliation. Au sein des 1872 gènes, un total de 152 gènes ne présentent pas d'annotation fonctionnelle de type *Gene Ontology* chez leurs orthologues humains, et constituent ainsi des cibles prometteuses à étudier de manière approfondie.

3. Intégration des résultats évolutifs et fonctionnels

Dans le but de prioriser l'ensemble des résultats que nous avons obtenus et d'obtenir des candidats multiciliés en vue d'une validation expérimentale, nous avons intégré des informations issues de différentes sources. Outre la comparaison des deux approches décrites précédemment, nous avons également comparé nos résultats à ceux de la base de connaissance CiliaCarta (van Dam et al., 2019), issue de l'approche intégrative décrite dans le chapitre 3. Il est à noter que cette base ne s'intéresse pas à la multiciliation de manière spécifique, mais à la ciliation en général.

3.1. Analyse globale

Nous avons commencé par une comparaison générale de l'ensemble de nos résultats : les 1361 protéines détectées par BLUR, les 1872 gènes détectés dans au moins deux jeux de données transcriptomiques, et les 936 gènes composants la base CiliaCarta (*Figure 6-8*). Comme l'on peut s'y attendre, l'analyse transcriptomique et CiliaCarta, toutes les deux issues d'expériences de génomique fonctionnelle, présentent le plus grand recoupement avec 362 gènes en commun, tandis que BLUR ne partage que 113 et 57 gènes avec l'analyse fonctionnelle et CiliaCarta respectivement. Enfin, 22 gènes sont détectés par les 3 méthodes. Aucun des gènes candidats identifiés par profilage phylogénétique ou par étude de la synténie (ANKRD61, C4orf50, GZMA et GZMK) n'est retrouvé par l'analyse fonctionnelle ou par CiliaCarta.

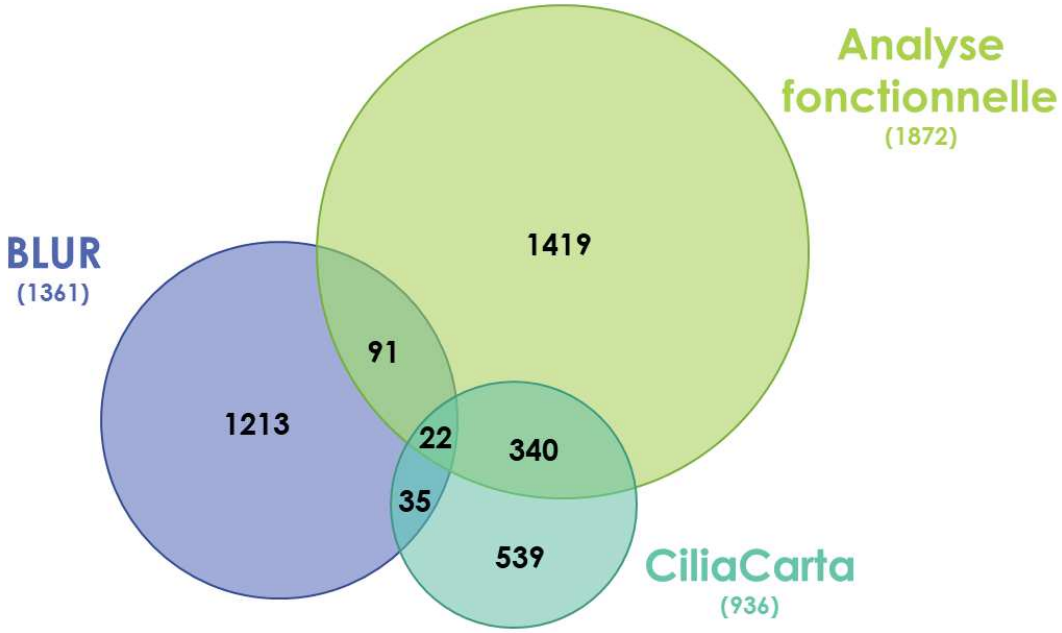

Figure 6-8: Diagramme de Venn montrant les recoupements entre l'approche évolutive, l'approche fonctionnelle et la base de connaissance CiliaCarta.

3.2. Comparaisons deux à deux

3.2.1. Analyse évolutive et analyse fonctionnelle

Parmi les 113 gènes communs entre BLUR et l'analyse fonctionnelle que nous avons réalisée, 8 font partie des gènes pour lesquels une conservation différentielle a été confirmée dans les séquences protéiques d'Otomorpha : AKAP9, CCDC92, CCP110, CEP164, FAM181A, IQCK, MCIDAS et PLAC8. Bien que la majorité d'entre eux soit déjà associée au cil ou à la multiciliation, trois gènes n'ont jusqu'à présent pas été liés à ces processus : PLAC8, qui semble être impliqué dans la régulation de l'apoptose et la prolifération cellulaire, IQCK et FAM181A, pour lesquels aucune fonction n'est pour l'instant annotée dans la base *Gene Ontology*. Une étude récente a néanmoins montré une interaction entre FAM181A et les facteurs de transcription TEAD, impliqués dans la voie de signalisation Hippo et dans le développement des organes (Bokhovchuk et al., 2020). En plus de ces cas de conservation différentielle, 41 des gènes communs entre les deux approches sont absents chez les Otomorpha, dont 5 pour lesquels aucun terme fonctionnel GO n'est associé : ANKRD60, C1orf189, C20orf85, CFAP47 et TEX43 (*Tableau 6-4*).

3.2.2. *Comparaisons à CiliaCarta*

Bien que CiliaCarta soit une ressource orientée vers le cil et non la multiciliation, nous avons tout de même comparé les résultats obtenus par BLUR et par notre analyse fonctionnelle à son contenu. Ainsi, l'intersection entre les résultats de BLUR et CiliaCarta comprend 35 gènes, dont 16 associés au cil, et 3 faisant partie des candidats identifiés précédemment comme présentant une conservation différentielle chez les Otomorpha (HAUS1, HAUS3 et CEP72). En ce qui concerne les 340 gènes communs entre notre analyse fonctionnelle et CiliaCarta, on retrouve encore une fois une large proportion de gènes liés au cil (193 ayant l'annotation GO '*cilium*'). 21 des 340 gènes ne possèdent aucune annotation GO chez l'Homme et représentent des candidats intéressant à étudier dans le cadre de la ciliation.

3.3. Comparaison des trois approches

Les 22 gènes communs aux trois approches sont pour la majorité déjà associés au cil d'une manière ou d'une autre, et confirment en un sens la pertinence de notre approche intégrative. 5 gènes ne sont néanmoins pas associés au cil à l'heure actuelle : C20orf85 et IQCK, dont nous avons déjà parlé précédemment, et ALDH1A1, DYDC1 et LRRC43 (*Tableau 6-4*).

Bien qu'ALDH1A1 fasse partie de la voie métabolique du rétinol, ce gène est absent chez les Otomorpha et semble être affecté par l'expression de p73 et MYB (*cluster* 25 dans notre analyse fonctionnelle), et son score CiliaCarta est égal à 1,88 (un score de 1 signifie qu'il est deux fois plus probable que le gène soit ciliaire que non-ciliaire). De manière intéressante, ALDH1A1 semble être impliqué dans la synthèse de l'acide rétinoïque (Labrecque et al., 1995), lui-même récemment impliqué dans un processus de régulation de la multiciliogénèse dans le pronephros de *Danio rerio* (Marra et al., 2019).

Les gènes DYDC1 et LRRC43 ont été tous les deux détectés par BLUR comme présentant une conservation différentielle de priorité moyenne, en revanche, l'analyse manuelle des alignements multiples ne permet pas de confirmer avec certitude l'existence d'une divergence nette, la qualité des séquences d'Otomorpha étant relativement discutable pour ces familles. DYDC1 semble être impliqué dans la méthylation d'histone de type H3-K4, néanmoins, son score CiliaCarta est de 1.55 et on le retrouve au sein du *cluster* 8 dans notre analyse fonctionnelle, *cluster* qui contient également la majorité des gènes multiciliés que nous avons étudiés. LRRC43 quant à lui ne présente pas d'annotation GO chez l'Homme, appartient au *cluster* 22 (cluster murin), et son score CiliaCarta est également de 1.55.

Dans l'ensemble, nous avons pu identifier 11 nouveaux gènes candidats prédits à la fois par notre approche évolutive et notre approche fonctionnelle, dont 5 également présents dans la base CiliaCarta (*Tableau 6-4*). De par leur absence d'annotation liée au cil, ces gènes représentent des candidats particulièrement prometteurs à la multiciliation qu'il serait intéressant d'investiguer par des protocoles expérimentaux.

Tableau 6-4: Tableau récapitulatif des candidats à la multiciliation les plus prometteurs. Méthode de détection : Analyse fonctionnelle (A) ; BLUR (B) ; CiliaCarta (C).

Nom de gène	Méthodes de détection	Annotations GO	Statut BLUR	Cluster fonctionnel	Score CiliaCarta
ALDH1A1	A;B;C	Retinoid metabolic process	Absent Otomorpha	25	1,88
ANKRD60	A;B	Aucune	Absent Otomorpha	17	-
C1orf189	A;B	Aucune	Absent Otomorpha	13	-
C20orf85	A;B;C	Aucune	Absent Otomorpha	18	4,34
CFAP47	A;B	Aucune	Absent Otomorpha	12	-
DYDC1	A;B;C	Histone H3-K4 methylation	Priorité moyenne	8	1,55
FAM181A	A;B	Aucune	Conservation différentielle	9	-
IQCK	A;B;C	Aucune	Conservation différentielle	12	1,55
LRRC43	A;B;C	Aucune	Priorité moyenne	22	1,55
PLAC8	A, B	Cell population proliferation	Haute priorité	9	-
TEX43	A;B	Aucune	Absent Otomorpha	18	-

4. Discussion

Les approches intégratives sont de plus en plus souvent employées pour étudier des processus et les caractériser de la façon la plus complète possible ; nous avons ici cherché à approfondir nos connaissances sur la multiciliation en couplant une étude de génomique comparative et des études de génomique fonctionnelle dans une approche intégrative multi-étapes.

Notre approche de génomique comparative, basée sur l'hypothèse selon laquelle la multiciliation incomplète observée chez les Otomorpha est due ou se reflète par la perte de gènes mais aussi par des divergences de séquence, a permis de faire ressortir des cas de conservation différentielle dans plusieurs protéines déjà connues pour être impliquées dans le cil ou la multiciliation. Notre application de BLUR à la multiciliation a également mis en évidence une des limitations du programme dont nous avons déjà discuté dans le chapitre précédent, à savoir sa forte dépendance à la qualité des protéomes des bases de données publiques. En effet, nous avons détecté de nombreux faux positifs dans notre analyse dus à des séquences protéiques de poisson osseux mal prédites. Notre approche de génomique fonctionnelle, quant à elle, a consisté en la comparaison de 10 expériences de transcriptomique visant l'étude de la multiciliation, et, après un *clustering*, a permis de faire ressortir un ensemble de 152 gènes détectés par au moins deux expériences, pour lesquels aucune annotation GO n'est associée chez l'Homme.

En couplant nos deux approches, nous avons pu faire ressortir 113 gènes communs, dont un grand nombre déjà liés au cil, confirmant ainsi que notre approche est pertinente et permet en effet de dégager des gènes impliqués dans ces processus. Nous avons également comparé nos résultats avec la base de connaissance CiliaCarta, elle-même issue d'une approche intégrative couplant de nombreuses expériences de génomique ciblées sur le cil. Nous avons ainsi mis en évidence 22 gènes détectés par l'ensemble des approches. Cela met également en avant une difficulté majeure de

l'étude de la multiciliation qui est de pouvoir dissocier la ciliation et la multiciliation à proprement parler. En effet, ces deux processus sont très largement interconnectés et il est difficile d'étudier les processus de multiciliation par des approches fonctionnelles sans impliquer de gènes ciliaires, qui en sont des composants intégraux. En ce sens, notre approche de génomique comparative basée sur une signature évolutive liée à une multiciliation incomplète chez les Otomorpha offre une opportunité unique de prioriser des gènes plus spécifiquement liés à la multiciliation.

L'ensemble de cette approche intégrative a permis de révéler de nombreux gènes cibles dont 11 gènes candidats particulièrement prometteurs n'étant jusqu'alors pas liés au cil et étant détectés par au moins 2 approches. Il serait à présent intéressant de valider l'ensemble de ces résultats de manière expérimentale. A l'avenir, nos approches pourraient être complétées par de nouvelles études de génomique fonctionnelle notamment temporelles, afin de caractériser les gènes dans le temps et ainsi identifier d'autres cibles potentielles. Dans une optique similaire, il est également envisageable d'intégrer des résultats issus d'expériences de *scRNAseq* pour identifier des gènes spécifiquement exprimés dans des cellules en cours de différenciation en MCC.

Chapitre 7 : Conclusions et Perspectives

Les avancées technologiques de ces dernières années sont à l'origine d'une augmentation exponentielle des données biologiques générées de façon quotidienne, qu'il devient de plus en plus complexe d'appréhender. Cela nous donne néanmoins l'opportunité d'étudier des processus jusqu'alors méconnus et de les caractériser de la façon la plus complète possible. Ces travaux de thèse nous ont permis, par l'étude de la multiciliation, d'établir la place de la génomique comparative à la fois dans l'ère des disciplines à haut débit, mais également dans les approches intégratives, majoritairement dédiées aux analyses fonctionnelles.

1. La multiciliation, un processus complexe

Nous l'avons vu tout au long de ce manuscrit, la multiciliation est un processus complexe, à la fois d'un point de vue biologique, mais également en tant que sujet d'étude. Nos connaissances à son sujet sont encore partielles, et peu de gènes lui sont associés ; on ne retrouve d'ailleurs que 8 gènes dans les banques publiques avec l'annotation fonctionnelle '*multi-ciliated epithelial cell differentiation*' (GO:1903251), bien que l'on en recense en réalité un plus grand nombre. De même, aucun gène n'est associé aux termes correspondants aux blépharoplastes des plantes, structures semblables aux deutérosome que l'on retrouve dans les spermatozoïdes multiciliés de certaines plantes (Hodges et al., 2012).

Au cours de notre étude approfondie de la multiciliation, nous avons été confrontés à plusieurs difficultés inhérentes à l'analyse de ce processus. Une de ces difficultés s'est révélée être l'étroite association de la multiciliation à la ciliation, qu'il est délicat de séparer, en particulier dans des analyses fonctionnelles. La seconde spécificité de la multiciliation est sa tendance à recycler les mécanismes préexistants et de les adapter à son besoin, ce qui souligne d'un côté toute la beauté du Vivant, mais rajoute une dimension de complexité supérieure pour le monde de la recherche. Enfin, il apparaît clairement que la multiciliation est un processus versatile dont la régulation et le fonctionnement ne semblent pas être identiques chez l'ensemble des espèces, voire l'ensemble des tissus.

1.1. Evolution des voies d'amplification chez les Métazoaires

Les travaux réalisés au cours de cette thèse nous ont permis de caractériser plus précisément les voies d'amplification des corps basaux à la lumière de l'évolution. Les résultats de notre bilan évolutif montrent que la voie d'amplification dépendante du centriole père remonterait à l'ancêtre des Métazoaires, avec une apparition précoce des gènes CEP63, CEP152 et de la famille E2F4/E2F5. D'un autre côté, la voie d'amplification liée au deutérosome est beaucoup plus récente, mais, contrairement à ce qui a pu être montré dans des études précédentes, ne se limite pas aux seuls Sarcoptérygiens (Tétrapodes, Cœlacanthes, Dipneustes). En effet, notre étude a révélé d'une part la présence de DEUP1, composant principale du deutérosome chez la chimère (*Callorhinchus milii*), ce qui suggère une présence chez l'ancêtre des Vertébrés. D'autre part, nos travaux ont également montré la présence de CCNO et de CCDC78, deux gènes participant à cette voie d'amplification, chez des espèces du clade Protostomiens, indiquant une apparition plus ancienne encore.

Dans l'ensemble, ces travaux ont permis d'éclaircir les mécanismes d'amplification des corps basaux indispensables à la multiciliation, mais soulèvent également de nombreuses questions qu'il reste à élucider. L'origine exacte de la voie d'amplification dépendante du deutérosome pourra notamment être précisée par l'étude de nouveaux génomes dans des groupes peu explorés tels que les Lophotrochozoaires, dans lesquels on retrouve plusieurs clades multiciliés (Mollusques, Annélides…). De la même manière, l'analyse fonctionnelle de ces espèces permettrait d'approfondir nos connaissances sur ces mécanismes ; une étude récente sur des espèces de Bryozoaires (clade appartement aux Lophotrochozoaires) a notamment mis en évidence l'existence de structures semblables aux deutérosomes chez ces espèces (Shunatova and Borisenko, 2020). Enfin, il reste un dernier point concernant l'amplification des corps basaux à éclaircir et que nous avons déjà soulevé dans le Chapitre 2, qui sont les voies d'amplification non canoniques, qui ne requièrent ni la présence de deutérosome, ni l'existence de centrioles père, et qui ont pu être observées à plusieurs reprises (Loncarek and Bettencourt-Dias, 2018).

1.2. La multiciliation chez les Vertébrés

Lors de la recherche de nouveaux gènes candidats liés à la multiciliation, nos travaux se sont majoritairement concentrés sur le clade des Vertébrés, chez qui nous avons pu montrer l'existence d'un 'locus multicilié' contenant notamment CCNO, MCIDAS, CDC20B et la famille miR-449, conservé au cours de l'évolution, et qui semble avoir été présent chez leur ancêtre commun. Nous avons par ailleurs montré qu'il existait au niveau de ce locus deux autres gènes conservés de façon similaire, GZMA et GZMK, suggérant un 'locus multicilié' plus étendu. Parmi les Vertébrés, on retrouve le groupe des Otomorpha, chez qui nous avons pu mettre en évidence un certain nombre de particularités. Outre le caractère réduit du phénotype multicilié rapporté par la littérature, nos travaux ont fait ressortir à la fois l'absence d'un gène et d'une famille de microARN clés de la multiciliation (CDC20B et miR-449a/b/c), mais également l'existence de divergences au niveau de séquences protéiques d'acteurs majeurs de la multiciliation. De plus, les Otomorpha représentent la seule instance où nous avons pu observer la perte de conservation du 'locus multicilié' au cours de notre étude. La présence de ces caractéristiques chez les Otomorpha nous a permis d'établir une signature évolutive de la multiciliation que nous avons par la suite exploitée lors de nos travaux.

Les résultats obtenus par notre analyse fonctionnelle soulèvent un point intéressant concernant la régulation de la multiciliation dans différentes espèces. En effet, nous avions mis en évidence l'existence de *clusters* 'murins' dont les gènes n'étaient surexprimés que dans des expériences menées sur la souris. La question se pose alors de savoir si la spécificité de ces clusters est en fait expliquée par des défauts de *mapping* entre les identifiants des différentes espèces et les noms de gènes humains, ou s'il existe effectivement des différences de régulation entre le xénope et la souris, voire avec l'homme. On note également un manque d'expériences de transcriptomique de type puce à ADN ou *RNAseq* dédiée à la multiciliation et réalisée sur un tissu humain. Il existe néanmoins la possibilité de créer des organoïdes de tissus multiciliés qu'il serait intéressant d'utiliser pour des analyses dédiées à la régulation de la multiciliation chez l'Homme (Kessler et al., 2015). D'autres études seraient nécessaires à la caractérisation plus précise de la régulation espèce-spécifique de la multiciliation, et dans un second temps, tissu-spécifique s'il y a lieu.

L'application d'une approche intégrative couplant génomique comparative et génomique fonctionnelle a mené à l'identification de nouveaux gènes candidats à la multiciliation que nous avons priorisés selon leur surexpression dans plusieurs expériences de transcriptomique ainsi que la présence de divergences dans leurs séquences protéiques chez les Otomorpha. Parmi les nouveaux candidats que nous avons trouvés, 5 sont particulièrement prometteurs : C1orf189, CFAP47, DYDC1, IQCK et PLAC8. En effet, en plus d'avoir été détectés par nos deux approches, ces gènes ne présentent pour la plupart aucune annotation fonctionnelle, et ont été classifiés au sein des *clusters* ciliés/multiciliés que nous avons identifiés lors de notre analyse fonctionnelle. L'ensemble de nos candidats pourra par la suite être caractérisé sur le plan phylogénétique en établissant leur distribution chez les Métazoaires pour déterminer leur histoire évolutive et, le cas échéant, de quelle voie d'amplification ils relèvent. Leur rôle au sein de la multiciliation pourra ensuite être confirmé ou infirmé par des approches expérimentales visant à inhiber leur expression afin de voir leur impact sur ce processus. S'ils s'avèrent être impliqués, il serait également intéressant de voir s'ils représentent des causes possibles de multiciliopathies en les prenant en compte lors d'études de génomique médicale sur des patients atteints de RGMC. L'ensemble des résultats de notre approche intégrative fera l'objet d'une publication actuellement en cours de préparation.

2. Vers une génomique comparative de précision

Au cours de nos travaux, nous avons employé un large panel d'approches de génomique comparative à plusieurs niveaux, disposant chacune de ses forces et de ses limitations, nous ayant mené à terme au développement d'une nouvelle approche destinée aux analyses évolutives à un niveau de granularité plus fin.

2.1. La puissance du parent pauvre de la génomique

L'application de différentes méthodes de génomique comparative nous a permis de caractériser avec plus de précision la multiciliation chez les Vertébrés, et ce en explorant différents aspects évolutifs. Notre analyse de la distribution des gènes impliqués dans la multiciliation a éclairci en partie les origines évolutives des différentes voies d'amplification des corps basaux, bien que le manque d'espèces variées dans certains clades ne nous ait pas permis d'affirmer avec certitude le moment exact de l'apparition de chacune. L'étude détaillée des familles de protéines a révélé une signature évolutive de la multiciliation chez les Otomorpha, jusqu'alors ignorée. Enfin, l'analyse du contexte génomique des gènes de la multiciliation a mis en évidence l'existence d'un 'locus multicilié' conservé au cours de l'évolution sur lequel nous nous sommes appuyés pour identifier de nouveaux candidats potentiels à la multiciliation (GZMA et GZMK), bien que ceux-ci n'aient pas été retenus par notre approche fonctionnelle. Enfin, l'application du profilage phylogénétique à la multiciliation, que nous savons puissant de par des analyses réalisées sur le cil (voir Chapitre 3 section 1.5), a fait ressortir la complexité de l'étude de la multiciliation. En effet, le profil multicilié que nous avons établi n'est pas assez atypique pour permettre de détecter avec précision des gènes multiciliés ; ceci a pu être confirmé par un enrichissement en gènes liés à d'autre processus, tels que l'immunité ou encore l'autophagie.

Nous avons par la suite recherché des gènes candidats multiciliés sur la base des divergences observées chez les Otomorpha, et l'absence de méthode adaptée à notre stratégie de recherche nous a poussés au développement d'un nouvel outil de génomique comparative. Le manque d'outils

et, de manière plus générale, le peu d'études évolutives portant sur la multiciliation et sur de nombreux autres processus, montrent qu'il existe une claire sous-exploitation de la génomique comparative et des caractéristiques évolutives des familles de gènes qui peuvent être informatives dans le cadre d'études de processus spécifiques. Il existe encore de nos jours trop de frontières entre les disciplines, au sein même de la génomique, où le domaine de la génomique comparative est bien trop souvent laissé pour compte, en particulier dans les approches intégratives.

2.2. Un nouvel outil pour des études plus fines

Les travaux de cette thèse ont mené au développement et à la mise à disposition de BLUR, un nouvel outil de génomique comparative multi-niveaux, sous forme de ressource accessible en ligne. Cet outil de profilage phylogénétique permet de prendre en compte des unités évolutives de différents niveaux : des protéines complètes, mais également des régions protéiques étendues ou des motifs composés de quelques acides aminés. Nous avons développé cette ressource dans le but de permettre l'étude de processus complexes par l'exploration des relations génotype/phénotype de façon plus subtile que le simple profilage phylogénétique basé sur la présence ou l'absence de gènes entiers. L'existence de bases de données dédiées contenant une large et diverse sélection de protéomes d'organismes appartenant aux trois domaines du Vivant confère à BLUR l'avantage de pouvoir être appliqué à des questions biologiques multiples, par la comparaison de protéomes sans *a priori* ou par l'étude des relations génotype/phénotype. Il existe néanmoins un point complexe relatif à l'utilisation de BLUR qui est le choix des organismes à comparer ainsi que l'organisme de référence à employer. Pour faciliter et encourager son utilisation, nous envisageons de mettre à disposition un set de comparaisons pré-calculées à titre d'exemple.

Malgré les avancées en termes d'outils ces dernières années auxquelles nous avons contribué par le développement de BLUR, il reste un territoire de la biologique systématiquement ignoré lors de la conception de nouvelles approches. En effet, la majeure partie du temps, l'existence de plusieurs isoformes d'une protéines est mise de côté et seules les isoformes principales sont utilisées dans les analyses. L'exploitation de ces isoformes et la prise en compte des transcrits alternatifs notamment par la génomique comparative représente un nouveau défi qui permettrait la caractérisation détaillée des mécanismes de régulation d'expression dans différents tissus, différentes conditions, ou différentes espèces et pourrait, à terme, approfondir notre connaissance du Vivant.

3. Les disciplines à haut débit et les approches intégratives

3.1. Le nouvel âge de la biologie intégrative

Les travaux de cette thèse ont abouti à la conception et à l'application d'une approche de type intégrative à l'étude de la multiciliation. Ces approches, de plus en plus employées, permettent non seulement de caractériser les processus selon plusieurs axes biologiques, mais également de faire ressortir des signaux parfois trop faibles pour être détectés de manière significative par une seule approche. L'analyse intégrative de la ciliation ayant abouti à la ressource CiliaCarta (van Dam et al., 2019) a ainsi permis l'identification de 285 nouveaux candidats ciliaires, démontrant la puissance de l'approche utilisée.

Nous avons pu apprécier pleinement la force des approches intégratives d'une part *via* la comparaison de multiples expériences de transcriptomique, pour lesquels le *clustering* a pu réunir des gènes d'expression similaire avec peu d'ambigüité, mais également par le couplage de ces résultats à un message évolutif qui a permis de faire ressortir des candidats multiciliés. En effet, bien que la multiciliation soit complexe à étudier d'un point de vue évolutif, la détermination d'une signature évolutive par des approches de génomique comparative nous a permis de prioriser les résultats de notre analyse fonctionnelle et d'orienter la séparation entre le cil et la multiciliation, ce qui aurait été beaucoup plus complexe voire impossible en se basant uniquement sur des résultats fonctionnels.

3.2. Une qualité des données insuffisante

Un problème fréquemment rencontré dans les approches multi-omiques mais également dans les approches plus classiques est la qualité des données, parfois très bruitées, pouvant être à l'origine de faux résultats positifs ou négatifs. Bien que ce problème concerne l'ensemble des disciplines à haut-débit, nous avons dans nos travaux été confrontés à des données de génomique de faible qualité, en particulier au niveau des protéomes employés pour nos diverses approches. En effet, la qualité des protéomes dans les banques de données montre un manque d'intérêt et de moyens certain pour beaucoup d'espèces dont les séquences sont fréquemment mal prédites. Ceci reflète non seulement le manque d'exploitation du pan évolutif de la biologie par des approches de génomique comparative mais également de la biodiversité disponible dont l'étude pourrait apporter de nombreuses informations. Il semble en effet que seules les espèces modèles et fréquemment utilisées pour les études médicales ou biologiques soient de bonne qualité et annotées de manière satisfaisante, tandis que certaines espèces sont systématiquement mises de côté malgré leur intérêt potentiel pour la recherche.

L'afflux des données rend difficile le contrôle de leur qualité c'est pourquoi il existe maintenant dans certains domaines des critères de qualité minimum à atteindre. Dans le cas des génomes et des protéomes par exemple, plusieurs indicateurs, tels que BUSCO (Waterhouse et al., 2018) ou DOGMA (Kemena et al., 2019) ont été développés dans le but d'évaluer la complétude des données, en particulier des protéomes. Ces indicateurs ont le désavantage de ne pas prendre en compte la qualité des annotations, ni la complétude des séquences protéiques en elle-même. Pour pallier ce manque, lors de la création de la dernière version d'OrthoInspector, un filtre a été mis en place pour la sélection des protéomes basé sur la longueur moyenne des séquences protéiques, et sur le pourcentage de séquences commençant par une méthionine. Si ce filtre constitue un progrès, il reste encore trop permissif et un nouvel indicateur prenant en considération la qualité des séquences est actuellement en cours de développement dans l'équipe, la nécessité d'avoir des données de qualité étant absolue pour la réalisation d'études de génomique comparative. Dans cette optique, un nouvel axe majeur de recherche de l'équipe se concentre sur la recherche et la correction d'erreurs de prédictions dans des protéomes, ainsi que sur l'annotation de génomes par de nouvelles approches d'intelligence artificielle, domaine qui, couplé à la biologie, va permettre de nombreuses avancées méthodologiques dans les années à venir.

Chapitre 8 : Matériel et Méthodes

1. Banques de données publiques

L'ensemble des travaux réalisés au cours de cette thèse est basé sur des données recueillies dans des bases de données publiques disponibles en ligne.

1.1. NCBI

En tant que ressource national d'information de biologie moléculaire, le *National Center for Biotechnology Information* (NCBI) héberge de nombreuses bases de données et d'outils dédiés à l'étude de la biologie. Au cours de cette thèse, trois de ces bases ont été exploitées.

1.1.1. RefSeq

La banque RefSeq (*Reference Sequence)* (O'Leary et al., 2016) est une collection de séquences nucléotidiques (ADN génomique et leurs transcrits) et protéiques visant à fournir des séquences curées, annotées et non redondantes, disponible en libre accès sur le site du NCBI. Les séquences contenues dans cette banque peuvent être prédites par un pipeline d'annotation, et portent alors des identifiants commençant par les préfixes de type 'XM_', 'XR_' et 'XP_', ou être curées et porter un identifiant commençant par le préfixe 'NM_', 'NR_' ou 'NP_'.

Les séquences issues de la banque RefSeq ont été utilisées à plusieurs reprises au cours de cette thèse, notamment lors de l'étude évolutive des gènes de la multiciliation, ainsi que lors de la création de la base de données BLUR. Les protéomes de poissons osseux n'existant pas dans la banque de protéomes de référence d'UniProt ont été récupérés dans la banque RefSeq.

1.1.2. Gene Expression Omnibus

GEO (*Gene Expression Omnibus*) est un dépôt publique de données de génomique fonctionnelle, tels que des résultats d'expérience de transcriptomique, hébergé par le NCBI. Cette banque nous a servi à récupérer les données issues d'expériences de génomique fonctionnelle liées à la multiciliation lors de notre analyse intégrative.

1.1.3. Taxonomy

La base de données *Taxonomy* du NCBI (Federhen, 2012) contient une classification standardisée et une nomenclature curée de tous les organismes existant dans les bases de données publiques de séquences. Elle permet de cette manière d'associer chaque séquence à un organisme et à l'ensemble des taxons auxquels il appartient, et ce de façon hiérarchique. Pour cela, chaque organisme et rang taxonomique est associé à un identifiant unique, le *taxonomy identifier* (*taxid*). Il est estimé que l'ensemble des espèces contenues dans la base *Taxonomy* ne représente que 10% des espèces vivantes décrites à ce jour.

La taxonomie du NCBI nous a très largement servi de base pour le développement à la fois du programme BLUR et du site web sur lequel il peut être exécuté. En effet, pour pouvoir comparer différents groupes d'organismes, les données de la base BLUR sont liées aux *taxids* issus de la base

Taxonomy. De même, la recherche de taxon d'intérêt sur le site BLUR est entièrement dépendante de la taxonomie du NCBI, dont une instance locale est disponible et tenue à jour au laboratoire. Cette base locale a été implémentée de façon à optimiser la recherche de lignage complet d'espèces et, dans cette optique, a été conçue selon le modèle *nested set*. Ce dernier permet de représenter des données hiérarchisées dans une base de données en attribuant à chaque nœud deux bornes dont les valeurs seront comprises entre celles des bornes du nœud parent, permettant ainsi d'établir les liens de parentés.

1.2. UniProt

UniProt (*Universal Protein Resource*) (The Uniprot Consortium, 2019) est une ressource issue d'une collaboration entre l'EMBL-EBI, le *Swiss Institute of Bioinformatics* (SIB) et la *Protein Information Resource* (PIR), destinée à l'étude des protéines, à la fois par le stockage de leurs séquences mais également par leur annotation manuelle et curation par des experts. Tout comme c'est le cas pour les données du NCBI, on retrouve dans la base de connaissance UniProtKB deux types de séquences : les séquences curées manuellement composent la base de donnés SwissProt, tandis que les séquences annotées de manière automatique forment la base TrEMBL. A partir de ces séquences sont formés les protéomes Uniprot, qui regroupent l'ensemble des protéines pour chaque organisme, qu'il est possible de récupérer sous forme non-redondante, avec une seule séquence protéique par gène. Certains de ces protéomes font partie des protéomes de référence, sélectionnés selon plusieurs critères pour représenter la diversité taxonomique disponible dans la base UniProtKB. Nous avons utilisé ces protéomes sous forme non-redondante pour générer la base de données BLUR. La version actuelle de BLUR contient des protéomes de référence d'Eucaryotes, de Bactéries et d'Archées téléchargés au cours du mois de novembre 2016, initialement utilisés pour générer la base d'orthologie d'OrthoInspector 3.0.

1.3. Gene Ontology et Panther

Le consortium *Gene Ontology* (The Gene Ontology Consortium, 2019) a pour but de permettre une meilleure caractérisation et compréhension des systèmes biologiques, et a développé dans cette optique une ressource permettant de décrire les connaissances actuelles sur les fonctions des gènes par un vocabulaire standardisé et hiérarchique. Ainsi, chaque gène est associé à un certain nombre de termes permettant de le caractériser selon 3 aspects : sa ou ses fonction(s) moléculaire(s), le(s) processus biologique(s) au(x)quel(s) il participe, et la localisation cellulaire des protéines correspondantes. Ces termes sont employés à la fois dans OrthoInspector et dans BLUR pour permettre d'une part une caractérisation des protéines de manière individuelle sur chaque page, ainsi qu'une caractérisation fonctionnelle globale sur un ensemble de protéines, et ce par un test d'enrichissement en termes GO.

Ces tests de surreprésentation fonctionnelle sont réalisés à l'aide des *webservices* de Panther, une ressource créée dans le but de classifier de manière fonctionnelle et évolutive les protéines pour faciliter les analyses à haut-débit. La plateforme propose de réaliser des tests d'enrichissement fonctionnel pour un ensemble de 142 espèces appartenant aux 3 domaines du Vivant. Dans le cadre des deux outils développés dans l'équipe (OrthoInspector et BLUR), une première étape est réalisée afin de vérifier de manière dynamique que l'espèce étudiée est disponible dans Panther, sans quoi la fonction d'analyse d'enrichissement est désactivée. Lorsque le

test est applicable, il est effectué à la volée sous la forme de 3 requêtes (une pour chaque ontologie GO), et les résultats sont affichés dans un nouvel onglet, par catégorie de termes GO et sous la forme d'un tableau.

1.4. STRING

La base de données STRING (Szklarczyk et al., 2019) contient des informations relatives aux liens fonctionnels et aux interactions physiques entre protéines, qu'elles soient connues ou prédites. Ces informations proviennent de différentes sources. On en retrouve 7 principales : (1) co-localisation des gènes (gènes synténiques), (2) co-occurrence dans un ensemble d'espèces, (3) évènements de fusion dans un ensemble d'espèces concernant les gènes d'intérêt, (4) co-expression dans une même espèce, (5) preuves expérimentales, (6) données issues d'autre bases de données, et (7) *text mining* de résumés de publications scientifiques. A partir de l'ensemble de ces informations, STRING calcule un score de confiance relatif à chaque interaction prédite. Nous avons utilisé cette base de données pour générer des réseaux d'interactions entre protéines sur la ressource BLUR, en n'utilisant que les interactions pour lesquelles le score de confiance était supérieur à 0.7 (*high confidence*).

1.5. Ensembl

Pour comparer le contexte génomique de différentes espèces durant l'étude évolutive de la multiciliation, nous avons utilisé la plateforme Ensembl, issue d'un projet entre l'EMBL-EBI et le Wellcome Trust Sanger Institute (Yates et al., 2020). Cette ressource regroupe à l'heure actuelle 227 génomes annotés, majoritairement vertébrés, et permet d'obtenir un grand nombre d'informations sur les gènes de chaque espèce, telles que leur localisation, les variants qu'ils présentent, s'ils sont ou non associés à des pathologies, s'ils possèdent des orthologues et des paralogues….

1.6. InterPro

InterPro (Mitchell et al., 2019) est une ressource de l'EMBL-EBI permettant l'analyse fonctionnelle de protéines, en les classifiant sous formes de familles et en inférant leurs architectures en termes de domaines. Cette ressource se base sur les informations provenant de plusieurs bases de données pour caractériser aux mieux les protéines.

Sur chaque page de protéine, BLUR intègre les informations issues de la base InterPro pour permettre une caractérisation fonctionnelle des protéines. Les données sont récupérées de manière dynamique par l'utilisation de l'interface de programmation (*API*) de l'EBI et du numéro d'accès UniProt de chaque protéine.

2. Ressources bioinformatiques

L'ensemble des travaux réalisés au cours de cette thèse s'est largement appuyé sur l'utilisation de ressources bioinformatiques existantes, la principale étant BLAST, qui a servi de point de départ pour une partie de nos développements.

2.1. BLAST

BLAST (*Basic Local Alignment Search Tool*) (Altschul et al., 1997) est un outil d'alignement local permettant la recherche de similarité entre une séquence requête fournie par l'utilisateur et une banque de séquences. Il existe plusieurs programmes BLAST selon la nature des séquences données en entrée et la banque de séquences employée. Au cours de cette thèse, nous avons majoritairement utilisé BLASTp, qui permet la comparaison de séquences protéiques, et tBLASTn, qui permet la recherche de similarité à partir d'une protéine dans une banque de séquences nucléotidiques traduites en séquences protéiques.

Lors de la recherche de séquences protéiques pour l'étude évolutive des gènes de la multiciliation, BLASTp a été utilisé en ligne sur le site du NCBI ou sur le site d'Uniprot, avec les paramètres par défaut. Pour chaque recherche, la séquence humaine a été utilisée comme requête. Pour confirmer ou infirmer l'absence de séquences homologues dans certaines espèces, tBLASTn a été utilisé sur la dernière version du génome disponible de l'organisme en question avec les paramètres par défaut et une taille de mot de 3 acides aminés.

Dans le cadre de la création de la base de données BLUR, nous avons utilisé une instance locale de BLAST+ (Camacho et al., 2009) pour faire les différentes recherches de similarité pour les organismes requêtes (espèces de référence) que nous avons sélectionnés. Trois banques BLAST ont été générées à partir des protéomes de référence UniProt, une pour chaque domaine du Vivant, Pour chaque organisme de référence, la recherche a été effectuée sur la banque à laquelle l'espèce appartient. Nous avons utilisé les paramètres par défaut, un nombre maximal de séquences détectées de 5000, et une *E-value* maximale de 1.0E-3.

2.2. Alignements multiples et arbres phylogénétiques

Durant l'étude évolutive des gènes de la multiciliation, des alignements multiples ont été générés à partir de séquences homologues préalablement collectées sur des bases de données publiques, et ce à l'aide de PipeAlign2 [http://lbgi.fr/pipealign], une nouvelle version non publiée de la cascade de programmes PipeAlign (Plewniak, 2003). Les séquences sont dans un premier temps alignées par le programme MAFFT (Katoh et al., 2002) ou par DBClustal (Thompson et al., 2000) selon le nombre de séquences à traiter. L'alignement est ensuite corrigé par le programme RASCAL (Thompson et al., 2003), et les séquences trop divergentes sont éliminées par l'outil LEON-BIS (Vanhoutreve et al., 2016). Enfin, MACSIMS (Thompson et al., 2006) permet l'annotation automatique de l'alignement sur la base des conservations et des informations sur les séquences collectées dans des bases de données publiques. Les alignements ont ensuite été visualisés et corrigés manuellement grâce aux programmes d'édition d'alignement Jalview (Waterhouse et al., 2009) et Ordali (Moulinier et al., en préparation), ce dernier étant développé au sein de l'équipe. Les arbres phylogénétiques des familles CEP63/DEUP1 et E2F4/E2F5 ont été construits selon l'approche du maximum de vraisemblance. Les positions divergentes et ambigües ont été éliminées par le programme Gblocks (Talavera and Castresana, 2007). Les arbres ont ensuite été générés par PhyML 3.0 (Guindon et al., 2010) avec les paramètres par défaut. La robustesse des branches a été évaluée par la méthode aLRT (*Approximate Likelihood Ratio Test*) (Anisimova and Gascuel, 2006).

2.3. Phyligrane

Phyligrane est un programme développé au sein de l'équipe permettant de générer, à partir d'une base de données de relations d'orthologie OrthoInspector un fichier tabulé correspondant à une matrice phylogénétique. Cette matrice contient l'ensemble des protéines de l'espèce requête en lignes, et les espèces cibles en colonnes. Pour chaque protéine, ses orthologues sont récupérés dans la base OrthoInspector et, pour chaque espèce, la présence d'un orthologue est indiquée par la présence du chiffre 1, l'absence par un 0. Dans notre cas, nous avons généré les profils pour les protéines humaines dans un ensemble de 169 espèces Métazoaires.

3. Développements informatiques et traitement de données

3.1. BLUR

3.1.1. Base de données relationnelle

Le programme BLUR s'appuie sur une base de données relationnelle *SQLite* contenant les informations relatives aux recherches de similarité BLAST préalablement calculées. Cette base de données se divise en 4 tables : (1) une table *query_info* contenant les informations relatives aux séquences des protéomes de référence, (2) une table *homology*, contenant le premier hit homologue détecté pour chaque espèce de la banque par BLAST, et ce pour l'intégralité des recherches de similarité, (3) une table *orthology*, similaire à la précédente, mais contenant les hits correspondant à des orthologues sur la base des relations prédites par OrthoInspector, et (4) une table contenant les *taxid* des espèces présentes dans la base et le nom de l'espèce correspondante.

3.1.2. Programme

Le programme BLUR a été développé en Python 3 selon une approche orientée objet, et s'appuie sur quatre classes conçues pour permettre l'interaction avec les bases de données et le traitement des résultats de BLAST (*Figure 8-1*).

Taxo_DB_Query. La première classe sur laquelle se base BLUR permet l'interaction avec notre instance locale de la base *Taxonomy*, et a pour but de récupérer les informations nécessaires à BLUR pour le traitement de la requête de l'utilisateur. Ainsi, en utilisant la librairie Python *psycopg2*, elle permet de récupérer l'ensemble des espèces appartenant au(x) taxon(s) que l'utilisateur saisi lors de sa recherche.

Blur_DB_Query. Pour récupérer l'ensemble des informations relatives aux recherches BLAST préalablement calculées, BLUR se base sur la classe *Blur_DB_Query* et utilise la librairie *sqlite3*. Cette classe a deux fonctions, la première étant la création des bases BLUR lors d'une première instanciation, la seconde étant de récupérer les hits de chaque BLAST stocké dans la base. Pour faciliter l'exécution sur les différents serveurs composant l'architecture informatique du laboratoire, nous avons également employé la librairie *sqlitebck* pour créer des copies de la base en mémoire.

Blast. La classe majeure sur laquelle s'appuie BLUR est Blast, qui crée un objet pour chaque protéine et contient l'ensemble des hits issus de la recherche BLAST. Cette classe permet notamment d'analyser les BLAST lors d'une première instanciation de BLUR, afin d'en récupérer les premiers *hits*

de chaque organisme et de les stocker dans la base de données BLUR. Lors d'une exécution de BLUR, cette classe compare les deux groupes d'intérêts sélectionnés pour chaque famille de protéine et calcule les différents critères dont BLUR se sert pour déterminer des cas de conservation différentielle, à savoir : (1) le ratio de la moyenne des valeurs logarithmiques des *E-values* entre les deux groupes, (2) la différence de distance moyenne à la requête et (3) le pourcentage d'espèces de chaque groupe ayant un meilleur rang que le premier *hit* de l'autre groupe. Pour ce faire, la classe *Blast* se base sur les calculs préalablement effectués par la classe *HitGroup*.

HitGroup. Cette dernière classe permet d'analyser les groupes d'intérêts en créant un objet contenant l'ensemble des hits des espèces appartenant à chaque groupe, et en calculant le comportement moyen du groupe dans chaque famille de protéines. Ces comportements incluent la valeur moyenne d'*E-value*, le meilleur hit du groupe, la distance moyenne à la requête, ainsi que les rangs des différents *hits* dans le BLAST.

Enfin, la fonction principale composant le programme BLUR utilise la méthode de Tukey (Tukey, 1977) pour détecter les familles de protéines *outliers* pouvant correspondre à des cas de conservation différentielle, en comparant les valeurs de chaque famille de protéines aux valeurs moyennes calculées sur l'ensemble du protéome. Elle génère ensuite les différentes listes de résultats visibles sur le site web de BLUR (haute priorité, priorité moyenne, aucun orthologue…).

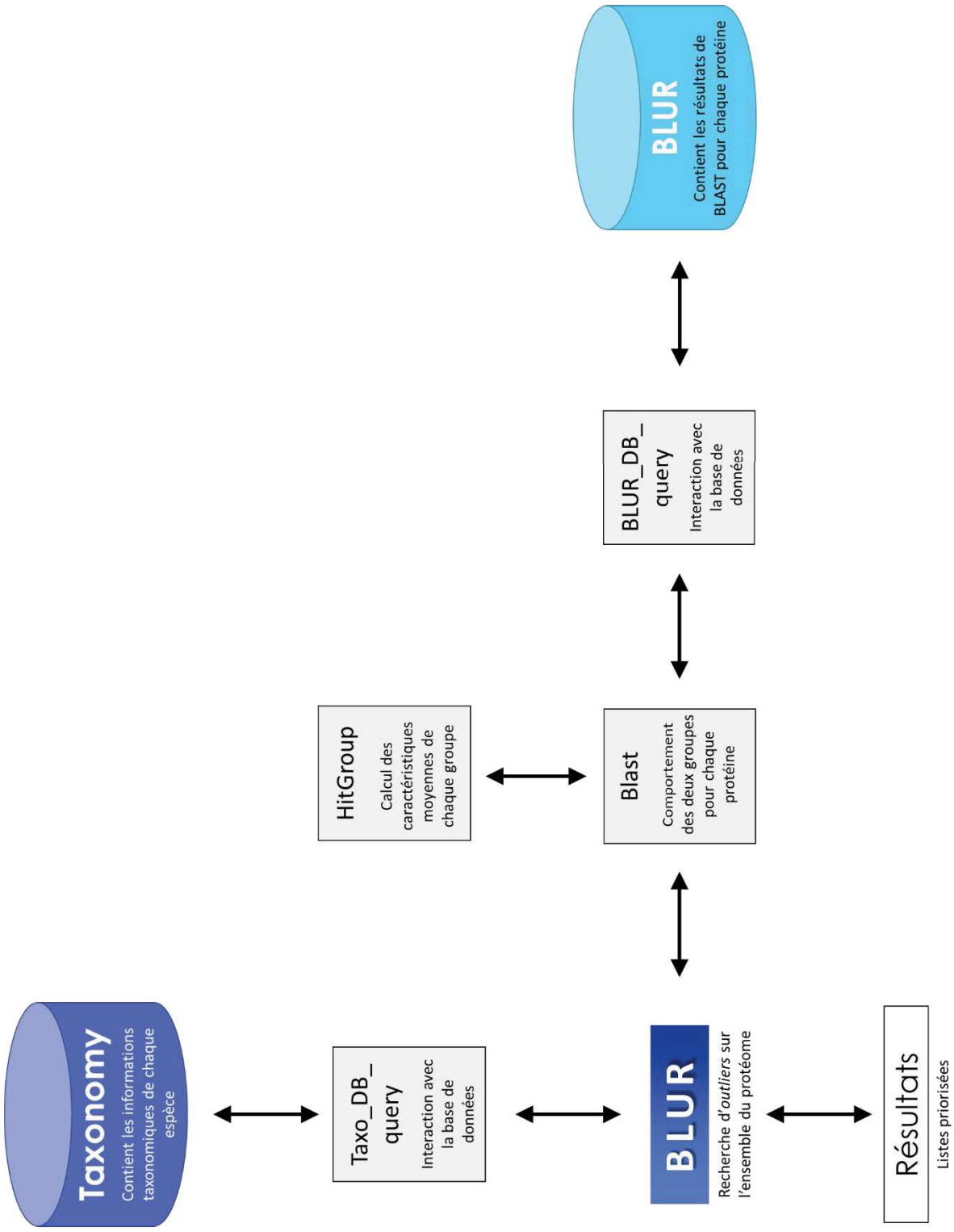

Figure 8-1: Schéma récapitulatif des classes et bases de données constitutives du programme BLUR.

3.2. Site web

Du côté serveur, le site web de BLUR est développé en PHP et repose sur le *framework Symfony* [https://symfony.com/], et le moteur de *template* Twig [https://twig.symfony.com/] qui permet de générer les pages HTML. L'interaction avec les diverses bases de données contenant les informations nécessaires à BLUR, telles que la base *Taxonomy* du NCBI, la base *Gene Ontology*, ou tout simplement la base BLUR, se fait *via* la librairie Doctrine, intégrée à Symfony.

Du côté client, le site web est basé sur des pages HTML avec un visuel géré par une feuille de style CSS personnalisée et la librairie Boostrap 3.0. L'ensemble des éléments dynamiques des différentes pages est généré par JavaScript et la librairie JQuery. Nous avons essayé de limiter au maximum l'utilisation de librairies externes pour faciliter le maintien du site web ; celui-ci se base néanmoins sur les librairies *vis.js* pour la génération des réseaux d'interactions [https://visjs.org/] et MSAViewer pour l'affichage des alignements multiples (Yachdav et al., 2016).

3.3. Comparaison de transcriptomes

3.3.1. Traitement des données brutes

Afin de pouvoir comparer les résultats issus des différentes expériences de transcriptomique, nous avons tout d'abord traité les données pour en faire ressortir les gènes surexprimés en condition multiciliée, dans un format standardisé indispensable à la comparaison.

Dans certains cas, nous avons été amenés à utiliser GEO2R, un outil analytique du NCBI qui permet de comparer des groupes d'échantillons d'expériences de génomique fonctionnelle. Celui-ci se base sur les *packages* R *GEOquery* (Davis and Meltzer, 2007) et *limma* (Smyth, 2005) du projet Bioconductor. *GEOquery* permet de structurer des données fonctionnelles en données exploitables par R, tandis que le *package limma* permet d'identifier des gènes différentiellement exprimés sur la base de tests statistiques. Pour l'ensemble des expériences nous avons sélectionné comme valeur seuil d'expression différentielle un *logFC* (*Fold Change*) de 1 et une *p-value* de 0.05.

GSE32452. Les valeurs d'expression sont disponibles sous forme brute et directement analysables par le module GEO2R du NCBI. Les réplicas de chaque expérience ont été comparés en utilisant les paramètres par défaut de GEO2R et l'ensemble des résultats a été téléchargé sous forme de fichier tabulé.

GSE59309. Les valeurs d'expression sont disponibles sous forme brute mais ne sont pas analysables par le module GEO2R. Suivant les indications des auteurs, nous avons traité les résultats en utilisant le *package* R *DESeq2* (Love et al., 2014). Les données comportent trois réplicas pour chaque condition, et trois points temporels (3h, 6h, 9h). Nous avons analysé chaque point individuellement et récupéré l'ensemble des gènes surexprimés dans les trois expériences.

GSE89271. Les résultats sont disponibles déjà traités sous forme de tableaux Excel, avec un tableau par expérience réalisée, contenant les gènes différentiellement exprimés de façon significative entre les différentes conditions, avec indications du différentiel d'expression.

GSE76342. L'expérience comporte trois points temporels (3h, 6h, 9h) et trois conditions, et les résultats sont disponibles déjà traités sous forme de tableau Excel, avec un fichier par point temporel. Chaque fichier comporte également une feuille contenant les gènes surexprimés dans l'ensemble des conditions. Nous avons récupéré ces derniers pour chaque point temporel et constitué une liste de gènes uniques surexprimés dans l'ensemble des expériences.

GSE60365. Les résultats sont disponibles sous forme brute et directement analysables par GEO2R. L'expérience comporte deux points temporels et trois réplicas pour chaque condition. Nous avons analysé chaque point temporel individuellement en utilisant les paramètres par défaut de GEO2R et récupéré l'ensemble des gènes surexprimés dans les deux expériences.

GSE75715. Les résultats sont disponibles déjà traités sous forme de tableau Excel contenant les comparaisons de valeurs d'expression entre les différentes conditions pour un ensemble de 38355 séquences murines.

GSE73331. Les valeurs d'expression sont directement analysables avec le module GEO2R. L'expérience consiste en deux conditions et 3 points temporels (*day 0, day 2, day 4*). Nous avons analysé chaque point temporel individuellement en utilisant les paramètres par défaut de GEO2R et récupéré l'ensemble des gènes surexprimés dans ces trois expériences.

GSE116690. Les résultats sont disponibles déjà traités sous forme de tableau Excel contenant les valeurs d'expression de 22556 gènes murins ainsi que les résultats de la comparaison de ces valeurs entre les deux conditions étudiées.

3.3.2. Clustering des gènes différentiellement exprimés

Pour comparer les différentes expériences mentionnées ci-dessus, nous avons été amenés à homogénéiser les identifiants de séquences, en utilisant les noms de gènes de leurs orthologues humains dans la mesure du possible. Pour les expériences menées sur la souris, nous avons employé la ressource *Mouse Genome Informatics* [http://www.informatics.jax.org/], et pour celles réalisées sur le xénope, la ressource Xenbase (Karimi et al., 2018).

Afin de pouvoir réaliser le *clustering* des résultats, nous avons ensuite généré une matrice sous forme de fichier tabulé à l'aide d'un script python. Cette matrice indique, pour chaque gène représenté par une ligne, l'existence (indiquée par 1) ou l'absence (indiquée par 0) de surexpression dans les expériences considérées, représentées par des colonnes.

Les distances entre les profils des différents gènes ont été calculées à l'aide de la fonction *Dist* du package *amap*, en utilisant la méthode '*binary*' [https://CRAN.R-project.org/package=amap]. Le *clustering* hiérarchique a ensuite été réalisé en utilisant la fonction *hclust* et la méthode '*ward.D2*'. Enfin, les *clusters* de gènes ont été définis suite à l'élagage dynamique du dendrogramme par la fonction *cutreeDynamic* du *package dynamicTreeCut* avec une profondeur de 2 et une méthode de coupe '*hybrid*' avec la matrice de distances entre les profils préalablement calculée [https://CRAN.R-project.org/package=dynamicTreeCut].

Références

Adl, S.M., Simpson, A.G.B., Lane, C.E., Lukeš, J., Bass, D., Bowser, S.S., Brown, M.W., Burki, F., Dunthorn, M., Hampl, V., et al. (2012). The Revised Classification of Eukaryotes. J. Eukaryot. Microbiol. *59*, 429–514.

Al Jord, A., Lemaître, A.-I., Delgehyr, N., Faucourt, M., Spassky, N., and Meunier, A. (2014). Centriole amplification by mother and daughter centrioles differs in multiciliated cells. Nature *516*, 104–107.

Al Jord, A., Shihavuddin, A., d'Aout, R.S., Faucourt, M., Genovesio, A., Karaiskou, A., Sobczak-Thépot, J., Spassky, N., and Meunier, A. (2017). Calibrated mitotic oscillator drives motile ciliogenesis. Science *358*, 803–806.

Albee, A.J., Kwan, A.L., Lin, H., Granas, D., Stormo, G.D., and Dutcher, S.K. (2013). Identification of Cilia Genes That Affect Cell-Cycle Progression Using Whole-Genome Transcriptome Analysis in *Chlamydomonas reinhardtti*. G3amp58 GenesGenomesGenetics *3*, 979–991.

Allen, R.D. (1969). The morphogenesis of basal bodies and accessory structures of the cortex of the ciliated protozoan tetrahymena pyriformis. J. Cell Biol. *40*, 716–733.

Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. *25*, 3389–3402.

Amirav, I., Wallmeier, J., Loges, N.T., Menchen, T., Pennekamp, P., Mussaffi, H., Abitbul, R., Avital, A., Bentur, L., Dougherty, G.W., et al. (2016). Systematic Analysis of *CCNO* Variants in a Defined Population: Implications for Clinical Phenotype and Differential Diagnosis. Hum. Mutat. *37*, 396–405.

Anisimova, M., and Gascuel, O. (2006). Approximate likelihood-ratio test for branches: A fast, accurate, and powerful alternative. Syst. Biol. *55*, 539–552.

Arnaiz, O., Cohen, J., Tassin, A.-M., and Koll, F. (2014). Remodeling Cildb, a popular database for cilia and links for ciliopathies. Cilia *3*, 9.

Auton, A., Abecasis, G.R., Altshuler, D.M., Durbin, R.M., Abecasis, G.R., Bentley, D.R., Chakravarti, A., Clark, A.G., Donnelly, P., Eichler, E.E., et al. (2015). A global reference for human genetic variation. Nature *526*, 68–74.

Bachmann-Gagescu, R. (2014). Complexité génétique des ciliopathies et identification de nouveaux gènes. médecine/sciences *30*, 1011–1023.

Barlocco, E.G., Valletta, E.A., Canciani, M., Lungarella, G., Gardi, C., Margherita De Santi, M., and Mastella, G. (1991). Ultrastructural ciliary defects in children with recurrent infections of the lower respiratory tract. Pediatr. Pulmonol. *10*, 11–17.

Bartlett, G.J., Borkakoti, N., and Thornton, J.M. (2003). Catalysing New Reactions during Evolution: Economy of Residues and Mechanism. J. Mol. Biol. *331*, 829–860.

Bayless, B.A., Navarro, F.M., and Winey, M. (2019). Motile Cilia: Innovation and Insight From Ciliate Model Organisms. Front. Cell Dev. Biol. *7*.

Berlucchi, M., de Santi, M.M., Bertoni, E., Spinelli, E., Timpano, S., and Padoan, R. (2012). Ciliary aplasia associated with hydrocephalus: an extremely rare occurrence. Eur. Arch. Otorhinolaryngol. *269*, 2295–2299.

Betancur-R, R., Broughton, R.E., Wiley, E.O., Carpenter, K., López, J.A., Li, C., Holcroft, N.I., Arcila, D., Sanciangco, M., Cureton Ii, J.C., et al. (2013). The tree of life and a new classification of bony fishes. PLoS Curr. *5*.

Blackburn, K., Bustamante-Marin, X., Yin, W., Goshe, M.B., and Ostrowski, L.E. (2017). Quantitative Proteomic Analysis of Human Airway Cilia Identifies Previously Uncharacterized Proteins of High Abundance. J. Proteome Res. *16*, 1579–1592.

Blacque, O.E., Perens, E.A., Boroevich, K.A., Inglis, P.N., Li, C., Warner, A., Khattra, J., Holt, R.A., Ou, G., Mah, A.K., et al. (2005). Functional Genomics of the Cilium, a Sensory Organelle. Curr. Biol. *15*, 935–941.

Bokhovchuk, F., Mesrouze, Y., Delaunay, C., Martin, T., Villard, F., Meyerhofer, M., Fontana, P., Zimmermann, C., Erdmann, D., Furet, P., et al. (2020). Identification of FAM181A and FAM181B as new interactors with the TEAD transcription factors. Protein Sci. *29*, 509–520.

Boon, M., Wallmeier, J., Ma, L., Loges, N.T., Jaspers, M., Olbrich, H., Dougherty, G.W., Raidt, J., Werner, C., Amirav, I., et al. (2014). MCIDAS mutations result in a mucociliary clearance disorder with reduced generation of multiple motile cilia. Nat. Commun. *5*, 4418.

Bornens, M. (2018). Cell polarity: having and making sense of direction—on the evolutionary significance of the primary cilium/centrosome organ in Metazoa. Open Biol. *8*, 180052.

Boury-Esnault, N., Efremova, S., Bézac, C., and Vacelet, J. (1999). Reproduction of a hexactinellid sponge: first description of gastrulation by cellular delamination in the Porifera. Invertebr. Reprod. Dev. *35*, 187–201.

Braun, D.A., Schueler, M., Halbritter, J., Gee, H.Y., Porath, J.D., Lawson, J.A., Airik, R., Shril, S., Allen, S.J., Stein, D., et al. (2016). Whole exome sequencing identifies causative mutations in the majority of consanguineous or familial cases with childhood-onset increased renal echogenicity. Kidney Int. *89*, 468–475.

Brent, M.R. (2008). Steady progress and recent breakthroughs in the accuracy of automated genome annotation. Nat. Rev. Genet. *9*, 62–73.

Brooks, E.R., and Wallingford, J.B. (2014). Multiciliated Cells. Curr. Biol. *24*, R973–R982.

Buljan, M., and Bateman, A. (2009). The evolution of protein domain families. Biochem. Soc. Trans. *37*, 751–755.

Bustamante-Marin, X.M., and Ostrowski, L.E. (2017). Cilia and Mucociliary Clearance. Cold Spring Harb. Perspect. Biol. *9*.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., and Madden, T.L. (2009). BLAST+: architecture and applications. BMC Bioinformatics *10*, 421.

Campbell, E.P., Quigley, I.K., and Kintner, C. (2016). Foxn4 promotes gene expression required for the formation of multiple motile cilia. Dev. Camb. Engl. *143*, 4654–4664.

Castro-Sánchez, S., Álvarez-Satta, M., Tohamy, M.A., Beltran, S., Derdak, S., and Valverde, D. (2017). Whole exome sequencing as a diagnostic tool for patients with ciliopathy-like phenotypes. PLOS ONE *12*, e0183081.

Chevalier, B., Adamiok, A., Mercey, O., Revinski, D.R., Zaragosi, L.-E., Pasini, A., Kodjabachian, L., Barbry, P., and Marcet, B. (2015). miR-34/449 control apical actin network formation during multiciliogenesis through small GTPase pathways. Nat. Commun. *6*, 8386.

Chivukula, R.R., Montoro, D.T., Leung, H.M., Yang, J., Shamseldin, H.E., Taylor, M.S., Dougherty, G.W., Zariwala, M.A., Carson, J., Daniels, M.L.A., et al. (2020). A human ciliopathy reveals essential functions for NEK10 in airway mucociliary clearance. Nat. Med. *26*, 244–251.

Cho, R.J., Campbell, M.J., Winzeler, E.A., Steinmetz, L., Conway, A., Wodicka, L., Wolfsberg, T.G., Gabrielian, A.E., Landsman, D., Lockhart, D.J., et al. (1998). A Genome-Wide Transcriptional Analysis of the Mitotic Cell Cycle. Mol. Cell *2*, 65–73.

Chu, Q., Yao, C., Qi, X., Stripp, B.R., and Tang, N. (2019). STK11 is required for the normal program of ciliated cell differentiation in airways. Cell Discov. *5*, 1–16.

Chung, M.-I., Kwon, T., Tu, F., Brooks, E.R., Gupta, R., Meyer, M., Baker, J.C., Marcotte, E.M., and Wallingford, J.B. (2014). Coordinated genomic control of ciliogenesis and cell movement by RFX2. ELife *3*.

Cromar, G.L., Zhao, A., Xiong, X., Swapna, L.S., Loughran, N., Song, H., and Parkinson, J. (2016). PhyloPro2.0: a database for the dynamic exploration of phylogenetically conserved proteins and their domain architectures across the Eukarya. Database *2016*, baw013.

Curtis, C., Shah, S.P., Chin, S.-F., Turashvili, G., Rueda, O.M., Dunning, M.J., Speed, D., Lynch, A.G., Samarajiwa, S., Yuan, Y., et al. (2012). The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. Nature *486*, 346–352.

van Dam, T.J.P., Kennedy, J., Lee, R. van der, Vrieze, E. de, Wunderlich, K.A., Rix, S., Dougherty, G.W., Lambacher, N.J., Li, C., Jensen, V.L., et al. (2019). CiliaCarta: An integrated and validated compendium of ciliary genes. PLOS ONE *14*, e0216705.

Davis, S., and Meltzer, P.S. (2007). GEOquery: a bridge between the Gene Expression Omnibus (GEO) and BioConductor. Bioinforma. Oxf. Engl. *23*, 1846–1847.

Dey, G., Jaimovich, A., Collins, S.R., Seki, A., and Meyer, T. (2015). Systematic discovery of human gene function and principles of modular organization through phylogenetic profiling. Cell Rep. *10*, 993–1006.

Didon, L., Zwick, R.K., Chao, I.W., Walters, M.S., Wang, R., Hackett, N.R., and Crystal, R.G. (2013). RFX3 Modulation of FOXJ1 regulation of cilia genes in the human airway epithelium. Respir. Res. *14*, 70.

Dohmen, E., Klasberg, S., Bornberg-Bauer, E., Perrey, S., and Kemena, C. (2020). The modular nature of protein evolution: domain rearrangement rates across eukaryotic life. BMC Evol. Biol. *20*.

Eddy, S., Mariani, L.H., and Kretzler, M. (2020). Integrated multi-omics approaches to improve classification of chronic kidney disease. Nat. Rev. Nephrol. 1–12.

El-Gebali, S., Mistry, J., Bateman, A., Eddy, S.R., Luciani, A., Potter, S.C., Qureshi, M., Richardson, L.J., Salazar, G.A., Smart, A., et al. (2019). The Pfam protein families database in 2019. Nucleic Acids Res. *47*, D427–D432.

Elgeti, J., and Gompper, G. (2013). Emergence of metachronal waves in cilia arrays. Proc. Natl. Acad. Sci. *110*, 4470–4475.

El Shamieh, S., Neuillé, M., Terray, A., Orhan, E., Condroyer, C., Démontant, V., Michiels, C., Antonio, A., Boyard, F., Lancelot, M.-E., et al. (2014). Whole-Exome Sequencing Identifies KIZ as a Ciliary Gene Associated with Autosomal-Recessive Rod-Cone Dystrophy. Am. J. Hum. Genet. *94*, 625–633.

Epting, D., Slanchev, K., Boehlke, C., Hoff, S., Loges, N.T., Yasunaga, T., Indorf, L., Nestel, S., Lienkamp, S.S., Omran, H., et al. (2015). The Rac1 regulator ELMO controls basal body migration and docking in multiciliated cells through interaction with Ezrin. Development *142*, 174–184.

Fahy, J.V., and Dickey, B.F. (2010). Airway Mucus Function and Dysfunction. N. Engl. J. Med. *363*, 2233–2247.

Falk, N., Lösl, M., Schröder, N., and Gießl, A. (2015). Specialized Cilia in Mammalian Sensory Systems. Cells *4*, 500–519.

Federhen, S. (2012). The NCBI Taxonomy database. Nucleic Acids Res. *40*, D136–D143.

Fitch, W.M. (1970). Distinguishing Homologous from Analogous Proteins. Syst. Biol. *19*, 99–113.

Forslund, K., Pekkari, I., and Sonnhammer, E.L. (2011). Domain architecture conservation in orthologs. BMC Bioinformatics *12*, 326.

Frayling, T. (2014). Genome-wide association studies: the good, the bad and the ugly. Clin. Med. *14*, 428–431.

Friedman, N., and Rando, O.J. (2015). Epigenomics and the structure of the living genome. Genome Res. *25*, 1482–1490.

Funk, M.C., Bera, A.N., Menchen, T., Kuales, G., Thriene, K., Lienkamp, S.S., Dengjel, J., Omran, H., Frank, M., and Arnold, S.J. (2015). Cyclin O (Ccno) functions during deuterosome-mediated centriole amplification of multiciliated cells. EMBO J. *34*, 1078–1089.

Garcia, G., Raleigh, D.R., and Reiter, J.F. (2018). How the Ciliary Membrane Is Organized Inside-Out to Communicate Outside-In. Curr. Biol. *28*, R421–R434.

Ghosh, A., Syed, S.M., and Tanwar, P.S. (2017). *In vivo* genetic cell lineage tracing reveals that oviductal secretory cells self-renew and give rise to ciliated cells. Development *144*, 3031–3041.

Giribet, G. (2016). New animal phylogeny: future challenges for animal phylogeny in the age of phylogenomics. Org. Divers. Evol. *16*, 419–426.

Guigó, R., Flicek, P., Abril, J.F., Reymond, A., Lagarde, J., Denoeud, F., Antonarakis, S., Ashburner, M., Bajic, V.B., Birney, E., et al. (2006). EGASP: the human ENCODE Genome Annotation Assessment Project. Genome Biol. 31.

Guindon, S., Dufayard, J.-F., Lefort, V., Anisimova, M., Hordijk, W., and Gascuel, O. (2010). New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0. Syst. Biol. *59*, 307–321.

Halbert, S.A., Patton, D.L., Zarutskie, P.W., and Soules, M.R. (1997). Function and structure of cilia in the fallopian tube of an infertile woman with Kartagener's syndrome. Hum. Reprod. *12*, 55–58.

Han, X. (2016). Lipidomics for studying metabolism. Nat. Rev. Endocrinol. *12*, 668–679.

Han, X., Guo, J., Pang, E., Song, H., and Lin, K. (2020). Ab Initio Construction and Evolutionary Analysis of Protein-Coding Gene Families with Partially Homologous Relationships: Closely Related Drosophila Genomes as a Case Study. Genome Biol. Evol. *12*, 185–202.

Hansen, A., and Zeiske, E. (1993). Development of the olfactory organ in the zebrafish,Brachydanio rerio. J. Comp. Neurol. *333*, 289–300.

Hansen, A., and Zielinski, B.S. (2005). Diversity in the olfactory epithelium of bony fishes: development, lamellar arrangement, sensory neuron cell types and transduction components. J. Neurocytol. *34*, 183–208.

Harrow, J., Nagy, A., Reymond, A., Alioto, T., Patthy, L., Antonarakis, S.E., and Guigó, R. (2009). Identifying protein-coding genes in genomic sequences. Genome Biol. *10*, 201.

Henrique, D., and Schweisguth, F. (2019). Mechanisms of Notch signaling: a simple logic deployed in time and space. Development *146*, dev172148.

Herawati, E., Taniguchi, D., Kanoh, H., Tateishi, K., Ishihara, S., and Tsukita, S. (2016). Multiciliated cell basal bodies align in stereotypical patterns coordinated by the apical cytoskeleton. J. Cell Biol. *214*, 571–586.

Hodges, M.E., Wickstead, B., Gull, K., and Langdale, J.A. (2012). The evolution of land plant cilia. New Phytol. *195*, 526–540.

Horani, A., Druley, T.E., Zariwala, M.A., Patel, A.C., Levinson, B.T., Van Arendonk, L.G., Thornton, K.C., Giacalone, J.C., Albee, A.J., Wilson, K.S., et al. (2012). Whole-Exome Capture and Sequencing Identifies HEATR2 Mutation as a Cause of Primary Ciliary Dyskinesia. Am. J. Hum. Genet. *91*, 685–693.

Hoyer-Fender, S. (2013). Primary and Motile Cilia: Their Ultrastructure and Ciliogenesis. In Cilia and Nervous System Development and Function, K.L. Tucker, and T. Caspary, eds. (Dordrecht: Springer Netherlands), pp. 1–53.

Hurst, L.D., Pál, C., and Lercher, M.J. (2004). The evolutionary dynamics of eukaryotic gene order. Nat. Rev. Genet. *5*, 299–310.

Ishikawa, H., and Marshall, W.F. (2017). Intraflagellar Transport and Ciliary Dynamics. Cold Spring Harb. Perspect. Biol. *9*, a021998.

Ishikawa, H., Thompson, J., Yates, J.R., and Marshall, W.F. (2012). Proteomic Analysis of Mammalian Primary Cilia. Curr. Biol. *22*, 414–419.

Jean, Y., Langlais, J., Roberts, K.D., Chapdelaine, A., and Bleau, G. (1979). Fertility of a Woman with Nonfunctional Ciliated cells in the Fallopian Tubes. Fertil. Steril. *31*, 349–350.

Karczewski, K.J., Francioli, L.C., Tiao, G., Cummings, B.B., Alföldi, J., Wang, Q., Collins, R.L., Laricchia, K.M., Ganna, A., Birnbaum, D.P., et al. (2020). The mutational constraint spectrum quantified from variation in 141,456 humans. Nature *581*, 434–443.

Karimi, K., Fortriede, J.D., Lotay, V.S., Burns, K.A., Wang, D.Z., Fisher, M.E., Pells, T.J., James-Zorn, C., Wang, Y., Ponferrada, V.G., et al. (2018). Xenbase: a genomic, epigenomic and transcriptomic model organism database. Nucleic Acids Res. *46*, D861–D868.

Katoh, K., Misawa, K., Kuma, K., and Miyata, T. (2002). MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res. *30*, 3059–3066.

Katz, S.M., and Morgan, J.J. (1984). Cilia in the Human Kidney. Ultrastruct. Pathol. *6*, 285–294.

Kemena, C., Dohmen, E., and Bornberg-Bauer, E. (2019). DOGMA: a web server for proteome and transcriptome quality assessment. Nucleic Acids Res. *47*, W507–W510.

Kessler, M., Hoffmann, K., Brinkmann, V., Thieck, O., Jackisch, S., Toelle, B., Berger, H., Mollenkopf, H.-J., Mangler, M., Sehouli, J., et al. (2015). The Notch and Wnt pathways regulate stemness and differentiation in human fallopian tube organoids. Nat. Commun. *6*, 8989.

Khan, S., and Scholey, J.M. (2018). Assembly, Functions and Evolution of Archaella, Flagella and Cilia. Curr. Biol. *28*, R278–R292.

Khasawneh, A.H., Garling, R.J., and Harris, C.A. (2018). Cerebrospinal fluid circulation: What do we know and how do we know it? Brain Circ. *4*, 14–18.

Kim, D.-H., Kim, Y.-S., Son, N.-I., Kang, C.-K., and Kim, A.-R. (2017). Recent omics technologies and their emerging applications for personalised medicine. IET Syst. Biol. *11*, 87–98.

Klos Dehring, D.A., Vladar, E.K., Werner, M.E., Mitchell, J.W., Hwang, P., and Mitchell, B.J. (2013). Deuterosome-Mediated Centriole Biogenesis. Dev. Cell *27*, 103–112.

Knowles, M.R., Daniels, L.A., Davis, S.D., Zariwala, M.A., and Leigh, M.W. (2013a). Primary Ciliary Dyskinesia. Recent Advances in Diagnostics, Genetics, and Characterization of Clinical Disease. Am. J. Respir. Crit. Care Med. *188*, 913–922.

Knowles, M.R., Leigh, M.W., Ostrowski, L.E., Huang, L., Carson, J.L., Hazucha, M.J., Yin, W., Berg, J.S., Davis, S.D., Dell, S.D., et al. (2013b). Exome Sequencing Identifies Mutations in CCDC114 as a Cause of Primary Ciliary Dyskinesia. Am. J. Hum. Genet. *92*, 99–106.

Koonin, E.V. (2005). Orthologs, Paralogs, and Evolutionary Genomics. Annu. Rev. Genet. *39*, 309–338.

Koonin, E.V., Fedorova, N.D., Jackson, J.D., Jacobs, A.R., Krylov, D.M., Makarova, K.S., Mazumder, R., Mekhedov, S.L., Nikolskaya, A.N., Rao, B.S., et al. (2004). A comprehensive evolutionary classification of proteins encoded in complete eukaryotic genomes. Genome Biol. *5*, R7.

Kramer-Zucker, A.G., Olale, F., Haycraft, C.J., Yoder, B.K., Schier, A.F., and Drummond, I.A. (2005). Cilia-driven fluid flow in the zebrafish pronephros, brain and Kupffer's vesicle is required for normal organogenesis. Development *132*, 1907–1921.

Kress, A., Lecompte, O., Poch, O., and Thompson, J.D. (2018). PROBE: analysis and visualization of protein block-level evolution. Bioinformatics *34*, 3390–3392.

Kumar, K.R., Cowley, M.J., and Davis, R.L. (2019). Next-Generation Sequencing and Emerging Technologies. Semin. Thromb. Hemost. *45*, 661–673.

Kyrousi, C., Arbi, M., Pilz, G.-A., Pefani, D.-E., Lalioti, M.-E., Ninkovic, J., Go tz, M., Lygerou, Z., and Taraviras, S. (2015). Mcidas and GemC1 are key regulators for the generation of multiciliated ependymal cells in the adult neurogenic niche. Development *142*, 3661–3674.

Labrecque, J., Dumas, F., Lacroix, A., and Bhat, P.V. (1995). A novel isoenzyme of aldehyde dehydrogenase specifically involved in the biosynthesis of 9-cis and all-trans retinoic acid. Biochem. J. *305*, 681–684.

Lee, J.M., and Sonnhammer, E.L.L. (2003). Genomic Gene Clustering Analysis of Pathways in Eukaryotes. Genome Res. *13*, 875–882.

Lees, J.G., Dawson, N.L., Sillitoe, I., and Orengo, C.A. (2016). Functional innovation from changes in protein domains and their combinations. Curr. Opin. Struct. Biol. *38*, 44–52.

Lemullois, M., Boisvieux-Ulrich, E., Laine, M.-C., Chailley, B., and Sandoz, D. (1988). Development and functions of the cytoskeleton during ciliogenesis in metazoa. Biol. Cell *63*, 195–208.

Leon-Mimila, P., Wang, J., and Huertas-Vazquez, A. (2019). Relevance of Multi-Omics Studies in Cardiovascular Diseases. Front. Cardiovasc. Med. *6*.

Li, J.B., Gerdes, J.M., Haycraft, C.J., Fan, Y., Teslovich, T.M., May-Simera, H., Li, H., Blacque, O.E., Li, L., Leitch, C.C., et al. (2004). Comparative Genomics Identifies a Flagellar and Basal Body Proteome that Includes the BBS5 Human Disease Gene. Cell *117*, 541–552.

Li, Y., Calvo, S.E., Gutman, R., Liu, J.S., and Mootha, V.K. (2014). Expansion of Biological Pathways Based on Evolutionary Inference. Cell *158*, 213–225.

Linard, B., Thompson, J.D., Poch, O., and Lecompte, O. (2011). OrthoInspector: comprehensive orthology analysis and visual exploration. BMC Bioinformatics *12*, 11.

Lindemann, C.B., and Lesich, K.A. (2016). Functional anatomy of the mammalian sperm flagellum. Cytoskeleton *73*, 652–669.

Liu, Q., Tan, G., Levenkova, N., Li, T., Pugh, E.N., Rux, J.J., Speicher, D.W., and Pierce, E.A. (2007a). The Proteome of the Mouse Photoreceptor Sensory Cilium Complex. Mol. Cell. Proteomics MCP *6*, 1299–1317.

Liu, Y., Pathak, N., Kramer-Zucker, A., and Drummond, I.A. (2007b). Notch signaling controls the differentiation of transporting epithelia and multiciliated cells in the zebrafish pronephros. Development *134*, 1111–1122.

Löhr, U., Yussa, M., and Pick, L. (2001). Drosophila fushi tarazu: a gene on the border of homeotic function. Curr. Biol. *11*, 1403–1412.

Loncarek, J., and Bettencourt-Dias, M. (2018). Building the right centriole for each cell type. J. Cell Biol. *217*, 823–835.

Long, H., and Huang, K. (2020). Transport of Ciliary Membrane Proteins. Front. Cell Dev. Biol. *7*.

Love, M.I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. *15*, 550.

Lowe, R., Shirley, N., Bleackley, M., Dolan, S., and Shafee, T. (2017). Transcriptomics technologies. PLOS Comput. Biol. *13*, e1005457.

Lu, H., Anujan, P., Zhou, F., Zhang, Y., Chong, Y.L., Bingle, C.D., and Roy, S. (2019). *Mcidas* mutant mice reveal a two-step process for the specification and differentiation of multiciliated cells in mammals. Development *146*, dev172643.

Lv, J., Zhang, Z., Pan, L., and Zhang, Y. (2019). MicroRNA-34/449 family and viral infections. Virus Res. *260*, 1–6.

Ma, L., Quigley, I., Omran, H., and Kintner, C. (2014). Multicilin drives centriole biogenesis via E2f proteins. Genes Dev. *28*, 1461–1471.

Machemer, H. (1972). Ciliary Activity and the Origin of Metachrony in Paramecium: Effects of Increased Viscosity. J. Exp. Biol. *57*, 239–259.

Marcet, B., Chevalier, B., Luxardi, G., Coraux, C., Zaragosi, L.-E., Cibois, M., Robbe-Sermesant, K., Jolly, T., Cardinaud, B., Moreilhon, C., et al. (2011). Control of vertebrate multiciliogenesis by miR-449 through direct repression of the Delta/Notch pathway. Nat. Cell Biol. *13*, 694–701.

Marra, A.N., Cheng, C.N., Adeeb, B., Addiego, A., Wesselman, H.M., Chambers, B.E., Chambers, J.M., and Wingert, R.A. (2019). Iroquois transcription factor irx2a is required for multiciliated and transporter cell fate decisions during zebrafish pronephros development. Sci. Rep. *9*.

Marshall, C.B., Mays, D.J., Beeler, J.S., Rosenbluth, J.M., Boyd, K.L., Santos Guasch, G.L., Shaver, T.M., Tang, L.J., Liu, Q., Shyr, Y., et al. (2016). p73 Is Required for Multiciliogenesis and Regulates the Foxj1-Associated Gene Network. Cell Rep. *14*, 2289–2300.

Matwijiw, I., Thliveris, J.A., and Faiman, C. (1987). Aplasia of Nasal Cilia with Situs Inversus, Azoospermia and Normal Sperm Flagella: A Unique Variant of the Immotile Cilia Syndrome. J. Urol. *137*, 522–524.

McClintock, T.S., Glasser, C.E., Bose, S.C., and Bergman, D.A. (2008). Tissue expression patterns identify mouse cilia genes. Physiol. Genomics *32*, 198–206.

McKusick, V.A., and Ruddle, F.H. (1987). A new discipline, a new name, a new journal. Genomics *1*, 1–2.

Mercey, O., Levine, M.S., LoMastro, G.M., Rostaing, P., Brotslaw, E., Gomez, V., Kumar, A., Spassky, N., Mitchell, B.J., Meunier, A., et al. (2019). Massive centriole production can occur in the absence of deuterosomes in multiciliated cells. Nat. Cell Biol. *21*, 1544–1552.

Metzker, M.L. (2010). Sequencing technologies — the next generation. Nat. Rev. Genet. *11*, 31–46.

Mi, H., Muruganujan, A., Ebert, D., Huang, X., and Thomas, P.D. (2019). PANTHER version 14: more genomes, a new PANTHER GO-slim and improvements in enrichment analysis tools. Nucleic Acids Res. *47*, D419–D426.

Mitchell, D.R. (2017). Evolution of Cilia. Cold Spring Harb. Perspect. Biol. *9*.

Mitchell, A.L., Attwood, T.K., Babbitt, P.C., Blum, M., Bork, P., Bridge, A., Brown, S.D., Chang, H.-Y., El-Gebali, S., Fraser, M.I., et al. (2019). InterPro in 2019: improving coverage, classification and access to protein sequence annotations. Nucleic Acids Res. *47*, D351–D360.

Mizukami, I., and Gall, J. (1966). Centriole replication. II. Sperm Formation in the Fern, Marsilea, and the Cycad, Zamia. J. Cell Biol. *29*, 97–111.

Moore, A.D., and Bornberg-Bauer, E. (2012). The Dynamics and Evolutionary Potential of Domain Loss and Emergence. Mol. Biol. Evol. *29*, 787–796.

Moore, A.D., Grath, S., Schüler, A., Huylmans, A.K., and Bornberg-Bauer, E. (2013). Quantification and functional analysis of modular protein evolution in a dense phylogenetic tree. Biochim. Biophys. Acta BBA - Proteins Proteomics *1834*, 898–907.

Mori, M., Hazan, R., Danielian, P.S., Mahoney, J.E., Li, H., Lu, J., Miller, E.S., Zhu, X., Lees, J.A., and Cardoso, W.V. (2017). Cytoplasmic E2f4 forms organizing centres for initiation of centriole amplification during multiciliogenesis. Nat. Commun. *8*.

Mukherjee, S., Stamatis, D., Bertsch, J., Ovchinnikova, G., Katta, H.Y., Mojica, A., Chen, I.-M.A., Kyrpides, N.C., and Reddy, T. (2019). Genomes OnLine database (GOLD) v.7: updates and new features. Nucleic Acids Res. *47*, D649–D659.

Munkholm, M., and Mortensen, J. (2014). Mucociliary clearance: pathophysiological aspects. Clin. Physiol. Funct. Imaging *34*, 171–177.

Nanjundappa, R., Kong, D., Shim, K., Stearns, T., Brody, S.L., Loncarek, J., and Mahjoub, M.R. (2019). Regulation of cilia abundance in multiciliated cells. ELife *8*, e44039.

Nemajerova, A., Kramer, D., Siller, S.S., Herr, C., Shomroni, O., Pena, T., Gallinas Suazo, C., Glaser, K., Wildung, M., Steffen, H., et al. (2016). TAp73 is a central transcriptional regulator of airway multiciliogenesis. Genes Dev. *30*, 1300–1312.

Nevers, Y., Defosset, A., and Lecompte, O. (In press). Orthology: promises and challenges. In Evolutionary Biology: Transdisciplinary Approach., P. Pontarotti, ed. (Springer International Publishing), p.

Nevers, Y., Prasad, M.K., Poidevin, L., Chennen, K., Allot, A., Kress, A., Ripp, R., Thompson, J.D., Dollfus, H., Poch, O., et al. (2017). Insights into Ciliary Genes and Evolution from Multi-Level Phylogenetic Profiling. Mol. Biol. Evol. *34*, 2016–2034.

Nevers, Y., Kress, A., Defosset, A., Ripp, R., Linard, B., Thompson, J.D., Poch, O., and Lecompte, O. (2019). OrthoInspector 3.0: open portal for comparative genomics. Nucleic Acids Res. *47*, D411–D418.

Nielsen, C. (2012). Animal evolution: interrelationships of the living phyla (Oxford ; New York: Oxford University Press).

Nikolaev, S.I. (2006). Phylogenetic position of Multicilia marina and the evolution of Amoebozoa. Int. J. Syst. Evol. Microbiol. *56*, 1449–1458.

Obernier, K., and Alvarez-Buylla, A. (2019). Neural stem cells: origin, heterogeneity and regulation in the adult mammalian brain. Development *146*, dev156059.

O'Leary, N.A., Wright, M.W., Brister, J.R., Ciufo, S., Haddad, D., McVeigh, R., Rajput, B., Robbertse, B., Smith-White, B., Ako-Adjei, D., et al. (2016). Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res. *44*, D733-745.

Onoufriadis, A., Shoemark, A., Munye, M.M., James, C.T., Schmidts, M., Patel, M., Rosser, E.M., Bacchelli, C., Beales, P.L., Scambler, P.J., et al. (2014). Combined exome and whole-genome sequencing identifies mutations in *ARMC4* as a cause of primary ciliary dyskinesia with defects in the outer dynein arm. J. Med. Genet. *51*, 61–67.

Pala, R., Alomari, N., and Nauli, S.M. (2017). Primary Cilium-Dependent Signaling Mechanisms. Int. J. Mol. Sci. *18*, 2272.

Pan, J., You, Y., Huang, T., and Brody, S.L. (2007). RhoA-mediated apical actin enrichment is required for ciliogenesis and promoted by Foxj1. J. Cell Sci. *120*, 1868–1876.

Pan, J., Adair-Kirk, T.L., Patel, A.C., Huang, T., Yozamp, N.S., Xu, J., Reddy, E.P., Byers, D.E., Pierce, R.A., Holtzman, M.J., et al. (2014). Myb permits multilineage airway epithelial cell differentiation. Stem Cells Dayt. Ohio *32*, 3245–3256.

Pefani, D.-E., Dimaki, M., Spella, M., Karantzelis, N., Mitsiki, E., Kyrousi, C., Symeonidou, I.-E., Perrakis, A., Taraviras, S., and Lygerou, Z. (2011). Idas, a Novel Phylogenetically Conserved Geminin-related Protein, Binds to Geminin and Is Required for Cell Cycle Progression. J. Biol. Chem. *286*, 23234–23246.

Pellegrini, M., Marcotte, E.M., Thompson, M.J., Eisenberg, D., and Yeates, T.O. (1999). Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. Proc. Natl. Acad. Sci. *96*, 4285–4288.

Persson, E., Kaduk, M., Forslund, S.K., and Sonnhammer, E.L.L. (2019). Domainoid: domain-oriented orthology inference. BMC Bioinformatics *20*, 523.

Pinu, F.R., Beale, D.J., Paten, A.M., Kouremenos, K., Swarup, S., Schirra, H.J., and Wishart, D. (2019). Systems Biology and Multi-Omics Integration: Viewpoints from the Metabolomics Research Community. Metabolites *9*.

Plewniak, F. (2003). PipeAlign: a new toolkit for protein family analysis. Nucleic Acids Res. *31*, 3829–3832.

Praetorius, H.A., and Spring, K.R. (2003). The renal cell primary cilium functions as a flow sensor: Curr. Opin. Nephrol. Hypertens. *12*, 517–520.

Que, J. (2015). The initial establishment and epithelial morphogenesis of the esophagus: a new model of tracheal–esophageal separation and transition of simple columnar into stratified squamous epithelium in the developing esophagus. Wiley Interdiscip. Rev. Dev. Biol. *4*, 419–430.

Quigley, I.K., and Kintner, C. (2017). Rfx2 stabilizes Foxj1 binding at chromatin loops to enable multiciliated cell gene expression. PLoS Genet. *13*, e1006538.

Rajagopala, S.V., and Uetz, P. (2011). Analysis of Protein–Protein Interactions Using High-Throughput Yeast Two-Hybrid Screens. In Network Biology, G. Cagney, and A. Emili, eds. (Totowa, NJ: Humana Press), pp. 1–29.

Rappoport, N., and Shamir, R. (2018). Multi-omic and multi-view clustering algorithms: review and cancer benchmark. Nucleic Acids Res. *46*, 10546–10562.

Reiter, J.F., and Leroux, M.R. (2017). Genes and molecular pathways underpinning ciliopathies. Nat. Rev. Mol. Cell Biol. *18*, 533–547.

Reiter, J.F., Blacque, O.E., and Leroux, M.R. (2012). The base of the cilium: roles for transition fibres and the transition zone in ciliary formation, maintenance and compartmentalization. EMBO Rep. *13*, 608–618.

Revinski, D.R., Zaragosi, L.-E., Boutin, C., Ruiz-Garcia, S., Deprez, M., Thomé, V., Rosnet, O., Gay, A.-S., Mercey, O., Paquet, A., et al. (2018). CDC20B is required for deuterosome-mediated centriole production in multiciliated cells. Nat. Commun. *9*, 4668.

Ridge, R.W., Hori, T., and Miyamura, S. (1997). Analysis of Flagellar Movement in Ginkgo biloba Sperm by High Speed Video Microscopy. In Ginkgo Biloba A Global Treasure, T. Hori, R.W. Ridge, W. Tulecke, P. Del Tredici, J. Trémouillaux-Guiller, and H. Tobe, eds. (Tokyo: Springer Japan), pp. 99–107.

Ritchie, M.D., Holzinger, E.R., Li, R., Pendergrass, S.A., and Kim, D. (2015). Methods of integrating data to uncover genotype–phenotype interactions. Nat. Rev. Genet. *16*, 85–97.

Rock, J.R., Gao, X., Xue, Y., Randell, S.H., Kong, Y.-Y., and Hogan, B.L. (2011). Notch-dependent differentiation of adult airway basal stem cells. Cell Stem Cell *8*, 639–648.

Ronshaugen, M., McGinnis, N., and McGinnis, W. (2002). Hox protein mutation and macroevolution of the insect body plan. Nature *415*, 914–917.

Ross, A.J., Dailey, L.A., Brighton, L.E., and Devlin, R.B. (2007). Transcriptional profiling of mucociliary differentiation in human airway epithelial cells. Am. J. Respir. Cell Mol. Biol. *37*, 169–185.

Rubin, G.M., Yandell, M.D., Wortman, J.R., Gabor, G.L., Miklos, Nelson, C.R., Hariharan, I.K., Fortini, M.E., Li, P.W., Apweiler, R., et al. (2000). Comparative Genomics of the Eukaryotes. Science *287*, 2204–2215.

Ruiz García, S., Deprez, M., Lebrigand, K., Cavard, A., Paquet, A., Arguel, M.-J., Magnone, V., Truchi, M., Caballero, I., Leroy, S., et al. (2019). Novel dynamics of human mucociliary differentiation revealed by single-cell RNA sequencing of nasal epithelial cultures. Dev. Camb. Engl. *146*.

Sanderson, M.J., and Sleigh, M.A. (1981). Ciliary activity of cultured rabbit tracheal epithelium: beat pattern and metachrony. J. Cell Sci. *47*, 331.

Sardiu, M.E., and Washburn, M.P. (2011). Building Protein-Protein Interaction Networks with Proteomics and Informatics Tools. J. Biol. Chem. *286*, 23645–23651.

Satir, P., and Christensen, S.T. (2007). Overview of Structure and Function of Mammalian Cilia. Annu. Rev. Physiol. *69*, 377–400.

Schmidt, T.I., Kleylein-Sohn, J., Westendorf, J., Le Clech, M., Lavoie, S.B., Stierhof, Y.-D., and Nigg, E.A. (2009). Control of Centriole Length by CPAP and CP110. Curr. Biol. *19*, 1005–1011.

Schulte, E. (1972). Untersuchungen an der Regio olfactoria des Aals, Anguilla anguilla L. Z. Für Zellforsch. Mikrosk. Anat. *125*, 210–228.

Sebastiani, P., Timofeev, N., Dworkis, D.A., Perls, T.T., and Steinberg, M.H. (2009). Genome-wide association studies and the genetic dissection of complex traits. Am. J. Hematol. *84*, 504–515.

Shiga, Y., Yasumoto, R., Yamagata, H., and Hayashi, S. (2002). Evolving role of Antennapedia protein in arthropod limb patterning. Development *129*, 3555–3561.

Shunatova, N., and Borisenko, I. (2020). Proliferating activity in a bryozoan lophophore. PeerJ *8*, e9179.

Singla, V., and Reiter, J.F. (2006). The Primary Cilium as the Cell's Antenna: Signaling at a Sensory Organelle. Science *313*, 629–633.

Smyth, G.K. (2005). limma: Linear Models for Microarray Data. In Bioinformatics and Computational Biology Solutions Using R and Bioconductor, R. Gentleman, V.J. Carey, W. Huber, R.A. Irizarry, and S. Dudoit, eds. (New York, NY: Springer), pp. 397–420.

Song, R., Walentek, P., Sponer, N., Klimke, A., Lee, J.S., Dixon, G., Harland, R., Wan, Y., Lishko, P., Lize, M., et al. (2014). miR-34/449 miRNAs are required for motile ciliogenesis by repressing cp110. Nature *510*, 115–120.

Sonnhammer, E.L.L., and Koonin, E.V. (2002). Orthology, paralogy and proposed classification for paralog subtypes. Trends Genet. *18*, 619–620.

Sonnhammer, E.L.L., and Östlund, G. (2015). InParanoid 8: orthology analysis between 273 proteomes, mostly eukaryotic. Nucleic Acids Res. *43*, D234–D239.

Sonnhammer, E.L.L., Gabaldon, T., Sousa da Silva, A.W., Martin, M., Robinson-Rechavi, M., Boeckmann, B., Thomas, P.D., Dessimoz, C., and the Quest for Orthologs consortium (2014). Big data and other challenges in the quest for orthologs. Bioinformatics *30*, 2993–2998.

Spassky, N. (2005). Adult Ependymal Cells Are Postmitotic and Are Derived from Radial Glial Cells during Embryogenesis. J. Neurosci. *25*, 10–18.

Spassky, N., and Meunier, A. (2017). The development and functions of multiciliated epithelia. Nat. Rev. Mol. Cell Biol. *18*, 423–436.

Standring, S. (2016). Gray's anatomy: the anatomical basis of clinical practice (New York: Elsevier Limited).

Stolzer, M., Siewert, K., Lai, H., Xu, M., and Durand, D. (2015). Event inference in multidomain families with phylogenetic reconciliation. BMC Bioinformatics *16*, S8.

Straube, N., Li, C., Mertzen, M., Yuan, H., and Moritz, T. (2018). A phylogenomic approach to reconstruct interrelationships of main clupeocephalan lineages with a critical discussion of morphological apomorphies. BMC Evol. Biol. *18*, 158.

Stubbs, J.L., Vladar, E.K., Axelrod, J.D., and Kintner, C. (2012). Multicilin promotes centriole assembly and ciliogenesis during multiciliate cell differentiation. Nat. Cell Biol. *14*, 140–147.

Subramanian, I., Verma, S., Kumar, S., Jere, A., and Anamika, K. (2020). Multi-omics Data Integration, Interpretation, and Its Application. Bioinforma. Biol. Insights *14*.

Szklarczyk, D., Gable, A.L., Lyon, D., Junge, A., Wyder, S., Huerta-Cepas, J., Simonovic, M., Doncheva, N.T., Morris, J.H., Bork, P., et al. (2019). STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. Nucleic Acids Res. *47*, D607–D613.

Talavera, G., and Castresana, J. (2007). Improvement of Phylogenies after Removing Divergent and Ambiguously Aligned Blocks from Protein Sequence Alignments. Syst. Biol. *56*, 564–577.

Tan, F.E., Vladar, E.K., Ma, L., Fuentealba, L.C., Hoh, R., Espinoza, F.H., Axelrod, J.D., Alvarez-Buylla, A., Stearns, T., Kintner, C., et al. (2013). Myb promotes centriole amplification and later steps of the multiciliogenesis program. Development *140*, 4277–4286.

Tang, C.-J.C., Lin, S.-Y., Hsu, W.-B., Lin, Y.-N., Wu, C.-T., Lin, Y.-C., Chang, C.-W., Wu, K.-S., and Tang, T.K. (2011). The human microcephaly protein STIL interacts with CPAP and is required for procentriole formation. EMBO J. *30*, 4790–4804.

Terré, B., Piergiovanni, G., Segura-Bayona, S., Gil-Gómez, G., Youssef, S.A., Attolini, C.S.-O., Wilsch-Bräuninger, M., Jung, C., Rojas, A.M., Marjanović, M., et al. (2016). GEMC1 is a critical regulator of multiciliated cell differentiation. EMBO J. *35*, 942–960.

The Cancer Genome Atlas Network (2012). Comprehensive molecular portraits of human breast tumors. Nature *490*, 61–70.

The Gene Ontology Consortium (2019). The Gene Ontology Resource: 20 years and still GOing strong. Nucleic Acids Res. *47*, D330–D338.

The Uniprot Consortium (2019). UniProt: a worldwide hub of protein knowledge. Nucleic Acids Res. *47*, D506–D515.

Thompson, J.D., Plewniak, F., Thierry, J.-C., and Poch, O. (2000). DbClustal: rapid and reliable global multiple alignments of protein sequences detected by database searches. Nucleic Acids Res. *28*, 2919–2926.

Thompson, J.D., Thierry, J.C., and Poch, O. (2003). RASCAL: rapid scanning and correction of multiple sequence alignments. Bioinformatics *19*, 1155–1161.

Thompson, J.D., Muller, A., Waterhouse, A., Procter, J., Barton, G.J., Plewniak, F., and Poch, O. (2006). MACSIMS : multiple alignment of complete sequences information management system. BMC Bioinformatics *7*, 318.

Tilley, A.E., Walters, M.S., Shaykhiev, R., and Crystal, R.G. (2015). Cilia Dysfunction in Lung Disease. Annu. Rev. Physiol. *77*, 379–406.

Tukey, J.W. (1977). Exploratory Data Analysis (Addison-Wesley Publishing Company Reading, Mass.).

Uchiyama, I. (2006). Hierarchical clustering algorithm for comprehensive orthologous-domain classification in multiple genomes. Nucleic Acids Res. *34*, 647–658.

Uchiyama, I., Mihara, M., Nishide, H., Chiba, H., and Kato, M. (2019). MBGD update 2018: microbial genome database based on hierarchical orthology relations covering closely related and distantly related comparisons. Nucleic Acids Res. *47*, D382–D389.

Uetz, P., Freed, P., and Hošek, J. (2020). The Reptile Database, http://www.reptile-database.org, consulté le 15 juillet 2020.

Vanhoutreve, R., Kress, A., Legrand, B., Gass, H., Poch, O., and Thompson, J.D. (2016). LEON-BIS: multiple alignment evaluation of sequence neighbours using a Bayesian inference system. BMC Bioinformatics *17*.

Vladar, E.K., Bayly, R.D., Sangoram, A.M., Scott, M.P., and Axelrod, J.D. (2012). Microtubules Enable the Planar Cell Polarity of Airway Cilia. Curr. Biol. *22*, 2203–2212.

Walentek, P., and Quigley, I.K. (2017). What we can learn from a tadpole about ciliopathies and airway diseases – Using systems biology in Xenopus to study cilia and mucociliary epithelia. Genes. N. Y. N 2000 *55*.

Wallingford, J.B. (2010). Planar cell polarity signaling, cilia and polarized ciliary beating. Curr. Opin. Cell Biol. *22*, 597–604.

Wallmeier, J., Al-Mutairi, D.A., Chen, C.-T., Loges, N.T., Pennekamp, P., Menchen, T., Ma, L., Shamseldin, H.E., Olbrich, H., Dougherty, G.W., et al. (2014). Mutations in CCNO result in congenital mucociliary clearance disorder with reduced generation of multiple motile cilia. Nat. Genet. *46*, 646–651.

Wallmeier, J., Frank, D., Shoemark, A., Nöthe-Menchen, T., Cindric, S., Olbrich, H., Loges, N.T., Aprea, I., Dougherty, G.W., Pennekamp, P., et al. (2019). De Novo Mutations in FOXJ1 Result in a Motile Ciliopathy with Hydrocephalus and Randomization of Left/Right Body Asymmetry. Am. J. Hum. Genet. *105*, 1030–1039.

Wang, L., Fu, C., Fan, H., Du, T., Dong, M., Chen, Y., Jin, Y., Zhou, Y., Deng, M., Gu, A., et al. (2013). miR-34b regulates multiciliogenesis during organ formation in zebrafish. Development *140*, 2755–2764.

Ward, J.H. (1963). Hierarchical Grouping to Optimize an Objective Function. J. Am. Stat. Assoc. *58*, 236–244.

Waterhouse, A.M., Procter, J.B., Martin, D.M.A., Clamp, M., and Barton, G.J. (2009). Jalview Version 2—a multiple sequence alignment editor and analysis workbench. Bioinformatics *25*, 1189–1191.

Waterhouse, R.M., Seppey, M., Simão, F.A., Manni, M., Ioannidis, P., Klioutchnikov, G., Kriventseva, E.V., and Zdobnov, E.M. (2018). BUSCO Applications from Quality Assessments to Gene Prediction and Phylogenomics. Mol. Biol. Evol. *35*, 543–548.

Werner, M.E., Hwang, P., Huisman, F., Taborek, P., Yu, C.C., and Mitchell, B.J. (2011). Actin and microtubules drive differential aspects of planar cell polarity in multiciliated cells. J. Cell Biol. *195*, 19–26.

Wheway, G., Parry, D.A., and Johnson, C.A. (2014). The role of primary cilia in the development and disease of the retina. Organogenesis *10*, 69–85.

Wu, Y.-C., Rasmussen, M.D., and Kellis, M. (2012). Evolution at the Subgene Level: Domain Rearrangements in the Drosophila Phylogeny. Mol. Biol. Evol. *29*, 689–705.

Yachdav, G., Wilzbach, S., Rauscher, B., Sheridan, R., Sillitoe, I., Procter, J., Lewis, S.E., Rost, B., and Goldberg, T. (2016). MSAViewer: interactive JavaScript visualization of multiple sequence alignments. Bioinformatics *32*, 3501–3503.

Yang, Q., Zhang, A., Miao, J., Sun, H., Han, Y., Yan, G., Wu, F., and Wang, X. (2019). Metabolomics biotechnology, applications, and future trends: a systematic review. RSC Adv. *9*, 37245–37257.

Yates, A.D., Achuthan, P., Akanni, W., Allen, J., Allen, J., Alvarez-Jarreta, J., Amode, M.R., Armean, I.M., Azov, A.G., Bennett, R., et al. (2020). Ensembl 2020. Nucleic Acids Res. *48*, D682–D688.

Yi, Y., Lv, Y., You, X., Chen, J., Bian, C., Huang, Y., Xu, J., Deng, L., and Shi, Q. (2019). High throughput screening of small immune peptides and antimicrobial peptides from the Fish-T1K database. Genomics *111*, 215–221.

You, Y., Huang, T., Richer, E.J., Schmidt, J.-E.H., Zabner, J., Borok, Z., and Brody, S.L. (2004). Role of f-box factor foxj1 in differentiation of ciliated airway epithelial cells. Am. J. Physiol. - Lung Cell. Mol. Physiol. *286*, L650–L657.

Yuan, S., Liu, Y., Peng, H., Tang, C., Hennig, G.W., Wang, Z., Wang, L., Yu, T., Klukovich, R., Zhang, Y., et al. (2019). Motile cilia of the male reproductive system require miR-34/miR-449 for development and function to generate luminal turbulence. Proc. Natl. Acad. Sci. *116*, 3584–3593.

Yuasa, A. (1933). Studies in the Cytology of Pteridophyta:IV. On the Spermatozoids of Selaginella, Isoetes and Salvinia. Shokubutsugaku Zasshi *47*, 697–709.

Zhang, Z., Wu, S., Stenoien, D.L., and Paša-Tolić, L. (2014). High-Throughput Proteomics. Annu. Rev. Anal. Chem. *7*, 427–454.

Zhao, H., Zhu, L., Zhu, Y., Cao, J., Li, S., Huang, Q., Xu, T., Huang, X., Yan, X., and Zhu, X. (2013). The Cep63 paralogue Deup1 enables massive de novo centriole biogenesis for vertebrate multiciliogenesis. Nat. Cell Biol. *15*, 1434–1444.

Zheng, M., Hu, Y., Gou, R., Wang, J., Nie, X., Li, X., Liu, Q., Liu, J., and Lin, B. (2019). Integrated multi-omics analysis of genomics, epigenomics, and transcriptomics in ovarian carcinoma. Aging *11*, 4198–4215.

Zheng, X., Ramani, A., Soni, K., Gottardo, M., Zheng, S., Ming Gooi, L., Li, W., Feng, S., Mariappan, A., Wason, A., et al. (2016). Molecular basis for CPAP-tubulin interaction in controlling centriolar and ciliary length. Nat. Commun. *7*, 11874.

Zhou, F., Narasimhan, V., Shboul, M., Chong, Y.L., Reversade, B., and Roy, S. (2015). Gmnc Is a Master Regulator of the Multiciliated Cell Differentiation Program. Curr. Biol. *25*, 3267–3273.

ANNEXES

Annexe 1
Orthology: promises and challenges

Orthology: promises and challenges

Yannis Nevers[1,2,3,4], Audrey Defosset[1] and Odile Lecompte*,[1]

[1]Complex Systems and Translational Bioinformatics, ICube UMR 7357, Université de Strasbourg, Strasbourg, France

[2]SIB Swiss Institute of Bioinformatics, Lausanne, Switzerland

[3]Department of Computational Biology, University of Lausanne, Lausanne, Switzerland

[4]Center for Integrative Genomics, University of Lausanne, Lausanne, Switzerland

*Corresponding author: E-mail: odile.lecompte@unistra.fr

Co-authors: yannis.nevers@unil.ch, adefosset@etu.unistra.fr

Abstract

Orthology is a cornerstone of comparative genomics and has numerous applications in current biology. In this chapter, we first introduce the concepts of orthology and paralogy. We then present the currently available orthology inference methods and the community-led efforts of standardization and benchmarking accompanying these developments. The large panel of available orthology resources is compared in terms of species coverage, access, contextual data and tools proposed to end-users to facilitate the analysis and exploitation of orthology data. We then review the importance of orthology applications, ranging from the study of protein families and information transfer to the comparison of genomes and genotype/phenotype correlations. Finally, we discuss the current challenges in the orthology field, faced with an ever-increasing number of proteomes of particularly heterogeneous quality. We highlight the urgent need of considering orthology at the protein domain and transcript levels and the conceptual and practical difficulties that this raises.

Introduction

Homology is a central concept in biology, and is essential for any intra-species or interspecies sequence comparison. Originally employed to compare phenotypic traits, it is now mainly used to define relationships between genomic regions, genes and, by extension, between proteins or even sub-protein regions. In this context, homology describes the relationship between two molecular entities (usually genes or proteins) that descend from the same ancestor. Two main categories of homologs were distinguished in the early days of molecular biology (Fitch 1970): paralogs that derive from a common ancestor by a duplication event and orthologs that emerge after a speciation event (Figure 1a). *Stricto sensu*, these definitions only refer to the evolutionary history of genes. However, it is commonly accepted that orthologs tend to retain a similar function while paralogs may have different fates in the course of evolution. Indeed, the paralogous copies may develop more specialized functions compared to the ancestral gene (tissue/stage specific expression, complementation of functions initially performed by a single gene) or one copy may evolve a new function under the reduced selection pressure or even degenerate into a pseudogene (Force et al. 1999). The ‘orthology conjecture', stating that orthologs frequently retain ancestral function while paralogs tend to diversify is widely used to transfer functional information between orthologs. Although this hypothesis is commonly accepted by the community, it has been challenged in some cases (Studer and Robinson-Rechavi 2009; Nehrt et al. 2011), especially among highly similar genes. Nevertheless, it still generally holds (Altenhoff et al. 2012; Chen and Zhang 2012). Notably, it has been shown that the organization of introns (Henricson et al. 2010), the three-dimensional structure of proteins (Peterson et al. 2009), and domain architecture (Forslund et al. 2011) tend to be more conserved between orthologs than paralogs. In addition, orthologs are generally expressed in the same tissues in contrast to paralogs (Kryuchkova-Mostacci and Robinson-Rechavi 2015).

The debate around the orthology conjecture underlines the importance of taking into account the chronology of speciation and duplication events to establish functional links between homologous genes. Indeed, paralogs that derive from a ‘recent’ duplication event may still share the same function in contrast to distant paralogs separated over millions of years of evolution. Unfortunately, there is no objective threshold to define recent versus ancient paralogs and in fact, it all depends on the evolutionary distance between compared species. This has been conceptualized with the terms ‘outparalogs’ and ‘inparalogs’ coined in 2002 (Sonnhammer and Koonin 2002). When comparing two species, paralogs deriving from a duplication event that occurred prior to the speciation event are called outparalogs, while paralogs originating from a duplication event subsequent to the speciation event are called inparalogs. Inparalogs are considered to be co-orthologs of genes descending from the speciation event in the other species (Figure 1b). Hence, inparalogy and outparalogy are relative notions: the same paralogous sequences can be considered inparalogs or outparalogs depending on the speciation referred to. The co-orthology concept also introduces different orthology relationships: 1-to-1, 1-to-many and many-to-many orthologs (Figure 1b).

The characterization of these intricate homology relationships is far from trivial since there is no direct record of past speciation or duplication events, and evolutionary scenarios can be further complicated by lineage-specific gene losses, whole genome duplications (WGD) and

horizontal gene transfers (HGT). WGD or polyploidy can arise within a single species by the doubling of the chromosome set (autopolyploidy) or can result from the merging of the chromosome sets of two different species and subsequent genome doubling (allopolyploidy) (see Van de Peer et al. 2017 for a recent review). Homologs arising by autopolyploidy are called ohnologs (Wolfe 2000) and constitute a special case of paralogs, since both copies evolved originally in the same genomic context. Homeologs that result from an allopolyploidy event are more complex to define (reviewed in Glover et al. 2016) but are observed in many plants. Like orthologs, they originally emerge after a speciation event but they are subsequently integrated in a single genome through autopolyploidization. Thus, homeologs experience a mosaic fate by initially evolving like orthologs and then after hybridization, undergoing an evolutionary pressure usually exerted on paralogs.

In HGT, the relationship does not rely on vertical transmission of genes but on acquisition of genetic material from another species. Genes whose history since their common ancestor involves an horizontal transfer are called xenologs (Gray and Fitch 1983; Fitch 2000). Xenology is especially prevalent in prokaryotes with HGT frequently involving mobile genetic elements, but it can also occur between prokaryotes and eukaryotes (notably in the case of endosymbiosis or endoparasitism) or even between eukaryotes (reviewed in Soucy et al. 2015). Xenology relationships encompass a wide range of evolutionary histories and xenolog classes have been proposed to reflect the events associated with the divergence of xenologs and the relative timing of transfer and speciation events (Darby et al. 2017).

The first step in the process of characterization of homology relations is based on sequence comparison. It is assumed that genes/proteins are homologous if they exhibit a higher similarity than would be expected by chance. Thus, homology detection usually relies on similarity searches, typically a BLAST search (Altschul et al. 1997; Camacho et al. 2009), with a fixed threshold of score, percentage identity, expect-value, etc. The distinction at the genome scale between the different types of homology (1-to-1 orthology, co-orthology, inparalogy, outparalogy, xenology) then requires dedicated approaches. The methods used to infer orthology and the corresponding available resources are presented in the first section of this chapter. We then review the main applications of orthology in biology. In the last section, we highlight the practical and conceptual challenges around the notion of orthology and its uses.

1 Orthology inference and resources

1.1 Orthology inference methods

An exhaustive description of the plethora of available programs is beyond the scope of this review (for a recent review on methods, see (Altenhoff et al. 2019)). However, these different programs can be classified into four main categories: graph-based, tree-based, hybrid and meta-prediction methods, that are presented briefly below.

In graph-based methods, genes/proteins are represented by nodes and homology relationships by edges in the graph. The graph construction relies on all-against-all similarity searches between genes/proteins from two genomes. The simplest approach, called Reciprocal Best Hit (RBH), will predict an orthology relationship between proteins A and B from two genomes if A is the genome-wide closest relative of B and *vice versa* (Overbeek et

al. 1999). This approach only considers 1-to-1 orthology relationships, thus overlooking one-to-many and many-to-many orthologs. To circumvent this problem and offer a more comprehensive view of evolutionary relationships, other algorithms have been developed where inparalogy relations are inferred and included during graph construction. Examples of such methods include COG (Tatusov et al. 1997), Inparanoid (Remm et al. 2001), OrthoMCL (Li et al. 2003), OMA (Roth et al. 2008), EggNOG (Jensen et al. 2008), OrthoInspector (Linard et al. 2011) and OrthoFinder (Emms and Kelly 2015). The homology relationships predicted between a pair of genomes can then be extended to a set of species, in order to define groups of orthologs (also called orthogroups) present in these species. The groups are delineated on the basis of the structure of the graph by transitivity or clustering. For instance, OrthoMCL uses Markov clustering to partition the homology graph into orthogroups containing highly connected orthologs and recent paralogs. OMA Groups are based on cliques, i.e. fully connected subgraphs corresponding to genes that are all orthologs to each other, thus *de facto* excluding orthologs involved in 1-to-many or many-to-many relations.

Tree-based methods infer orthologs based on the gene's evolutionary history, which is reconstructed by reconciling the gene family tree with the species tree. First, a multiple alignment of homologous sequences is constructed to generate a phylogenetic tree of the gene family. Then, the nodes of this gene-tree are labeled as duplication or speciation events by comparison to the species-tree during the reconciliation step, allowing the prediction of orthology and paralogy relationships. This type of approach is implemented in numerous programs, including RIO (Zmasek and Eddy 2002), Orthostrapper (Storm and Sonnhammer 2002), PhylomeDB (Huerta-Cepas et al. 2007), Ensembl Compara (Vilella et al. 2009), PANTHER(Mi et al. 2010). These methods produce hierarchical ortholog groups, i.e. groups of orthologs and inparalogs deriving from a common ancestor, in the form of trees. These hierarchical groups are more informative than simple orthology relationships between pairs of species or flat groups of orthologs without evolutionary information about intra-group relations. Unfortunately, tree-based methods are highly dependent on the construction of correct multiple alignments and trees and are computationally demanding, preventing their application to very large datasets.

Although hierarchical groups are naturally produced by tree-based methods, they can also be generated by a post-processing of orthogroups obtained by graph-based methods. As an example, EggNog and OrthoDB explicitly delineate the hierarchy of ortholog groups by identifying orthogroups at different taxonomic levels of the species tree. Hybrid methods go further by using attributes of graph-based and tree-based methods in the inference of orthology relationships itself. The method of OMA Hierarchical Orthologous Groups (HOG) (Altenhoff et al. 2013) uses an orthology graph of pairwise relations to form groups, starting with the most specific taxonomic level and progressively merging groups toward the root of the species tree. Hieranoid (Schreiber and Sonnhammer 2013) progressively calculates pairwise orthology relationships using RBH at each node of a guide tree from the leaves to the ancestor. At each node, a consensus or a profile is built from the child nodes and used for subsequent pairwise comparisons, which considerably reduces the number of required pairwise comparisons. OrthoFinder 2 (Emms and Kelly 2019) first identifies orthogroups among a set of species using the OrthoFinder graph-based approach (Emms and Kelly 2015) and then uses the orthogroups to infer approximate gene trees and a species tree. Finally, each gene tree is compared to the species tree to infer duplication events and refine prediction of orthology and paralogy relations.

Meta-prediction methods are designed to exploit predictions generated by different programs and thus, can potentially highlight false positives and negatives. As an example, DIOPT (Hu et al. 2011) assigns a score to each orthology relationship according to the number of independent methods predicting this relation. The MARIO program (Pereira et al. 2014) goes further by delineating a group of orthologs from predictions of several methods, and constructing a Hidden Markov Model (HMM) profile of these orthologous sequences. This profile is then used to evaluate the predictions made by each individual method. MetaPhOrs (Pryszcz et al. 2011) integrates phylogenetic trees constructed by several methods to predict orthology relations and assigns a score depending on the number of predictions. This filters unreliable results linked to poor resolution of phylogenetic trees. The WORMHOLE program (Sutphin et al. 2016) uses a classifier based on support vector machines (SVM) trained on a positive set of validated orthology relationships and a negative set of non-orthology gene pairs. The algorithm assigns a weight to each prediction method depending on its performance in different test cases (e.g. according to the proximity of the species under consideration). This weight is then used to combine predictions on a complete dataset and extract reliable orthology relations.

1.2 Standardization and benchmarking

Given the multiplicity of orthology inference methods available, it is crucial to cross-reference, compare and evaluate their predictions in different biological contexts in order to choose the relevant program for a given biological question and to improve prediction methods. This requires a standardization of orthology prediction formats and an objective benchmarking. These topics are the central goals of the Quest For Orthologs (QFO) consortium (Gabaldón et al. 2009). QFO addresses both conceptual issues and technical challenges in orthology prediction. For example, community efforts led to the development of the standardized OrthoXML format (Schmitt et al. 2011) designed to represent orthology predictions for both graph- and tree-based methods. An ontology (Fernández-Breis et al. 2016) has also been developed to formalize the representation of orthology relationships. This ontology allows the representation of data according to a semantic Web standard, RDF (Resource Descriptions Framework), that facilitates interoperability between resources.

The QFO consortium has also defined a QFO reference proteome dataset to allow the comparison of methods on a common set of species and proteins. The dataset is updated every year and currently comprises 78 UniProt Reference proteomes. It includes sequences from model organisms, species of interest for biomedical or agronomic research, or species of interest from a phylogenetic point of view (Sonnhammer et al. 2014). In parallel, a variety of benchmarks have been developed to evaluate orthology prediction methods according to phylogenetic and functional criteria. A large-scale benchmarking study (Altenhoff et al. 2016) comparing 15 orthology methods highlighted a trade-off between sensitivity and specificity and clearly showed that the best approach is highly dependent on the biological context. Overall, the orthogroup predictions of OMA are characterized by high specificity, whereas the tree-based method used in PANTHER has high sensitivity. However, there is no systematic difference between tree-based and graph-based methods. Finally, Inparanoid, Hieranoid and OrthoInspector as well as OrthoFinder in the most recent version of the benchmark (results available at https://orthology.benchmarkservice.org), show a good balance between specificity and sensitivity over all benchmarks. Orthology predictions from the best methods identified by the benchmarking are now integrated in the Alliance of Genome Resources (Alliance) portal (Alliance of Genome Resources Consortium 2020). The

Alliance aims to facilitate exploration of orthologous genes in human and well-studied model organisms in order to exploit the wealth of genetic and genomic studies available in these organisms.

1.3 Orthology resources

Most orthology inference programs can be installed and executed locally on a user-defined set of proteomes, but many of them are also used to generate databases of orthology relationships. These resources are essential for the routine use of the orthology concept by non-experts. The databases differ in terms of number and diversity of represented species (Table 1), which determines the granularity with which orthology relationships can be exploited. Some generalist databases cover a large panel of species such as EggNog (Huerta-Cepas et al. 2016), HOGENOM (Penel et al. 2009), Inparanoid (Sonnhammer and Östlund 2015), MBGD (Uchiyama et al. 2019), OMA (Altenhoff et al. 2018), OrthoDb (Kriventseva et al. 2019) and OrthoInspector (Nevers et al. 2019). EggNog and OrthoDB also include viral genomes. Other resources are clade-specific, including TreeFam (for Metazoa) (Schreiber et al. 2014), FungiPath (for Fungi) (Grossetête et al. 2010), and GreenPhylDB (Rouard et al. 2011) and PLAZA (Van Bel et al. 2018) that focus on plants. With the exception of MetaPhOrs (Pryszcz et al. 2011), the resources based on meta-predictions generally focus on a small number of model species (Table 1). In addition to the databases dedicated to orthology, orthology relationships are also provided in more general biological portals, such as PANTHER (Mi et al. 2019) , Ensembl Compara (Herrero et al. 2016) and HomoloGene (NCBI Resource Coordinators 2016).

Orthology databases offer diverse access to information, via web interfaces for manual exploration or using programmatic access through web services or SPARQL (SPARQL Protocol and RDF Query Language) interfaces. Users can search for orthologs of a given gene using genes/proteins or orthogroup identifiers, or perform a sequence similarity search. Information can also be accessed through functional annotation of the gene of interest (keywords, description or GO annotations). For instance, OrthoInspector (Nevers et al. 2019) allows users to retrieve all proteins of a given species associated with a given GO term and visualize their evolutionary histories. OrthoMCL (Chen et al. 2006) and GreenPhylDB (Rouard et al. 2011) propose searches for groups with a given protein domain. Genes can also be retrieved on the basis of their phylogenetic distribution, i.e. the presence or absence of an ortholog in different taxa. This phylogenetic profiling search is implemented in MBGD (Uchiyama et al. 2019), OrthoDb (Kriventseva et al. 2019), OrthoInspector (Nevers et al. 2019), OrthologeDB (Whiteside et al. 2013), OrthoMCL (Chen et al. 2006) and GreenPhylDB (Rouard et al. 2011). It can be used to perform genotype/phenotype studies as discussed in the Applications section.

All orthology databases provide orthology predictions in the form of a list of orthologs in the covered species, but many of them contextualize this minimum information by adding relevant data and tools to analyse and exploit the evolutionary information (Table 1). Hence, they frequently provide additional information about the function (GO term annotation, Enzyme classification numbers...) or architecture (protein domains) of the predicted orthologs as illustrated in Table 1. This functional information most often comes from automatic annotations that must be handled with care. However, viewing the annotations for all the orthologs of a protein makes it easier to detect inconsistencies and spurious annotations. For example, OMA (Altenhoff et al. 2018) offers a synthetic representation of the GO annotations of the detected orthologs with a color code that distinguishes between automatic annotation, annotation validated by an expert and annotation based on experimental data. Multiple sequence alignment (MSA) and phylogenetic trees also

constitute an essential analytical tool for a more in-depth understanding of the relationships between orthologs and paralogs. As such, they are often made available, in particular by tree-based methods. They are either pre-calculated and available directly on the web interface, or can be constructed 'on the fly' for a selection of predicted orthologous sequences. In addition, some resources provide information about the genomic context of the query gene and its orthologs, allowing to detect syntenic stretches of genes that can be helpful for the validation of orthology relations and may be indicative of a functional link between syntenic genes. Finally, orthology resources can provide the taxonomic distribution of detected orthologs in each species represented in the orthology database. This is suitable for clade-specific resources such as GreenPhylDb (Rouard et al. 2011) and PLAZA (Van Bel et al. 2018). For generalist orthology resources, a synthetic view of distributions is required as exemplified by OrthoInspector (Nevers et al. 2019) that provides schematic representations of phylogenetic distributions at different granularity levels.

2 Orthology: the Swiss army knife of genomics

2.1 Exploration of gene and protein families

Since their definition in the early seventies, orthologs and paralogs have been traditionally used to study gene and protein families, in particular in the framework of multiple alignment analysis. By placing a gene or a protein sequence in its evolutionary context, the multiple alignment reveals selection pressure existing at particular sequence positions, allowing the straightforward detection of conserved motifs, localization signals or key functional residues for a considered family of orthologs or a superfamily regrouping several paralogous families (Lecompte et al. 2001). Such evolutionary analyses are essential for the determination of catalytic sites or residues involved in protein interactions for example. This can be exploited to decipher residues, motifs or domains involved in the specificity of paralogous families, for instance to identify residues responsible for the enzyme substrate specificity in a multienzymatic family. In addition, alignments of orthologs or homologs are exploited in both 2D and 3D structure prediction methods by comparative protein modeling (reviewed in Khan et al. 2016). With the increase of experimentally determined structures, a wide range of accurate models are now available that can be used to predict protein binding sites, effects of protein mutations, and for structure-guided virtual screening (Liu et al. 2011; Leelananda and Lindert 2016).

Orthologous sequences are directly exploited by many mutation analysis tools, such as PolyPhen (Adzhubei et al. 2010) or SIFT (Vaser et al. 2016), to predict the phenotypic effects of variants. Pairwise or multiple alignments of orthologous sequences are also used at the genomic level to highlight conserved regions that may reflect the existence of functional elements. Orthologs are also the cornerstone of phylogenetic studies aimed at deciphering the evolutionary history of a gene family or, more generally, phylogenetic relationships between species. The reconstruction of phylogenetic relationships between species has for a long time relied on a single family of genes, typically 16S/18S rRNA genes or well conserved housekeeping protein genes. Today, species phylogenies can be built using comparisons of several protein families, including genome-wide comparisons (Crawford et al. 2012). These studies generally focus on widely conserved protein families exhibiting one-to-one orthology relationships. Orthofinder directly exploits orthogroups within

a species set to construct a phylogenomics species tree using the Species Tree from All Genes (STAG) algorithm (Emms and Kelly 2018). With the multiplication of available genomes and metagenomes, such phylogenomics analyses have renewed our vision of the tree of life, for instance by highlighting the bacterial diversification (Hug et al. 2016), reshaping the eukaryotic tree (Burki et al. 2020), and revealing a new group of Archaea, the Asgard that questioned the position of Eukaryotes in the tree of life (Spang et al. 2015; Zaremba-Niedzwiedzka et al. 2017).

Orthologous sequences and phylogenetic trees can also be exploited for ancestral sequence reconstruction. The leaves of the phylogenetic tree represent the extant sequences of the family, while the root corresponds to the extinct common ancestor. The ancestor can be synthesized to experimentally explore its biochemical properties. This approach allows to resurrect an ancestral precursor with selected properties, such as thermostability, in order to initiate synthetic evolution experiments (Gumulya and Gillam 2017). It can also be used to decipher past environmental conditions. For example, the reconstruction of translation elongation factors from organisms that lived 3.5 billion years ago revealed that the thermostability of these factors declines in the course of evolution and suggested a 30°C decrease in environmental temperature (Gaucher et al. 2008). Ancestral sequence reconstruction methods also deduce the sequences present at each internal node of the tree. These intermediate states can help to elucidate evolutionary processes, in particular the main mutations involved in the distinct properties of extant proteins (Straub and Merkl 2019). Applied to whole genomes, ancestral reconstruction offers a partial view of ancestral gene repertoires, from the known repertoires of extant species. Such a resource is available on the ancestral genome portal, constructed from PANTHER inferences (Huang et al. 2019).

2.2 Information transfer

As stated above, orthologous genes tend to retain equivalent functions across species and are thus widely used to transfer information from model species to poorly characterized ones. Typically, the functional annotation of genes in a newly sequenced genome is carried out by identifying annotated orthologs using similarity searches in protein databases such as UniProt (The UniProt Consortium 2019) or through the Gene Ontology (The Gene Ontology Consortium 2019), and then transferring these annotations to genes of unknown function. Several protocols (compared in (Amar et al. 2014)) allow this automated annotation transfer. Although this approach is time efficient, it can also lead to bias since the orthology conjecture is not an absolute law and the ortholog/paralog distinction is not trivial, especially in superfamilies (Schnoes et al. 2009). The problem of misannotation is also particularly severe, with multi-domain proteins exhibiting a differential conservation of some domains (discussed in section 3.3 Beyond gene level orthology). In addition, automated transfer can propagate annotation errors. It is therefore wise to rely on closely related orthologs with expert-curated annotations, whenever possible, to avoid the ‘percolation of annotation errors’ modeled by Gilks and colleagues and its deleterious effects on database quality (Gilks et al. 2002).

More generally, orthology can be used to transfer experimentally evidenced information obtained from one species to another, provided that the organisms are sufficiently close. This principle is used by the Gene Ontology Consortium (Ashburner et al. 2000; The Gene Ontology Consortium 2019) to propagate standardized annotations not only on

protein molecular function but also on their sub-cellular localization and the biological processes in which they are involved. The resulting annotations receive the IEA evidence code (Inferred from Electronic Annotation) in the case of an automatic transfer between orthologs. The Gene Ontology also integrates a semi-automated transfer protocol (Gaudet et al. 2011), taking into account annotations from several orthologs and the phylogenetic relationships between the corresponding species. These annotations are labeled with the IBA code (Inferred from Biological ancestry).

Information about protein-protein interactions (PPIs) can also be transferred from one species to another through the concept of interologs. The term 'interolog' (Walhout et al. 2000) refers to the conserved interaction between two pairs of proteins A1 and B1 from a first species and A2 and B2 from a second species. The A1/B1 interaction is considered as an interolog of the A2/B2 interaction if A1 and A2 are orthologs to B1 and B2 respectively. The concept of interology can be exploited in a predictive way: orthologs of interacting proteins in one species are identified and the PPI information is transferred to the pair of orthologs. To avoid false-positive errors, interology inferences are usually combined with other data, as illustrated by the STRING interaction database (Szklarczyk et al. 2019) that relies on a large panel of diverse evidence (experiments, text mining, co-expression, synteny, etc.).

Finally, when working on human genes, orthology relationships are key elements to consider when choosing a relevant model species for experimental studies. In addition to practical considerations (duration, cost, etc.), the model species should be chosen to avoid 1-to-many or many-to-1 orthology relations between the human and the model species, since the existence of additional inparalogs in one species would considerably complicate the interpretation of experimental results.

2.3 Comparison of genomes and proteomes

Comparisons of complete genomes and proteomes are intrinsically linked to the proper delineation of orthologs and paralogs. Comparisons of orthologs at the sequence level are used to evaluate the selection pressure acting to model evolutionary rates in different species. One of the first examples of such genome-wide assessment of evolution rates was carried out in mammalian and nematode lineages (Castillo-Davis et al. 2004). This study showed that strong purifying selection seems to act on the same central cellular processes (such as translation and protein transport) in mammals and nematodes, whereas positive selection acts on different biological processes in each lineage (DNA-dependent transcriptional regulation in nematodes, signal transduction via receptors and host immune response in mammals). Such comparative analyses are also performed for non-coding RNA genes such as microRNA. For example, the study of microRNA substitution rates in human and chimpanzee genomes revealed that primate-specific microRNAs have twice as many substitutions as older microRNA families (Santpere et al. 2016).

Comparison of proteomes in terms of gene content has become a quasi-obligatory step when sequencing a new genome. It requires the prediction of orthology and paralogy relations between the proteomes under consideration and reveals the set of conserved protein families, but also the acquisitions and losses that have taken place independently in each lineage. These comparisons have highlighted the extraordinary plasticity of the gene repertoire among species. This is particularly striking in the case of prokaryotic genomes. In

a comparison of more than 500 bacterial species, Lapierre and Gogarten (Lapierre and Gogarten 2009) showed that the conserved bacterial core was reduced to about 250 gene families, with the notable exception of certain symbionts exhibiting a particularly reduced genome. This diversity of gene repertoire observed even among closely related species can be explained by lineage-specific expansion of gene families, acquisition of genes by horizontal transfer (xenologs) and differential gene losses. In some prokaryotes, the genomic versatility is so important that large differences in gene content can occur between different strains of the same species. This led to the definition of the pangenome concept, i.e. the set of all genes present in a given species, that can be divided into the conserved core and the accessory genome (reviewed in (Brockhurst et al. 2019)). In species with an 'open' pangenome, the core genome conserved in all strains represents only a small fraction of the pangenome, questioning the concept of species in Prokaryotes. For instance in *Escherichia coli*, the core genome is restricted to about 3,000 gene families while the pangenome reaches a total of about 90,000 families (Land et al. 2015).

Comparisons of orthologous genomic regions or complete chromosomes decipher the evolution of genome architecture by revealing differential gains/losses of genomic regions, segmental duplications and balanced rearrangements. These comparisons can be made at the nucleotide level using, for example BLASTZ (Schwartz et al. 2003) or LASTZ and chaining/netting programs (Kent et al. 2003) to discriminate between orthologous and paralogous alignments. Alternatively, the comparison of genomic regions can be based on the comparison of genomic location of orthologs in different genomes to identify conserved syntenic blocks, i.e. a stretch of genes with a conserved gene order in different species. Such comparisons delineate syntenic genes frequently linked by functional relations and allow the detection of elements involved in genomic plasticity at the syntenic regions boundaries. They are also used to reconstruct ancestral genomes with distance/event-based or homology/adjacency-based methods (reviewed in (Feng et al. 2017)).

2.4 Functional inferences and genotype/phenotype correlations

Comparisons of complete proteomes based on orthology relationships can be exploited to perform functional inferences between genes or to detect genes potentially involved in a phenotype. The rationale behind this approach is that functionally linked genes are preserved or lost in a correlated manner over the course of evolution, and thus are found in the same species (Pellegrini et al. 1999). This assumption can be exploited in different ways. Subtractive analysis aims to identify genes restricted to species with a given phenotype. In practice, this means comparing the gene repertoire of at least two species (species A and B) possessing the phenotypic trait of interest and one or several related species (species C) lacking the considered phenotype. The set of genes with orthologs in species A and B but without orthologs in species C is likely to be enriched in genes associated with the phenotypic trait of interest. This approach was introduced by Huynen (Huynen et al. 1998) in the early days of comparative genomics in order to compare the genome of the pathogen *Helicobacter pylori* with that of another pathogen *Haemophilus influenzae* and a benign strain of *Escherichia coli*. They identified 17 gene families restricted to the pathogenic species and potentially involved in virulence and host-pathogen interactions.

The subtractive method is applicable to the search for genes linked to a phenotypic trait or biological process that has been lost/acquired in some species during evolution. This approach can be extended to the comparison of tens or hundreds of genomes to allow a

precise definition of the phenotypic distribution. The comparison of phylogenetically distinct lineages that have independently acquired (or lost) a given phenotype limits false positive predictions by eliminating genome differences simply due to random gains and losses of genes. For instance, Hecker and colleagues (Hecker et al. 2019) compared mammalian genomes to identify convergent gene losses associated with dietary adaptations in 6 independent herbivore lineages (16 species) and five independent carnivore lineages (15 species). Regarding the small evolutionary distances separating these placental mammals, they considered not only loss of entire genes or exons but also gene-inactivating mutations, using a genomic approach that combines the identification of orthologous regions and the CESAR program, a coding exon-structure aware realigner (Sharma et al. 2016).

At a larger evolutionary scale, another methodological framework is required. Phylogenetic profiles represent a generalization of subtractive analysis allowing the comparison of a large number of genomes that can be evolutionary distant. A phylogenetic profile of a gene represents the presence or absence of orthologs of that gene in the genomes of several species (Tatusov et al. 1997). Phylogenetic profiles were first used to infer the function of uncharacterized genes and the method has been successfully applied to the annotation of genes, mainly prokaryotes (see (Kensche et al. 2008) for examples). They are also exploited to predict functional links between genes, notably in the STRING (Szklarczyk et al. 2019) and OrthoInspector databases (Nevers et al. 2019).

Phylogenetic profiles can not only be compared to each other but also to all types of presence-absence distributions, including phenotypic traits. Phylogenetic profiling can thus be exploited to perform phenotype-genotype association studies. One of the first studies of this type was carried out on 86 prokaryotic genomes to identify genes associated with thermophily (Jim et al. 2004). Since then, many similar studies have been performed, notably to identify genes involved in human diseases thanks to the huge increase of available eukaryotic genomes that allows a detailed exploration of the distribution of human genes. For instance, Tabach and colleagues (Tabach et al. 2013) identified 54 clusters of phylogenetic profiles associated with a specific class of symptoms. More recently, the profiling of human genes in 100 eukaryotic species revealed 274 human genes exhibiting a phylogenetic distribution correlated with the distribution of cilia in eukaryotic lineages (Nevers et al. 2017). This set of predicted ciliary genes includes 87 new candidates. Among them, 21 have already been experimentally validated as ciliary genes.

3 Challenges

3.1 Keeping up with the data flow

As seen above, orthology is the cornerstone of a plethora of applications in comparative genomics and biology, and orthology resources provide numerous contextual data and analytical tools to facilitate orthology exploitation. Coming into a new decade, they are now gearing up to adapt to new challenges, a data flow brought by the next generation sequencing and a need to assess orthology at different granularity levels. The last two decades have seen a massive increase in sequencing capacities, leading to the acquisition of numerous genomes from across the Tree of Life. These genomes have obvious

usefulness for studying evolution at a broad scale and are increasingly incorporated into orthology resources. Nevertheless, they also lead to important challenges linked to the management and analysis of the ever increasing volume of data and the heterogeneous data quality.

Genomic data, hence genome annotations, have been increasing at an exponential rate with the advent of high-throughput sequencing technologies. As of today, 19,163 complete genomes are registered in the Genome Online Database (Mukherjee et al. 2019), as well as 215,613 genomes in the permanent draft state. This increase in data generation represents a challenge for orthology resources. It is especially true for tree-based approaches, which are commonly more computationally intensive as they rely on phylogenetic tree inference tools and are traditionally limited in the number of species they can include. While less computationally intensive, the data increase is still onerous for graph-based approaches, as they rely on all-vs-all sequence comparisons, which grow quadratically with the number of sequences. The legacy tools for these kinds of comparison, namely BLAST (Altschul et al. 1990; Camacho et al. 2009) or Smith-Waterman (Smith and Waterman 1981) alignment, do not scale well, and resources that use them rely heavily on high-performance computing clusters. Other tools and resources use faster but generally less sensitive solutions: MMSeq2 (standard modes) (Steinegger and Söding 2017), DIAMOND (Buchfink et al. 2015) or *ad-hoc* methods as in SwiftOrtho (Hu and Friedberg 2019) for instance can perform all-vs-all comparisons with better performances.

Nonetheless, solutions bypassing computationally intensive all-vs-all computations are increasingly being investigated, in anticipation of an even bigger surge in data. These approaches, such as EggNog-Mapper (Huerta-Cepas et al. 2017) aim to reduce the computation required to adding new proteomes by exploiting already precomputed ortholog groups, that are assumed to be stable over time. Their goal is to use fast methods, e.g. Hidden Markov Models (Eddy 2011) or k-mer based sequence similarity searches, to identify likely existing orthologous groups in which each sequence fits. While fast, these methods rely on existing databases with sufficient clade coverage to be efficient.

Another aspect of data management, linked to computational time, is the size of databases produced. Storing a high number of orthologous relations or orthologous size implies storing Terabytes of data, and induces longer access times to the data. Consequently, it is not necessarily optimal for orthology resources to include all available genomes and a choice is often made concerning which data to select, with high variability of species chosen in each orthology resource. This is reflected by the number of species available in different resources and variable representation in terms of clades or domains of life. Notably, some resources specialize in specific clades such as Plaza (Van Bel et al. 2018) for plants or FungiPath for fungi (Grossetête et al. 2010). Even among the databases with a large number of species, a wide diversity of species is preferred rather than sheer number, as diversity is generally more important than number in comparative studies (Škunca and Dessimoz 2015). This can be achieved by limiting additional species to new taxa of interest or by limiting inter-clade computations to fewer species (Nevers et al. 2019) with several levels of taxonomic resolution. The decision to add or keep a species in an existing database is a product of multiple factors but may be informed by indicators of how the addition of one species affects the diversity. For example, the rarefaction curve proposed by the KinFin analysis tool (Laetsch and Blaxter 2017) (compatible with some orthology inference software suites),

provides an objective measure of the novelties in terms of orthogroups added by each included species. Favoring diversity is also beneficial for the fast-placing strategies mentioned above, moving toward resources with a limited number of species computed directly with the all-versus-all strategy, and other species added to existing groups using less computationally intensive strategies.

3.2 Addressing proteome quality

Another aspect of high-throughput data is the associated data quality issues and the genomic data used in comparative genomics studies are no exception. Proteomes, i.e. the genome annotations of protein coding genes from genomic data, result from a multi-step process ranging from genome sequencing to the actual annotation of the final assembled sequences, with multiple possible sources of error. Consequently, a proteome may have missing proteins (either being permanent draft or a misannotated complete genome), or contain proteins that are either fragmented or actually erroneous. All of these cases may in turn induce errors in orthology inference, that rely heavily on sequence comparison and in comparative genomics approaches that assume data completeness.

Missing proteins, for instance, lead to missing orthology relationships between species with incomplete genomes and other species. Most orthology pipelines assume data completeness when inferring orthology, and while they are in principle robust to gene losses, incomplete gene sets may lead to errors in orthology inference and in orthogroup reconstruction. Some methods, e.g. Hamstr (Ebersberger et al. 2009) and OrthoGraph (Petersen et al. 2017), are designed to avoid this assumption by first excluding incomplete datasets (for example issued from RNA-seq data) during orthogroup construction. Sequences from the incomplete datasets are then mapped to the pre-computed robust orthogroups. Even with correct orthology inference, incomplete genomes impact the phylogenetic placement of species, as fewer marker genes are available. This is particularly detrimental when relations between species are hard to resolve. More spectacularly, artificially missing proteins constitute a significant source of errors for comparative genomics methods relying on comparison of entire species gene repertoires, e.g. phylogenetic profiling.

Fragmented proteins are another matter, and initially have an impact on orthology prediction *via* sequence similarity comparisons. For example, if a fragmented protein sequence corresponds to a single domain, reciprocal best hit methods may infer a false positive pairwise relation with a protein in another species having a homologous domain, although the full-length protein would not be identified as orthologous. Conversely, if the protein fragment corresponds to a low complexity, repeat-containing or divergent region, similarity based orthology prediction methods will miss it, leading to false negatives and in the worst case, may even be responsible for spurious relations (false positives). It is worth noting that issues caused by this kind of region, amplified in the presence of fragments, constitute a general limit of similarity based orthology inference methods in any organism.

A stark difference in proteome data quality is revealed by analysis of the distribution of protein length between publicly available proteomes. For example, Figure 2 shows the protein length distribution, normalized for proteome size, in four vertebrate species. Most proteomes share a distribution centered on a peak in the range of 200-400 amino acids and

a decreasing number of long proteins, as illustrated by *Homo sapiens* and *Danio rerio* (Figure 2). In contrast, some proteomes present a peak for small proteins (less than 100 amino acid long), as exemplified by the other proteomes presented on Figure 2. Strikingly, all manually curated proteomes of model species have the former distribution and both distributions are distributed across the species tree, ruling out biological exceptions (Nevers et al., *in prep*). Instead, it indicates a high number of truncated or erroneous proteins.

One must thus be cautious when providing annotations of genomic data to public databases or using these data for orthology inference and comparative genomics. Quality measures exist to indicate the quality of genome assembly, N50 being a standard indicator of genome contiguity that is commonly provided with published genome assemblies. However genome assembly quality does not necessarily correlate with proteome annotation quality. State of the art tools exist that provide an indication of data completeness and fragmentation. For instance, CEGMA (Parra et al. 2007, 2009) and its successor, BUSCO (Waterhouse et al. 2018) make use of known conserved gene families, so-called core orthologs, in single-copy in most species for the latter, to assess the completeness of the gene annotation for a given genome. The assumption being that the proportion of core orthologs found in a genome reflects the completeness of the gene annotation as a whole. BUSCO provides additional information about the state of the proteome, by indicating which proportion of core orthologs are found only in a fragmented state. Assessing BUSCO completeness is standard practice when publishing new genomes and this information is now available in UniProt (The UniProt Consortium 2019) for most available proteomes.

However, empirical data show that BUSCO completeness assessment is not always correlated to the standard protein length distribution, suggesting that it does not capture all cases of genome misannotation. A better proxy of this bias can be obtained in the form of summary statistics, such as the proportion of extremely short proteins in the genome, or the number of proteins annotated as not starting with a methionine (i.e. annotated genes for which no start codon was found by the annotation pipeline). These summary statistics can be used to filter genomes used in orthology analysis (Nevers et al. 2019), by setting thresholds under which proteomes are considered as not annotated. As these parameters are nearly orthogonal to core ortholog completeness, they can be used in parallel with methods like BUSCO and CEGMA to identify low quality proteomes. Despite these developments, work is still needed to further assess proteome quality and its impact on downstream applications, and this issue is an important target for future community efforts.

3.3 Beyond gene-level orthology

While most orthology prediction methods are based on full-length gene or protein sequences, in certain cases, functional domains might be a more pertinent entity to consider. Indeed, the majority of known proteins consist of multiple domains, especially in the Eukaryotic lineages, and it is known that multi-domain architectures tend to evolve over time as a result of different mechanisms, such as domain gains, losses and duplications, or gene fusion and fission (Buljan and Bateman 2009). The latter in particular can result in complex evolutionary histories for genes with domains of very different ancestral origins, which in turn makes orthology relations more complicated. In addition, domain architecture rearrangements have been observed several times between orthologs of species belonging to different phyla, possibly as a consequence of different organism complexity (Koonin et al.

2000, 2004). However, studies have shown that domain rearrangements can occur between relatively close species, such as mammals or members of the *Drosophila* genus, and it has been estimated that they could concern up to 50% of proteins (Forslund et al. 2011; Wu et al. 2012; Sonnhammer et al. 2014).

Divergences of domain content and/or order between orthologs can be challenging for traditional orthology inference methods. In some cases, parts of the protein sequence might be too highly divergent in some species to be properly detected as orthologs. In other cases one protein might have significant similarity to multiple different protein families, each due to a different domain of the query protein, making it hard to clearly establish orthologous relations. This shows a clear limitation of full-length analyses, as they ignore the natural tendency of proteins to be modular and to evolve not at the complete sequence level, but at the domain level. It would be beneficial to focus future improvements and developments on domain-aware orthology inference as a complement to full-length methods, in order to predict more precise ortholog relations and better understand architectural rearrangements in protein evolution. While it has been widely acknowledged that such methods are needed (Sjolander et al. 2011), very few currently take domains into account. Exceptions include the microbial genome database MBGD, which constructs ortholog groups at the domain level (Uchiyama et al. 2019), and Domainoid (Persson et al. 2019), a tool that uses Pfam (El-Gebali et al. 2019) defined domains to infer orthology relations at the single level domain. Domainoid has been shown to retrieve orthologs not detected by classical full-length approaches, thus showing the interest of combining both types of strategies.

Another hassle of focusing on gene level orthology is that, in Eukaryotes, a single gene may be transcribed into several isoforms with different exons combinations. This process, called alternative splicing, is especially prominent in Vertebrates (Keren et al. 2010). Its functional implication is debated, but it has been shown for particular genes that different isoforms may have different tissue expression and even sometimes produce proteins with antagonist cellular functions (Wang et al. 2008). This has direct implications on the way orthology is used to transfer function between genes, as two orthologous genes could display different splicing patterns and even two orthologous genes with orthologous exons may have substantially different transcripts. Integrating homology between alternative transcripts of orthologs will provide additional information on whether an evolutionary conserved isoform is more likely to be functional, and whether observations made in a model species on a particular isoform are likely to be applicable to other species.

Assessing orthology between alternative transcripts often relies on two conditions (Blanquart et al. 2016). Indeed, transcripts are orthologous if (1) they are transcripts of orthologous genes and (2) their exons are similar enough to assume they are orthologous and appear in the same order in the gene sequence. The first condition is a classical orthology inference problem. The second condition may be determined by spliced sequence alignment, using an exon-aware alignment method (Kapustin et al. 2008; Gotoh 2008; Sharma et al. 2016; Jammali et al. 2019). Transcript orthology prediction has been successfully employed to identify orthologous isoforms between the gene repertoires of mouse and human (Zambelli et al. 2010). Applying it to more species is trickier since it cannot be done with pairwise relations and requires the construction of gene trees, which is computationally demanding. Nonetheless, it has been used to study multiple gene families, mapping events of isoform gains and losses to the branches of the trees (Christinat and Moret 2012; Jammali et al.

2019). Nevertheless, one must still be cautious when using isoform orthology determination, and ensure that expression of both isoforms can be detected through experimental means in the species of interest, to avoid the pitfalls of erroneous annotation transfer.

As can be seen, despite the major advances made in recent years in orthology inference and resources, there is still a long way to go in the quest for orthologs. The practical and conceptual challenges are numerous and will require the efforts of the entire comparative genomics community to invent new solutions. Substantial progress will be needed both in the development of new indicators of proteome quality and for the formal representation of orthology relationships at different granularity levels.

Acknowledgments

The authors thank Julie Thompson for critical reading of the manuscript. The authors are also grateful to the anonymous referees for their useful suggestions.

References

Adzhubei IA, Schmidt S, Peshkin L, et al (2010) A method and server for predicting damaging missense mutations. Nat Methods 7:248–249. https://doi.org/10.1038/nmeth0410-248

Alliance of Genome Resources Consortium (2020) Alliance of Genome Resources Portal: unified model organism research platform. Nucleic Acids Res 48:D650–D658. https://doi.org/10.1093/nar/gkz813

Altenhoff AM, Boeckmann B, Capella-Gutierrez S, et al (2016) Standardized benchmarking in the quest for orthologs. Nat Methods 13:425–430. https://doi.org/10.1038/nmeth.3830

Altenhoff AM, Gil M, Gonnet GH, Dessimoz C (2013) Inferring hierarchical orthologous groups from orthologous gene pairs. PloS One 8:e53786. https://doi.org/10.1371/journal.pone.0053786

Altenhoff AM, Glover NM, Dessimoz C (2019) Inferring Orthology and Paralogy. In: Anisimova M (ed) Evolutionary Genomics: Statistical and Computational Methods. Springer, New York, NY, pp 149–175

Altenhoff AM, Glover NM, Train C-M, et al (2018) The OMA orthology database in 2018: retrieving evolutionary relationships among all domains of life through richer web and programmatic interfaces. Nucleic Acids Res 46:D477–D485. https://doi.org/10.1093/nar/gkx1019

Altenhoff AM, Studer RA, Robinson-Rechavi M, Dessimoz C (2012) Resolving the ortholog conjecture: orthologs tend to be weakly, but significantly, more similar in function than paralogs. PLoS Comput Biol 8:e1002514. https://doi.org/10.1371/journal.pcbi.1002514

Altschul SF, Gish W, Miller W, et al (1990) Basic local alignment search tool. J Mol Biol 215:403–410. https://doi.org/10.1016/S0022-2836(05)80360-2

Altschul SF, Madden TL, Schäffer AA, et al (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25:3389–3402. https://doi.org/10.1093/nar/25.17.3389

Amar D, Frades I, Danek A, et al (2014) Evaluation and integration of functional annotation pipelines for newly sequenced organisms: the potato genome as a test case. BMC Plant Biol 14:329. https://doi.org/10.1186/s12870-014-0329-9
Ashburner M, Ball CA, Blake JA, et al (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 25:25–29. https://doi.org/10.1038/75556
Blanquart S, Varré J-S, Guertin P, et al (2016) Assisted transcriptome reconstruction and splicing orthology. BMC Genomics 17:786. https://doi.org/10.1186/s12864-016-3103-6
Brockhurst MA, Harrison E, Hall JPJ, et al (2019) The Ecology and Evolution of Pangenomes. Curr Biol CB 29:R1094–R1103. https://doi.org/10.1016/j.cub.2019.08.012
Buchfink B, Xie C, Huson DH (2015) Fast and sensitive protein alignment using DIAMOND. Nat Methods 12:59–60. https://doi.org/10.1038/nmeth.3176
Buljan M, Bateman A (2009) The evolution of protein domain families. Biochem Soc Trans 37:751–755. https://doi.org/10.1042/BST0370751
Burki F, Roger AJ, Brown MW, Simpson AGB (2020) The New Tree of Eukaryotes. Trends Ecol Evol 35:43–55. https://doi.org/10.1016/j.tree.2019.08.008
Camacho C, Coulouris G, Avagyan V, et al (2009) BLAST+: architecture and applications. BMC Bioinformatics 10:421. https://doi.org/10.1186/1471-2105-10-421
Castillo-Davis CI, Kondrashov FA, Hartl DL, Kulathinal RJ (2004) The functional genomic distribution of protein divergence in two animal phyla: coevolution, genomic conflict, and constraint. Genome Res 14:802–811. https://doi.org/10.1101/gr.2195604
Chen F, Mackey AJ, Stoeckert CJ, Roos DS (2006) OrthoMCL-DB: querying a comprehensive multi-species collection of ortholog groups. Nucleic Acids Res 34:D363-368. https://doi.org/10.1093/nar/gkj123
Chen X, Zhang J (2012) The ortholog conjecture is untestable by the current gene ontology but is supported by RNA sequencing data. PLoS Comput Biol 8:e1002784. https://doi.org/10.1371/journal.pcbi.1002784
Christinat Y, Moret BME (2012) Inferring transcript phylogenies. BMC Bioinformatics 13 Suppl 9:S1. https://doi.org/10.1186/1471-2105-13-s9-s1
Crawford NG, Faircloth BC, McCormack JE, et al (2012) More than 1000 ultraconserved elements provide evidence that turtles are the sister group of archosaurs. Biol Lett 8:783–786. https://doi.org/10.1098/rsbl.2012.0331
Darby CA, Stolzer M, Ropp PJ, et al (2017) Xenolog classification. Bioinforma Oxf Engl 33:640–649. https://doi.org/10.1093/bioinformatics/btw686
Ebersberger I, Strauss S, von Haeseler A (2009) HaMStR: profile hidden markov model based search for orthologs in ESTs. BMC Evol Biol 9:157. https://doi.org/10.1186/1471-2148-9-157
Eddy SR (2011) Accelerated Profile HMM Searches. PLoS Comput Biol 7:e1002195. https://doi.org/10.1371/journal.pcbi.1002195
El-Gebali S, Mistry J, Bateman A, et al (2019) The Pfam protein families database in 2019. Nucleic Acids Res 47:D427–D432. https://doi.org/10.1093/nar/gky995
Emms DM, Kelly S (2018) STAG: Species Tree Inference from All Genes. bioRxiv 267914. https://doi.org/10.1101/267914
Emms DM, Kelly S (2015) OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. Genome Biol 16:157. https://doi.org/10.1186/s13059-015-0721-2
Emms DM, Kelly S (2019) OrthoFinder: phylogenetic orthology inference for comparative genomics. Genome Biol 20:238. https://doi.org/10.1186/s13059-019-1832-y
Feng B, Zhou L, Tang J (2017) Ancestral Genome Reconstruction on Whole Genome Level. Curr Genomics 18:306–315. https://doi.org/10.2174/1389202918666170307120943
Fernández-Breis JT, Chiba H, Legaz-García MDC, Uchiyama I (2016) The Orthology Ontology: development and applications. J Biomed Semant 7:34. https://doi.org/10.1186/s13326-016-0077-x

Fitch WM (1970) Distinguishing homologous from analogous proteins. Syst Zool 19:99–113
Fitch WM (2000) Homology a personal view on some of the problems. Trends Genet TIG 16:227–231. https://doi.org/10.1016/s0168-9525(00)02005-9
Force A, Lynch M, Pickett FB, et al (1999) Preservation of duplicate genes by complementary, degenerative mutations. Genetics 151:1531–1545
Forslund K, Pekkari I, Sonnhammer ELL (2011) Domain architecture conservation in orthologs. BMC Bioinformatics 12:326. https://doi.org/10.1186/1471-2105-12-326
Gabaldón T, Dessimoz C, Huxley-Jones J, et al (2009) Joining forces in the quest for orthologs. Genome Biol 10:403. https://doi.org/10.1186/gb-2009-10-9-403
Gaucher EA, Govindarajan S, Ganesh OK (2008) Palaeotemperature trend for Precambrian life inferred from resurrected proteins. Nature 451:704–707. https://doi.org/10.1038/nature06510
Gaudet P, Livstone MS, Lewis SE, Thomas PD (2011) Phylogenetic-based propagation of functional annotations within the Gene Ontology consortium. Brief Bioinforma 12:449–462
Gilks WR, Audit B, De Angelis D, et al (2002) Modeling the percolation of annotation errors in a database of protein sequences. Bioinforma Oxf Engl 18:1641–1649. https://doi.org/10.1093/bioinformatics/18.12.1641
Glover NM, Redestig H, Dessimoz C (2016) Homoeologs: What Are They and How Do We Infer Them? Trends Plant Sci 21:609–621. https://doi.org/10.1016/j.tplants.2016.02.005
Gotoh O (2008) Direct mapping and alignment of protein sequences onto genomic sequence. Bioinforma Oxf Engl 24:2438–2444. https://doi.org/10.1093/bioinformatics/btn460
Gray GS, Fitch WM (1983) Evolution of antibiotic resistance genes: the DNA sequence of a kanamycin resistance gene from Staphylococcus aureus. Mol Biol Evol 1:57–66. https://doi.org/10.1093/oxfordjournals.molbev.a040298
Grossetête S, Labedan B, Lespinet O (2010) FUNGIpath: a tool to assess fungal metabolic pathways predicted by orthology. BMC Genomics 11:81. https://doi.org/10.1186/1471-2164-11-81
Gumulya Y, Gillam EMJ (2017) Exploring the past and the future of protein evolution with ancestral sequence reconstruction: the “retro” approach to protein engineering. Biochem J 474:1–19. https://doi.org/10.1042/BCJ20160507
Hecker N, Sharma V, Hiller M (2019) Convergent gene losses illuminate metabolic and physiological changes in herbivores and carnivores. Proc Natl Acad Sci 116:3036–3041. https://doi.org/10.1073/pnas.1818504116
Henricson A, Forslund K, Sonnhammer ELL (2010) Orthology confers intron position conservation. BMC Genomics 11:412. https://doi.org/10.1186/1471-2164-11-412
Herrero J, Muffato M, Beal K, et al (2016) Ensembl comparative genomics resources. Database J Biol Databases Curation 2016:. https://doi.org/10.1093/database/baw053
Hu X, Friedberg I (2019) SwiftOrtho: A fast, memory-efficient, multiple genome orthology classifier. GigaScience 8:. https://doi.org/10.1093/gigascience/giz118
Hu Y, Flockhart I, Vinayagam A, et al (2011) An integrative approach to ortholog prediction for disease-focused and other functional studies. BMC Bioinformatics 12:357. https://doi.org/10.1186/1471-2105-12-357
Huang X, Albou L-P, Mushayahama T, et al (2019) Ancestral Genomes: a resource for reconstructed ancestral genes and genomes across the tree of life. Nucleic Acids Res 47:D271–D279. https://doi.org/10.1093/nar/gky1009
Huerta-Cepas J, Capella-Gutiérrez S, Pryszcz LP, et al (2014) PhylomeDB v4: zooming into the plurality of evolutionary histories of a genome. Nucleic Acids Res 42:D897-902. https://doi.org/10.1093/nar/gkt1177
Huerta-Cepas J, Dopazo H, Dopazo J, Gabaldón T (2007) The human phylome. Genome Biol 8:R109. https://doi.org/10.1186/gb-2007-8-6-r109
Huerta-Cepas J, Forslund K, Coelho LP, et al (2017) Fast Genome-Wide Functional Annotation through Orthology Assignment by eggNOG-Mapper. Mol Biol Evol

34:2115–2122. https://doi.org/10.1093/molbev/msx148
Huerta-Cepas J, Szklarczyk D, Forslund K, et al (2016) eggNOG 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences. Nucleic Acids Res 44:D286-293. https://doi.org/10.1093/nar/gkv1248
Hug LA, Baker BJ, Anantharaman K, et al (2016) A new view of the tree of life. Nat Microbiol 1:16048. https://doi.org/10.1038/nmicrobiol.2016.48
Huynen M, Dandekar T, Bork P (1998) Differential genome analysis applied to the species-specific features of Helicobacter pylori. FEBS Lett 426:1–5. https://doi.org/10.1016/s0014-5793(98)00276-2
Jammali S, Aguilar J-D, Kuitche E, Ouangraoua A (2019) SplicedFamAlign: CDS-to-gene spliced alignment and identification of transcript orthology groups. BMC Bioinformatics 20:133. https://doi.org/10.1186/s12859-019-2647-2
Jensen LJ, Julien P, Kuhn M, et al (2008) eggNOG: automated construction and annotation of orthologous groups of genes. Nucleic Acids Res 36:D250-254. https://doi.org/10.1093/nar/gkm796
Jim K, Parmar K, Singh M, Tavazoie S (2004) A cross-genomic approach for systematic mapping of phenotypic traits to genes. Genome Res 14:109–115. https://doi.org/10.1101/gr.1586704
Kapustin Y, Souvorov A, Tatusova T, Lipman D (2008) Splign: algorithms for computing spliced alignments with identification of paralogs. Biol Direct 3:20. https://doi.org/10.1186/1745-6150-3-20
Kensche PR, van Noort V, Dutilh BE, Huynen MA (2008) Practical and theoretical advances in predicting the function of a protein by its phylogenetic distribution. J R Soc Interface 5:151–170. https://doi.org/10.1098/rsif.2007.1047
Kent WJ, Baertsch R, Hinrichs A, et al (2003) Evolution's cauldron: duplication, deletion, and rearrangement in the mouse and human genomes. Proc Natl Acad Sci U S A 100:11484–11489. https://doi.org/10.1073/pnas.1932072100
Keren H, Lev-Maor G, Ast G (2010) Alternative splicing and evolution: diversification, exon definition and function. Nat Rev Genet 11:345–355. https://doi.org/10.1038/nrg2776
Khan FI, Wei D-Q, Gu K-R, et al (2016) Current updates on computer aided protein modeling and designing. Int J Biol Macromol 85:48–62. https://doi.org/10.1016/j.ijbiomac.2015.12.072
Koonin EV, Aravind L, Kondrashov AS (2000) The Impact of Comparative Genomics on Our Understanding of Evolution. Cell 101:573–576. https://doi.org/10.1016/S0092-8674(00)80867-3
Koonin EV, Fedorova ND, Jackson JD, et al (2004) A comprehensive evolutionary classification of proteins encoded in complete eukaryotic genomes. Genome Biol 5:R7
Kriventseva EV, Kuznetsov D, Tegenfeldt F, et al (2019) OrthoDB v10: sampling the diversity of animal, plant, fungal, protist, bacterial and viral genomes for evolutionary and functional annotations of orthologs. Nucleic Acids Res 47:D807–D811. https://doi.org/10.1093/nar/gky1053
Kryuchkova-Mostacci N, Robinson-Rechavi M (2015) Tissue-Specific Evolution of Protein Coding Genes in Human and Mouse. PloS One 10:e0131673. https://doi.org/10.1371/journal.pone.0131673
Laetsch DR, Blaxter ML (2017) KinFin: Software for Taxon-Aware Analysis of Clustered Protein Sequences. G3 Bethesda Md 7:3349–3357. https://doi.org/10.1534/g3.117.300233
Land M, Hauser L, Jun S-R, et al (2015) Insights from 20 years of bacterial genome sequencing. Funct Integr Genomics 15:141–161. https://doi.org/10.1007/s10142-015-0433-4
Lapierre P, Gogarten JP (2009) Estimating the size of the bacterial pan-genome. Trends Genet TIG 25:107–110. https://doi.org/10.1016/j.tig.2008.12.004
Lecompte O, Thompson JD, Plewniak F, et al (2001) Multiple alignment of complete

sequences (MACS) in the post-genomic era. Gene 270:17–30. https://doi.org/10.1016/s0378-1119(01)00461-9
Leelananda SP, Lindert S (2016) Computational methods in drug discovery. Beilstein J Org Chem 12:2694–2718. https://doi.org/10.3762/bjoc.12.267
Li L, Stoeckert CJ, Roos DS (2003) OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res 13:2178–2189. https://doi.org/10.1101/gr.1224503
Linard B, Thompson JD, Poch O, Lecompte O (2011) OrthoInspector: comprehensive orthology analysis and visual exploration. BMC Bioinformatics 12:11. https://doi.org/10.1186/1471-2105-12-11
Liu T, Tang GW, Capriotti E (2011) Comparative modeling: the state of the art and protein drug target structure prediction. Comb Chem High Throughput Screen 14:532–547. https://doi.org/10.2174/138620711795767811
Mi H, Dong Q, Muruganujan A, et al (2010) PANTHER version 7: improved phylogenetic trees, orthologs and collaboration with the Gene Ontology Consortium. Nucleic Acids Res 38:D204-210. https://doi.org/10.1093/nar/gkp1019
Mi H, Muruganujan A, Ebert D, et al (2019) PANTHER version 14: more genomes, a new PANTHER GO-slim and improvements in enrichment analysis tools. Nucleic Acids Res 47:D419–D426. https://doi.org/10.1093/nar/gky1038
Mukherjee S, Stamatis D, Bertsch J, et al (2019) Genomes OnLine database (GOLD) v.7: updates and new features. Nucleic Acids Res 47:D649–D659. https://doi.org/10.1093/nar/gky977
NCBI Resource Coordinators (2016) Database resources of the National Center for Biotechnology Information. Nucleic Acids Res 44:D7-19. https://doi.org/10.1093/nar/gkv1290
Nehrt NL, Clark WT, Radivojac P, Hahn MW (2011) Testing the ortholog conjecture with comparative functional genomic data from mammals. PLoS Comput Biol 7:e1002073. https://doi.org/10.1371/journal.pcbi.1002073
Nevers Y, Kress A, Defosset A, et al (2019) OrthoInspector 3.0: open portal for comparative genomics. Nucleic Acids Res 47:D411–D418. https://doi.org/10.1093/nar/gky1068
Nevers Y, Prasad MK, Poidevin L, et al (2017) Insights into Ciliary Genes and Evolution from Multi-Level Phylogenetic Profiling. Mol Biol Evol 34:2016–2034. https://doi.org/10.1093/molbev/msx146
Overbeek R, Fonstein M, D'Souza M, et al (1999) The use of gene clusters to infer functional coupling. Proc Natl Acad Sci U S A 96:2896–2901. https://doi.org/10.1073/pnas.96.6.2896
Parra G, Bradnam K, Korf I (2007) CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. Bioinforma Oxf Engl 23:1061–1067. https://doi.org/10.1093/bioinformatics/btm071
Parra G, Bradnam K, Ning Z, et al (2009) Assessing the gene space in draft genomes. Nucleic Acids Res 37:289–297. https://doi.org/10.1093/nar/gkn916
Pellegrini M, Marcotte EM, Thompson MJ, et al (1999) Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. Proc Natl Acad Sci U S A 96:4285–4288. https://doi.org/10.1073/pnas.96.8.4285
Penel S, Arigon A-M, Dufayard J-F, et al (2009) Databases of homologous gene families for comparative genomics. BMC Bioinformatics 10 Suppl 6:S3. https://doi.org/10.1186/1471-2105-10-S6-S3
Pereira C, Denise A, Lespinet O (2014) A meta-approach for improving the prediction and the functional annotation of ortholog groups. BMC Genomics 15 Suppl 6:S16. https://doi.org/10.1186/1471-2164-15-S6-S16
Persson E, Kaduk M, Forslund SK, Sonnhammer ELL (2019) Domainoid: domain-oriented orthology inference. BMC Bioinformatics 20:523. https://doi.org/10.1186/s12859-019-3137-2
Petersen M, Meusemann K, Donath A, et al (2017) Orthograph: a versatile tool for mapping coding nucleotide sequences to clusters of orthologous genes. BMC Bioinformatics

18:111. https://doi.org/10.1186/s12859-017-1529-8
Peterson ME, Chen F, Saven JG, et al (2009) Evolutionary constraints on structural similarity in orthologs and paralogs. Protein Sci Publ Protein Soc 18:1306–1315. https://doi.org/10.1002/pro.143
Pryszcz LP, Huerta-Cepas J, Gabaldón T (2011) MetaPhOrs: orthology and paralogy predictions from multiple phylogenetic evidence using a consistency-based confidence score. Nucleic Acids Res 39:e32. https://doi.org/10.1093/nar/gkq953
Rane RV, Oakeshott JG, Nguyen T, et al (2017) Orthonome - a new pipeline for predicting high quality orthologue gene sets applicable to complete and draft genomes. BMC Genomics 18:673. https://doi.org/10.1186/s12864-017-4079-6
Remm M, Storm CE, Sonnhammer EL (2001) Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. J Mol Biol 314:1041–1052. https://doi.org/10.1006/jmbi.2000.5197
Roth ACJ, Gonnet GH, Dessimoz C (2008) Algorithm of OMA for large-scale orthology inference. BMC Bioinformatics 9:518. https://doi.org/10.1186/1471-2105-9-518
Rouard M, Guignon V, Aluome C, et al (2011) GreenPhylDB v2.0: comparative and functional genomics in plants. Nucleic Acids Res 39:D1095-1102. https://doi.org/10.1093/nar/gkq811
Santpere G, Lopez-Valenzuela M, Petit-Marty N, et al (2016) Differences in molecular evolutionary rates among microRNAs in the human and chimpanzee genomes. BMC Genomics 17:528. https://doi.org/10.1186/s12864-016-2863-3
Schmitt T, Messina DN, Schreiber F, Sonnhammer ELL (2011) Letter to the editor: SeqXML and OrthoXML: standards for sequence and orthology information. Brief Bioinform 12:485–488. https://doi.org/10.1093/bib/bbr025
Schnoes AM, Brown SD, Dodevski I, Babbitt PC (2009) Annotation error in public databases: misannotation of molecular function in enzyme superfamilies. PLoS Comput Biol 5:e1000605. https://doi.org/10.1371/journal.pcbi.1000605
Schreiber F, Patricio M, Muffato M, et al (2014) TreeFam v9: a new website, more species and orthology-on-the-fly. Nucleic Acids Res 42:D922-925. https://doi.org/10.1093/nar/gkt1055
Schreiber F, Sonnhammer ELL (2013) Hieranoid: hierarchical orthology inference. J Mol Biol 425:2072–2081. https://doi.org/10.1016/j.jmb.2013.02.018
Schwartz S, Kent WJ, Smit A, et al (2003) Human-mouse alignments with BLASTZ. Genome Res 13:103–107. https://doi.org/10.1101/gr.809403
Sharma V, Elghafari A, Hiller M (2016) Coding exon-structure aware realigner (CESAR) utilizes genome alignments for accurate comparative gene annotation. Nucleic Acids Res 44:e103. https://doi.org/10.1093/nar/gkw210
Sjolander K, Datta RS, Shen Y, Shoffner GM (2011) Ortholog identification in the presence of domain architecture rearrangement. Brief Bioinform 12:413–422. https://doi.org/10.1093/bib/bbr036
Škunca N, Dessimoz C (2015) Phylogenetic profiling: how much input data is enough? PloS One 10:e0114701. https://doi.org/10.1371/journal.pone.0114701
Smith TF, Waterman MS (1981) Identification of common molecular subsequences. J Mol Biol 147:195–197. https://doi.org/10.1016/0022-2836(81)90087-5
Sonnhammer ELL, Gabaldón T, Sousa da Silva AW, et al (2014) Big data and other challenges in the quest for orthologs. Bioinforma Oxf Engl 30:2993–2998. https://doi.org/10.1093/bioinformatics/btu492
Sonnhammer ELL, Koonin EV (2002) Orthology, paralogy and proposed classification for paralog subtypes. Trends Genet TIG 18:619–620. https://doi.org/10.1016/s0168-9525(02)02793-2
Sonnhammer ELL, Östlund G (2015) InParanoid 8: orthology analysis between 273 proteomes, mostly eukaryotic. Nucleic Acids Res 43:D234-239. https://doi.org/10.1093/nar/gku1203
Soucy SM, Huang J, Gogarten JP (2015) Horizontal gene transfer: building the web of life. Nat Rev Genet 16:472–482. https://doi.org/10.1038/nrg3962

Spang A, Saw JH, Jørgensen SL, et al (2015) Complex archaea that bridge the gap between prokaryotes and eukaryotes. Nature 521:173–179. https://doi.org/10.1038/nature14447
Steinegger M, Söding J (2017) MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. Nat Biotechnol 35:1026–1028. https://doi.org/10.1038/nbt.3988
Storm CEV, Sonnhammer ELL (2002) Automated ortholog inference from phylogenetic trees and calculation of orthology reliability. Bioinforma Oxf Engl 18:92–99. https://doi.org/10.1093/bioinformatics/18.1.92
Straub K, Merkl R (2019) Ancestral Sequence Reconstruction as a Tool for the Elucidation of a Stepwise Evolutionary Adaptation. Methods Mol Biol Clifton NJ 1851:171–182. https://doi.org/10.1007/978-1-4939-8736-8_9
Studer RA, Robinson-Rechavi M (2009) How confident can we be that orthologs are similar, but paralogs differ? Trends Genet TIG 25:210–216. https://doi.org/10.1016/j.tig.2009.03.004
Sutphin GL, Mahoney JM, Sheppard K, et al (2016) WORMHOLE: Novel Least Diverged Ortholog Prediction through Machine Learning. PLoS Comput Biol 12:e1005182. https://doi.org/10.1371/journal.pcbi.1005182
Szklarczyk D, Gable AL, Lyon D, et al (2019) STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. Nucleic Acids Res 47:D607–D613. https://doi.org/10.1093/nar/gky1131
Tabach Y, Golan T, Hernández-Hernández A, et al (2013) Human disease locus discovery and mapping to molecular pathways through phylogenetic profiling. Mol Syst Biol 9:692. https://doi.org/10.1038/msb.2013.50
Tatusov RL, Koonin EV, Lipman DJ (1997) A genomic perspective on protein families. Science 278:631–637. https://doi.org/10.1126/science.278.5338.631
The UniProt Consortium (2019) UniProt: a worldwide hub of protein knowledge. Nucleic Acids Res 47:D506–D515. https://doi.org/10.1093/nar/gky1049
The Gene Ontology Consortium (2019) The Gene Ontology Resource: 20 years and still GOing strong. Nucleic Acids Res 47:D330–D338. https://doi.org/10.1093/nar/gky1055
Uchiyama I, Mihara M, Nishide H, et al (2019) MBGD update 2018: microbial genome database based on hierarchical orthology relations covering closely related and distantly related comparisons. Nucleic Acids Res 47:D382–D389. https://doi.org/10.1093/nar/gky1054
Van Bel M, Diels T, Vancaester E, et al (2018) PLAZA 4.0: an integrative resource for functional, evolutionary and comparative plant genomics. Nucleic Acids Res 46:D1190–D1196. https://doi.org/10.1093/nar/gkx1002
Van de Peer Y, Mizrachi E, Marchal K (2017) The evolutionary significance of polyploidy. Nat Rev Genet 18:411–424. https://doi.org/10.1038/nrg.2017.26
Vaser R, Adusumalli S, Leng SN, et al (2016) SIFT missense predictions for genomes. Nat Protoc 11:1–9. https://doi.org/10.1038/nprot.2015.123
Vilella AJ, Severin J, Ureta-Vidal A, et al (2009) EnsemblCompara GeneTrees: Complete, duplication-aware phylogenetic trees in vertebrates. Genome Res 19:327–335. https://doi.org/10.1101/gr.073585.107
Walhout AJ, Boulton SJ, Vidal M (2000) Yeast two-hybrid systems and protein interaction mapping projects for yeast and worm. Yeast Chichester Engl 17:88–94. https://doi.org/10.1002/1097-0061(20000630)17:2<88::AID-YEA20>3.0.CO;2-Y
Wang ET, Sandberg R, Luo S, et al (2008) Alternative isoform regulation in human tissue transcriptomes. Nature 456:470–476. https://doi.org/10.1038/nature07509
Waterhouse RM, Seppey M, Simão FA, et al (2018) BUSCO Applications from Quality Assessments to Gene Prediction and Phylogenomics. Mol Biol Evol 35:543–548. https://doi.org/10.1093/molbev/msx319
Whiteside MD, Winsor GL, Laird MR, Brinkman FSL (2013) OrthologeDB: a bacterial and

archaeal orthology resource for improved comparative genomic analysis. Nucleic Acids Res 41:D366-376. https://doi.org/10.1093/nar/gks1241
Wolfe K (2000) Robustness—it's not where you think it is. Nat Genet 25:3–4. https://doi.org/10.1038/75560
Wu Y-C, Rasmussen MD, Kellis M (2012) Evolution at the Subgene Level: Domain Rearrangements in the Drosophila Phylogeny. Mol Biol Evol 29:689–705. https://doi.org/10.1093/molbev/msr222
Zambelli F, Pavesi G, Gissi C, et al (2010) Assessment of orthologous splicing isoforms in human and mouse orthologous genes. BMC Genomics 11:534. https://doi.org/10.1186/1471-2164-11-534
Zaremba-Niedzwiedzka K, Caceres EF, Saw JH, et al (2017) Asgard archaea illuminate the origin of eukaryotic cellular complexity. Nature 541:353–358. https://doi.org/10.1038/nature21031
Zmasek CM, Eddy SR (2002) RIO: analyzing proteomes by automated phylogenomics using resampled inference of orthologs. BMC Bioinformatics 3:14. https://doi.org/10.1186/1471-2105-3-14

Figures

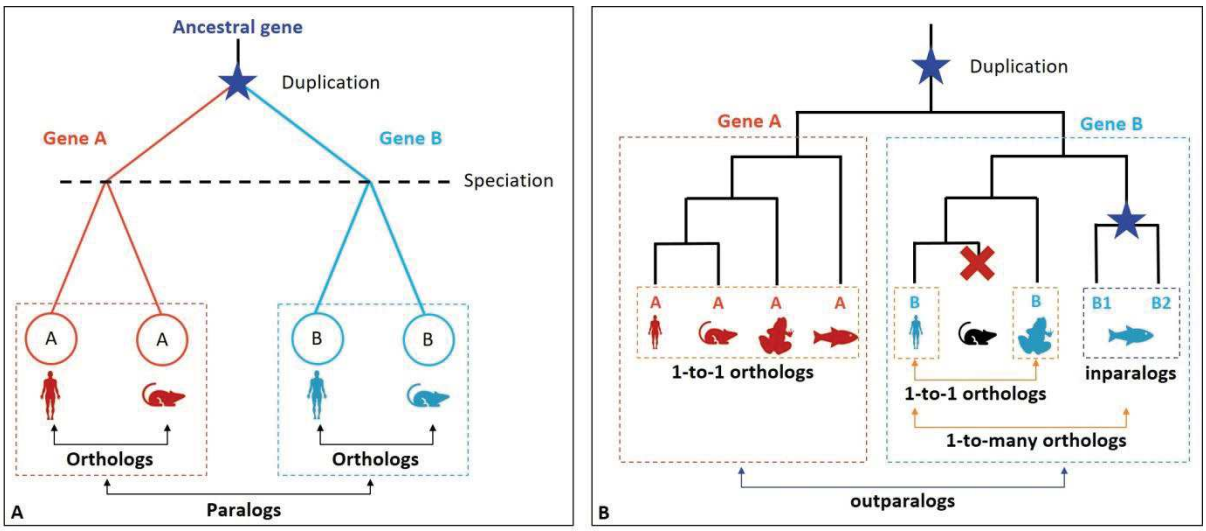

Figure 1. Homology relationships **a.** Evolutionary history of a gene family with duplication and speciation events. Genes A (in red) present in humans and mouse emerged after a speciation event, they are orthologous to each other. The same is true for genes B (in blue). Genes A and B are paralogous between each other because they are separated by a duplication event in their evolutionary history. **b.** Genes A (in red) are only separated by speciation events, they are 1-to-1 orthologs. The evolutionary history of genes B (in blue) is more complex with a lineage-specific loss in mouse and a 'recent' duplication in fish. Considering the evolutionary history of vertebrates, genes B1 and B2 are inparalogs to each other and co-orthologs to the human gene B. Thus, there is a 1-to-many orthology relation between the human gene B and the fish genes B1 and B2 genes. Considering Vertebrates, genes A and B are outparalogs between each other because they emerged after a duplication that occurred in the vertebrate ancestor, i.e. before speciations.

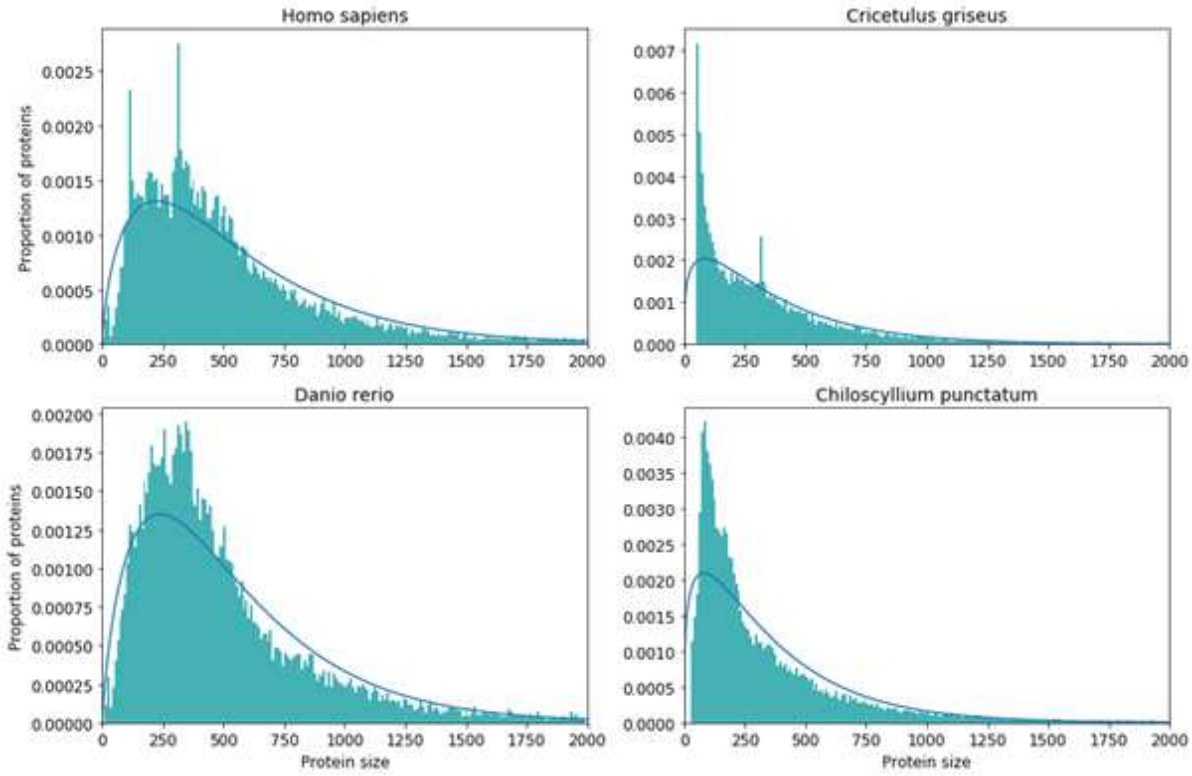

Figure 2. Protein length distribution in four proteomes, from various Vertebrate clades. On the left are examples of the distribution observed in well studied species (*Homo sapiens* and *Danio rerio*), similar to the one observed in most proteomes. On the left, examples of atypical distributions with high number of small proteins for the rodent *Cricetulus griseus* and the chondrichthyes *Chiloscyllium punctatum*.

Table

Table 1. Main orthology resources

Resource		Coverage					Exploration							Representation						
Type	Name	Genomes	Bacteria	Eukaryota	Archaea	Viruses	Gene Id	Group Id	Sequence	Function	Distribution	SPARQL	Webservice	Orthologues	Function	Domains	MSA	Tree	Synteny	Distribution
General	Inparanoid	273	/	/	/	0	✓	✗	✓	✓	✗	✗	✗	✓	✗	✗	✗	✗	✗	✗
	OMA	2 327	1688	485	154	0	✓	✓	✓	✓	✗	✗	✓	✓	✓	✓	✓	✗	✓	✓
	EggNOG	2 031	1678	115	238	352	✓	✓	✓	✗	✗	✗	✓	✓	✓	✓	✓	✓	✗	✓
	OrthoDb	7284	5609	1271	404	7963	✓	✗	✓	✓	✓	✓	✓	✓	✓	✓	✗	✗	✗	✗
	OrthoMCL	150	36	98	16	0	✓	✓	✓	✓	✓	✗	✓	✓	✓	✓	✓	✗	✗	✓
	Hieranoid	66	20	40	6	0	✓	✗	✓	✓	✓	✗	✗	✗	✗	✗	✓	✓	✗	✗
	OrthoInspector	4753	3863	711	179	0	✓	✗	✓	✓	✓	✗	✓	✓	✓	✗	✓	✗	✗	✓
	MBGD	6318	5861	203	254	0	✓	✗	✓	✓	✓	✓	✗	✓	✓	✗	✓	✓	✓	✓
	OtholugeDb	2069	/	0	/	0	✓	✗	✗	✗	✓	✗	✗	✓	✗	✗	✗	✗	✓	✗
	HOGENOM	13367	12326	593	224	0	✓	✓	✓	✗	✗	✗	✗	✓	✗	✗	✓	✓	✗	✗
	PhylomeDb	1 862	/	/	/	0	✓	✗	✓	✗	✗	✗	✗	✓	✗	✗	✓	✓	✗	✗
Specific	TreeFam	109	0	109	0	0	✓	✗	✓	✗	✗	✗	✗	✓	✓	✓	✓	✓	✗	✓
	FungiPath	165	0	165	0	0	✓	✓	✓	✓	✓	✗	✗	✓	✓	✗	✓	✓	✗	✗
	Greenphyl	37	0	37	0	0	✓	✓	✓	✓	✓	✗	✗	✓	✓	✓	✓	✓	✗	✓
	PLAZA	119	0	119	0	0	✓	✗	✓	✓	✓	✗	✗	✓	✓	✓	✓	✓	✓	✓
Meta-predictions	P-POD	12	1	11	0	0	✓	✗	✗	✓	✗	✗	✗	✓	✓	✗	✗	✓	✗	✓
	MetaPhOrs	2713	1 720	877	116	1	✓	✗	✓	✗	✗	✗	✗	✓	✓	✗	✗	✗	✗	✗
	WORMHOLE	6	0	6	0	0	✓	✗	✗	✗	✗	✗	✗	✓	✗	✗	✗	✗	✗	✗
	DIOPT	10	0	10	0	0	✓	✗	✗	✗	✗	✗	✗	✓	✗	✗	✗	✗	✗	✗
	YOGY	11	1	10	0	0	✓	✗	✗	✗	✗	✗	✗	✓	✗	✗	✗	✗	✗	✓
	HCOP	19	0	19	0	0	✓	✗	✗	✗	✗	✗	✗	✓	✗	✗	✗	✗	✗	✗
Other	Panther	142	35	99	8	0	✓	✗	✗	✗	✗	✗	✓	✓	✗	✓	✗	✓	✗	✗
	Ensembl	1191	123*	1068	/	0	✓	✗	✗	✗	✗	✗	✓	✓	✓	✓	✓	✓	✓	✓
	Homologene	21	0	21	0	0	✓	✗	✗	✗	✗	✗	✗	✓	✓	✗	✗	✗	✗	✓

References: EggNog (Huerta-Cepas et al. 2016), HOGENOM ((Penel et al. 2009)), Inparanoid (Sonnhammer and Östlund 2015), MBGD (Uchiyama et al. 2019), OMA (Altenhoff et al. 2018), OrthoDb (Kriventseva et al. 2019), OrthoInspector (Nevers et al. 2019), OrtholugeDB (Whiteside et al. 2013), OrthoMCL (Chen et al. 2006), PhylomeDB (Huerta-Cepas et al. 2014), TreeFam (Schreiber et al. 2014), FungiPath (Grossetête et al. 2010), GreenPhylDB (Rouard et al. 2011) and PLAZA (Van Bel et al. 2018), PANTHER(Mi et al. 2019), Ensembl Compara (Herrero et al. 2016), HomoloGene (NCBI Resource Coordinators 2016).

*123 prokaryotic species (mainly Bacteria but also some Archaea) are included in the Pan-Compara resource which includes a selection of prokaryotic and eukaryotic species.

Annexe 2

OrthoInspector 3.0: open portal for comparative genomics

Published online 31 October 2018 *Nucleic Acids Research, 2019, Vol. 47, Database issue* ***D411–D418***
doi: 10.1093/nar/gky1068

OrthoInspector 3.0: open portal for comparative genomics

Yannis Nevers[1], Arnaud Kress[1], Audrey Defosset[1], Raymond Ripp[1], Benjamin Linard[2,3,4], Julie D. Thompson[1], Olivier Poch[1] and Odile Lecompte[1,*]

[1]Department of Computer Science, ICube, UMR 7357, University of Strasbourg, CNRS, Fédération de Médecine Translationnelle de Strasbourg, Strasbourg, France, [2]LIRMM, Univ Montpellier, CNRS, Montpellier, France, [3]ISEM, Univ Montpellier, CNRS, IRD, EPHE, CIRAD, INRAP, Montpellier, France and [4]AGAP, Univ Montpellier, CIRAD, INRA, Montpellier Supagro, Montpellier, France

Received September 12, 2018; Revised October 17, 2018; Editorial Decision October 18, 2018; Accepted October 19, 2018

ABSTRACT

OrthoInspector is one of the leading software suites for orthology relations inference. In this paper, we describe a major redesign of the OrthoInspector online resource along with a significant increase in the number of species: 4753 organisms are now covered across the three domains of life, making OrthoInspector the most exhaustive orthology resource to date in terms of covered species (excluding viruses). The new website integrates original data exploration and visualization tools in an ergonomic interface. Distributions of protein orthologs are represented by heatmaps summarizing their evolutionary histories, and proteins with similar profiles can be directly accessed. Two novel tools have been implemented for comparative genomics: a phylogenetic profile search that can be used to find proteins with a specific presence-absence profile and investigate their functions and, inversely, a GO profiling tool aimed at deciphering evolutionary histories of molecular functions, processes or cell components. In addition to the re-designed website, the OrthoInspector resource now provides a REST interface for programmatic access. OrthoInspector 3.0 is available at http://lbgi.fr/orthoinspectorv3.

INTRODUCTION

Genes descending from a common ancestor, or homologs, are commonly divided into two classes: orthologs, that are derived from a speciation event, and paralogs, that are derived from a duplication event (1). According to the ortholog conjecture (2), which has been debated recently but still holds (3,4), orthologs generally conserve the same function in distinct species while paralogs can evolve different or specialized functions. Furthermore, a discrimination between outparalogs and inparalogs is needed when studying evolutionary and functional relationships between proteins (5). Outparalogs are produced by a duplication event anterior to a given speciation event, while inparalogs result from a 'recent' duplication, posterior to a speciation event. Thus, inparalogs in one species are assumed to be relatively close to each other and are considered co-orthologs to their counterparts in another species deriving from the considered speciation event.

These notions are key principles in current biology and inferring the true orthologs or co-orthologs of proteins is crucial for comparative genomics and molecular biology. For example, it is essential in the transfer of data from experimental studies between species, thus making it possible to study human health in model organisms. It is also the keystone of phylogenetic profiling, an approach that exploits the presence and absence of protein orthologs across multiple species (6). The method is based on the principle that two proteins that interact or are involved in the same biological process tend to be conserved and lost together (7). Applications of phylogenetic profiling include protein-protein interaction inference and genotype-phenotype correlation as genes associated with a certain phenotypic trait tend to have a profile correlated with that trait's phylogenetic distribution (8).

More than 30 resources have been developed to address the challenges of orthologous relation inference and community efforts have been directed towards standardization and benchmarking of these resources, in the form of the Quest for Orthologs consortium (9). OrthoInspector (10,11) was shown to be one of the three most balanced methods of orthology inference in terms of precision and recall in a standardized benchmarking test (12) and performed well in other comparative studies (13). The previous release of OrthoInspector (11) provided two precomputed databases (Prokaryotes and Eukaryotes) that could be queried from its website, however since the last release the

*To whom correspondence should be addressed. Tel: +33 03 68 85 32 96; Email: odile.lecompte@unistra.fr

number of available annotated genomes has significantly increased and standards for web interfaces have evolved.

Here, we present the third release of OrthoInspector that includes a number of important developments. First, we report a major increase in the number of species represented in the OrthoInspector precomputed databases across the three domains of cellular life, including both in-domain and cross-domain relations, making the OrthoInspector databases the most exhaustive orthology resource to date in terms of covered species. Second, to manage the massive increase of data, the OrthoInspector website has been entirely redesigned to provide a streamlined and intuitive experience for users, including a summary visualization of ortholog distributions and novel tools allowing powerful comparative genomics analyses.

RESULTS

Improved coverage of the tree of life

Proteome selection. When designing the OrthoInspector databases, we focused on providing a broad coverage of the tree of life, with a selection of organisms that are representative of the taxonomic diversity. In order to meet this goal, we used the Uniprot Reference Proteomes (14), which result from an effort to efficiently sample the tree of life and limit redundancy. Incomplete genomes, mispredicted or fragmentary protein sequences constitute an important source of errors in orthology inference. Therefore, we used a combination of filters (see supplementary materials and methods) to exclude proteomes with abnormally small proteome size, a high proportion of small proteins (<100 amino acids) or of proteins that do not start with a methionine. Specifically, we excluded proteomes of Archaea and Bacteria with >20% of small proteins and/or 10% of false-start proteins and/or >10% proteins annotated as fragments. For Eukaryotes, we kept the same threshold for small proteins and excluded proteomes with >55% of false start proteins.

Starting from the 5443 Reference Proteomes, the quality filtering step resulted in the exclusion of 690 proteomes (13%). The percentages of excluded proteomes were similar across domains: 119 out of 830 eukaryotes (14%), 537 out of the 4400 Bacteria (12%) and 34 out of 213 Archaea (16%). In one case, we privileged the coverage of the tree of life over quality measures and kept the proteome of *Lokiarchaeum* sp. *GC14_75* owing to the general interest for representatives of the Asgard group in comparative genomics (15,16).

Database architecture. The OrthoInspector 3.0 databases cover 4753 organisms (+144% compared with the previous release): 3863 bacteria (+146%), 711 (+174%) eukaryotes and 179 archaea (+49%) (Figure 1). This is, to our knowledge, the widest coverage available for an orthology inference resource in terms of species (excluding viruses).

The database architecture is designed to cover the essential use cases for orthology data. It relies on three main databases, one for each domain of life. Each database provides all the orthologous relations between proteins of each species within the domain. This exhaustive coverage of each domain is suitable for fine grained studies, as it provides a good resolution at low taxonomic levels.

We designed a fourth database to provide orthologous relationships across a wider evolutionary spectrum and specifically, to cover the three domains simultaneously. To facilitate handling and interpretation of these cross-domain comparisons, we defined a subset of significant species that we will refer to as 'model species' (see Supplementary Table S1). We selected these species according to their importance in the biological field (e.g. model species such as *Mus musculus* or *Caenorhabditis elegans*) and/or to ensure a good taxonomic sampling (Figure 1). This selection corresponds to 317 species: 144 eukaryotes, 142 bacteria and 31 archaea.

OrthoInspector can thus be used to find intra-domain orthologs in a large number of species and to find inter-domain orthologs in fewer, well-studied, species. Users interested in orthology relationships between non-model species from different domains can find them by transitivity, by first finding orthologs in close 'model species'. This original implementation involving the co-existence of databases with different levels of granularity implies that orthologs can be found in all our available species without the huge computational burden a 'full' inference would require.

Complete information about the database content is available in Supplementary Table S1 and in the database tab on the website.

A new information design

To cope with the massive increase in the number of species available in the OrthoInspector databases and the corresponding increase in the number of orthology relationships, we implemented a new website interface providing a smooth navigation in the new datasets.

Access to protein entries. The OrthoInspector website offers two main ways to access the data: by protein identifier and by sequence similarity searches.

The protein identifier search is accessible from the main page, or anywhere on the site using the navigation bar. The user should define the appropriate database by selecting the domain of life of the query protein. Typing in the search bar triggers autocompletion and dynamically proposes a list of clickable protein entries available in the selected OrthoInspector database. The identifier search currently supports both Uniprot identifiers and Uniprot access numbers.

A sequence similarity search is also available from the OrthoInspector webpage or by selecting 'BLAST search' on the database tab. This launches a BLASTp (17,18) search against all protein sequences in the OrthoInspector databases. The result is a formatted BLAST output of the 50 best hits along with their corresponding local alignments and links to the corresponding protein pages in OrthoInspector.

Protein page. The data in OrthoInspector can be explored from protein pages. The protein page header gives a quick summary of the protein (gene name, description, organism). All Gene Ontology (19) terms associated with this protein are displayed in an extendable panel when available, as well as the protein sequence and a schematic view of InterPro (20) domains found in the protein. The protein page is the core section of the website architecture and provides access

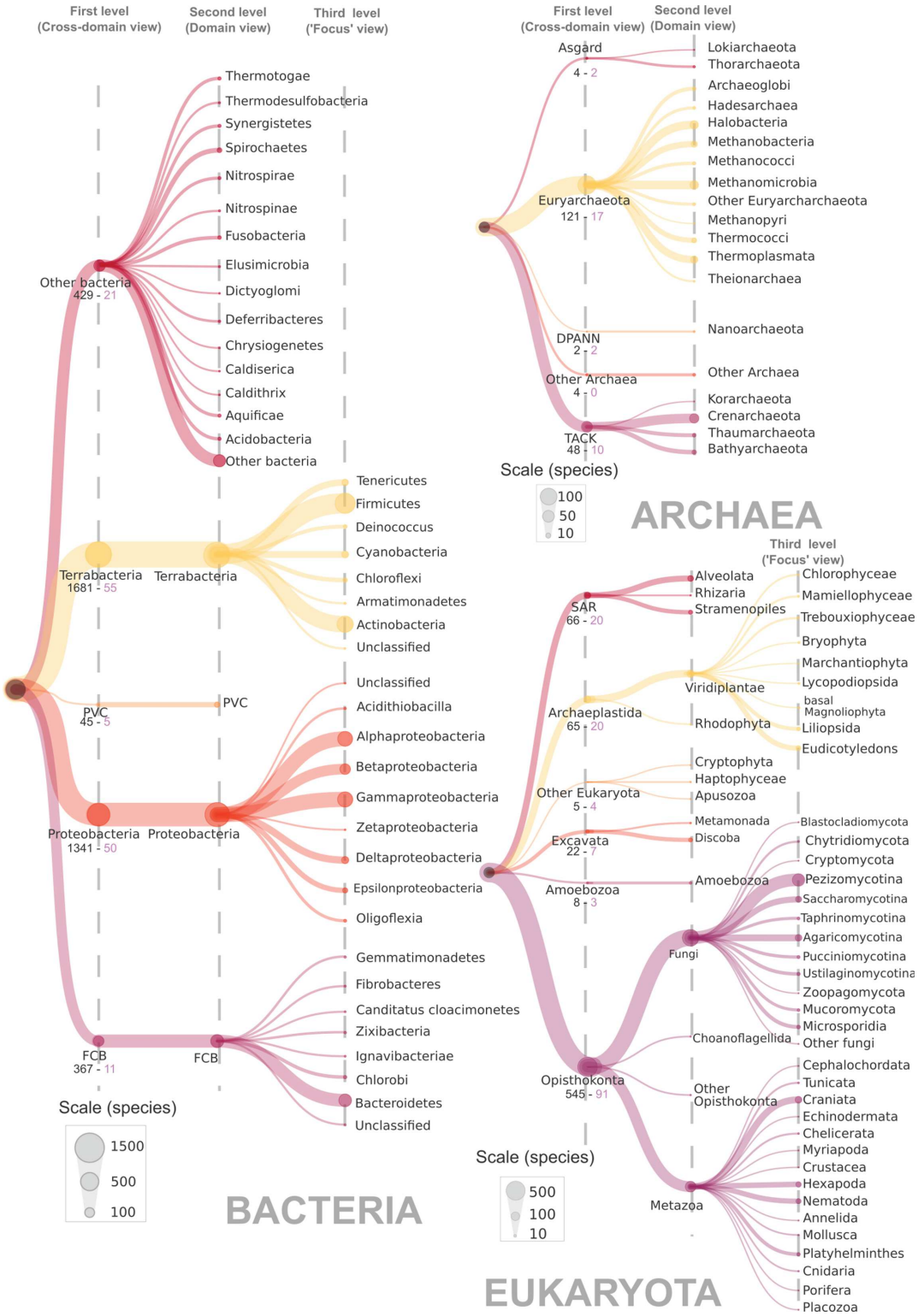

Figure 1. Taxonomic distribution of species represented in OrthoInspector. The domain trees are distributed on three 'levels'. The first level corresponds to the cross-domain taxonomic distribution heatmap shown when browsing the cross-domain database, the second level is shown on the heatmap for domain specific databases and the third level is the 'focus view' available for certain clades (see Figure 2). The size of a node is proportional to the number of species in the corresponding clades according to indicated scales. The number of species and model species in first-level clades are displayed in black and pink respectively.

to orthology relations, taxonomic distribution and proteins with similar distribution (detailed below).

Orthology data. Orthologous relationships are presented in the 'Orthologs and taxonomic distribution' section of the protein page. A menu allows users to choose display options, depending on their needs:

- **Domain's model organisms**: only orthologs found in the 'model organisms' of Eukaryotes, Bacteria or Archaea are shown in this tab. This view is used to find orthologs in popular species and avoids overwhelming the user with superfluous information. The page shown by default should meet the requirements of most users and thus serves as a suitable entry point.
- **Whole domain**: orthologs in all species of the in-domain databases are shown in this tab. This exhaustive view is suitable for an in-depth exploration of intra-domain relationships.
- **Three domains:** orthologs in 'model organisms' of the three domains of life are shown in this tab. This view, which provides orthologs across all domains of life, is relevant for broader comparative genomics studies. This tab is only available for proteins belonging to 'model organisms'.

All ortholog relations are shown in a table giving basic information: the type of relations (one-to-one, one-to-many, many-to-one, many-to-many), identifiers of all inparalogs (for many-to-*) and orthologs with links to their respective protein pages on the OrthoInspector and Uniprot web sites, the species name (linking to the NCBI taxonomy) and a summary of the species taxonomy. Additional information about orthologs (protein description and length) can be shown by customizing the output using the columns output button, in the top right corner.

By default, the table is ordered according to the taxonomic distance of the target species from the query species, as inferred from the NCBI taxonomy. Thus, except in the case of unusual evolutionary events, the first orthologs shown will be more closely related to the query protein. In the case of proteins with a large number of orthologs, a search bar allows the user to search specific results by identifier, species name, species taxid or even a specific clade name. For example, if a user is interested in orthologs of a human protein in representatives of the carnivore clade only, typing 'carnivora' on the search bar will achieve this.

Data export. From the protein page, multiple export options are available. Exports of the table itself are available in numerous formats (Excel, CSV, XML…) *via* the top right corner 'Export' button. User can also retrieve all sequences involved in selected relations (all inparalogs and orthologs) in FASTA format, which could serve as a starting point for further analyses.

OrthoInspector also offers the possibility to directly generate a multiple sequence alignment of the query protein and all its orthologs in selected species (and inparalogs, if any) using the latest version of the alignment workflow PipeAlign 2.0 (http://www.lbgi.fr/pipealign) (Kress, in prep).

Finally, on each protein page, the selected orthologous relations can be downloaded in the standardized OrthoXML format, as defined by the Quest for Orthologs consortium (21).

Taxonomic distribution summary. The orthologs table contains, as seen above, all information about orthology relations. However, making sense of such tables can be a daunting task, especially for proteins involved in many orthology relations. To facilitate knowledge extraction, the OrthoInspector protein page provides a summary view of the ortholog distribution at three levels of granularity: the domain's model organisms, the whole domain and all three domains.

This information appears in a banner above the orthologs table after complete loading and is displayed as a heatmap (see Figure 2) on a single row. Each tile of the heatmap corresponds to a major clade (Figure 1) of the selected domain, defined either from the NCBI taxonomy (22) or in some cases from the consensus in the literature (for example, 'Excavata ' appears in the cross-domains banner and is widely accepted by the community despite not existing as such in the NCBI taxonomy). For each clade, the corresponding tile is colored in green if orthologs are found in all its representatives and red if no orthologs are found, with intermediary states between these two colors if orthologs are found in a subset of representatives. The number of species in which orthologs are found and the total number of species belonging to the clade represented in the OrthoInspector database are both displayed when hovering over the tiles.

The clades on the heatmap are ordered according to the taxonomy: clades close to each other are side by side on the heatmap. The heatmap provides users with preliminary information about the evolutionary history (emergence and losses in major clades) of their protein family at a glance.

The clades displayed in this view depend on the granularity level selected by the user. In the cross-domain view, only high-level clades are indicated ('First level' in Figures 1 and 2A), for instance Opisthokonta. The domain of each clade is clearly indicated in the banner, by an indicator above the heatmap and by a color code. Some of the high-level clades are detailed in the 'domain's model organisms' and 'whole domain' views. For instance, Opisthokonta appear as Fungi, Choanoflagellida, Metazoa and Other Opisthokonta ('Second level' in Figures 1 and 2B). Additionally, major clades referencing many species can be further divided by clicking on the tile to display subtaxa and show a more nuanced version of the distribution (see 'Third level' in Figures 1 and 2C). For instance, 15 phyla or subphyla can be visualized for the Metazoa kingdom (156 species including 47 'model' species). These clickable tiles are identified by a blue frame.

Inparalogs distribution. Information about presence and absence of orthologs is fundamental when studying the evolutionary histories of proteins, but can miss some evolutionary events, notably duplication events. To address this issue, the taxonomic summary banner also provides a 'See inparalogs' button, that shows all inparalogs of the query protein relative to the considered clade. They are represented by ticks under the heatmap tiles that provide information

Figure 2. Taxonomic distribution heatmaps. Each labelled tile corresponds to a clade and is colored according to the proportion of species in the clade with at least one ortholog. Colors range from red (no species) to green (all species). (**A**) Heatmap corresponding to the cross-domain database. The domain of life of the clades is shown by an additional label and a color code. Inparalogs distribution is indicated by a tick under each clade. (**B**) Heatmap corresponding to the eukaryotic database. The box framed by a thin blue outline can be expanded to 'focus view'. (**C**) Heatmap corresponding to the 'focus view' of Metazoa.

about the timing of each duplication during the gene's evolutionary history (Figure 2A). For example, an inparalog of a human protein found in relation to all species except Opisthokonts may indicate a duplication of the ancestral gene in the Opisthokonta common ancestor.

Finally, the summary section also includes the list of species in which no orthologs were found.

Phylogenetic profiling tools

The presence and absence of orthologs summarized in the above section can be represented as detailed binary profiles, the phylogenetic profiles.

Searching for proteins with similar evolutionary histories. The OrthoInspector protein page can be used to find other proteins of the same species with similar phylogenetic profiles. This information is available under the 'Proteins with similar distribution' section on the Protein page. The data available in these sections are based on the Jaccard distance between all phylogenetic profiles of proteins in the same species (see supplementary materials and methods). The identifiers of proteins exhibiting a phylogenetic profile distance <0.4 are shown, along with a short description of their functions and the exact value of the distance. For clarity, only the five closest proteins are shown; additional proteins can be visualized by clicking 'See more'.

Distances are available both from a domain centric point of view (calculated on profiles limited to species of the same domain) or from a cross-domain point of view. While the domain specific section is available for all species in OrthoInspector, the cross-domain section is only available for 'model species'. Distances between intra-domain and cross-domain profiles may differ significantly only for proteins that are present in multiple domains.

Ciliary proteins are a good example of proteins whose phylogenetic profiles are clearly correlated to their function, since the cilium has a very specific evolutionary history in Eukaryotes including multiple independent losses (8). The cilium critically depends on molecular complexes to function properly, notably the intraflagellar transport (IFT) complexes (23). We searched a core protein of the IFT-A complex, IFT122 (IF122_HUMAN) on the OrthoInspector website. In the 'Proteins with similar distribution in Eukaryota' section, we found a list of 33 proteins, showing a significative enrichment in the GO term 'cilium' (P-value: 4.93×10^{-43}). This list includes 4 out of the 5 other components of the IFT-A complex and 8 out of the 16 components of IFT-B, most of them with a distance <0.3.

As illustrated by this example, these sections provide an original perspective when studying the function of proteins and can be used to obtain a list of other proteins with potentially similar functions and possible interaction partners.

Searching proteins with a known profile. Genes associated with a given phylogenetic trait tend to share the same distribution. The distribution of a trait can thus be exploited to identify associated genes. OrthoInspector offers an original tool for phylogenetic profiling, i.e. to search for proteins with orthologs present in a defined set of species or clades

and absent in others. This tool is available from the home page and under the 'Access/Search by profile' tab. Users should select their query species on the dropdown menu and then interact with a dynamic representation of the NCBI taxonomic tree to define the profile. Clicking once on a clade imposes the presence of orthologs in at least one species of the clade, double clicking imposes the absence in all species, a third click removes the constraints. Once the constraints are set, the database is queried to find all proteins meeting the user's requirements (Figure 3).

The resulting proteins are displayed as panels in the result windows with their distribution summary (see above) to facilitate identification of distribution subcategories within the results. Each protein panel also contains a short description of the protein along with the associated Gene Ontology terms. For a functional analysis of the complete protein list, a button can be clicked to run a GO term enrichment analysis using the Panther webservice (24). The full list of proteins obtained can be exported using the 'Download list' button, for further analysis.

Using this tool, we performed a phylogenetic profile search on the cross-domain database. Our objective was to identify Eukaryotic Signature Proteins (ESP) that were also present in Asgard Archaea, a clade whose discovery sparked interest due to its unexpected similarity to Eukaryotes (15,16). We searched for orthologs of *Homo sapiens* proteins present in Archaea of the Asgard group but absent in other Archaea and in Bacteria (Figure 3A). This operation resulted in a total of 69 proteins with the required distributions (Figure 3B). The list shows a strong enrichment in proteins with GTPase activity (P-value: 4.97×10^{-28}) and vesicle-mediated transport (P-value: 5.12×10^{-36}), in agreement with previous studies (16). We also retrieved actin-cytoskeleton proteins and ubiquitin-associated proteins, two iconic examples of ESP previously reported in the Asgard group (16). As shown here, the phylogenetic profile search rapidly provides both a list of genes associated with specific distributions and the tools required to extract functional knowledge.

Identifying profiles linked to a functional category. OrthoInspector provides an original tool to explore the evolutionary history of a biological function, process or component. This module, available on the home page or via the 'Access/GO profile' tab, provides the distribution of all proteins of a species associated with a given GO term. After selecting the database, species, and GO term of interest, the user retrieves the list of matching proteins, in the format described above with the summary of the distribution of each associated protein. In this way, users can derive the distribution associated with their function of interest and explore the different evolutionary histories of proteins involved in the same biological system.

Data and software accessibility

This database update is complemented by the release of the new version 3.0 of the OrthoInspector software suite, developed in Java, and available for download on the website in the download section (http://www.lbgi.fr/orthoinspectorv3/download_Package). This release does not involve changes to the main algorithm (10) but provides several software improvements.

Software improvement. Several code modifications were performed to optimize the management of the massive quantities of data. This implies type changes to handle larger datasets, code optimization by reducing loop redundancy, the use of more optimized data structures (library Fastutil, arXiv:1601.06919) and more efficient database access from the software (fewer SQL queries). This version of OrthoInspector runs faster than the precedent for large computations and can still be parallelized when installing a large database.

Improved accessibility. Following feedback from users, the new OrthoInspector version provides an easier accessibility for small datasets. Until now, fully supported database systems included MySQL and PostgreSQL, which require prior experience of SQL management systems. This version comes with full support for SQLite database, which eliminates most of the preliminary steps for computing a local database since no database server configuration is required. We recommend the use of the easily accessible SQLite database option when installing small local databases and, for performance reasons, the use of PostgreSQL and MySQL systems for larger databases (several hundred of species). Updated tutorials for the installation procedure are available on the website.

Precomputed databases. All four precomputed databases (Eukaryotes, Bacteria, Archaea, Cross-Domain) can be accessed *via* the website interface. Due to the data volume (up to multiple terabytes in a single database), the database dump is not available for direct download but could be made available on demand.

Quest for Ortholog consortium reference proteome. The Quest for Ortholog (QFO) consortium is part of an ongoing effort from the community pushing for standardization in orthology inferences. The QFO consortium published a list of 78 reference proteomes representing high quality proteomes and recommend using it for benchmarking purposes. The precomputed orthology relationship made using this benchmark are available on http://www.lbgi.fr/orthoinspectorv3/QFO.

Webservices. In addition to the web interface, a programming access is a major requirement for modern databases, as it allows more flexible use of data. In this release, we introduce a Representational State Transfer (REST) API providing access to most data available from the website, through the Swagger framework (https://swagger.io). The documentation is available on the website (http://www.lbgi.fr/orthoinspectorv3/API) where all endpoints and their parameters are described. All queries can be executed with custom parameters directly from the documentation page.

CONCLUSIONS AND FUTURE DIRECTIONS

With this new release of OrthoInspector, we provide improvement in two main areas: proteome coverage and information design.

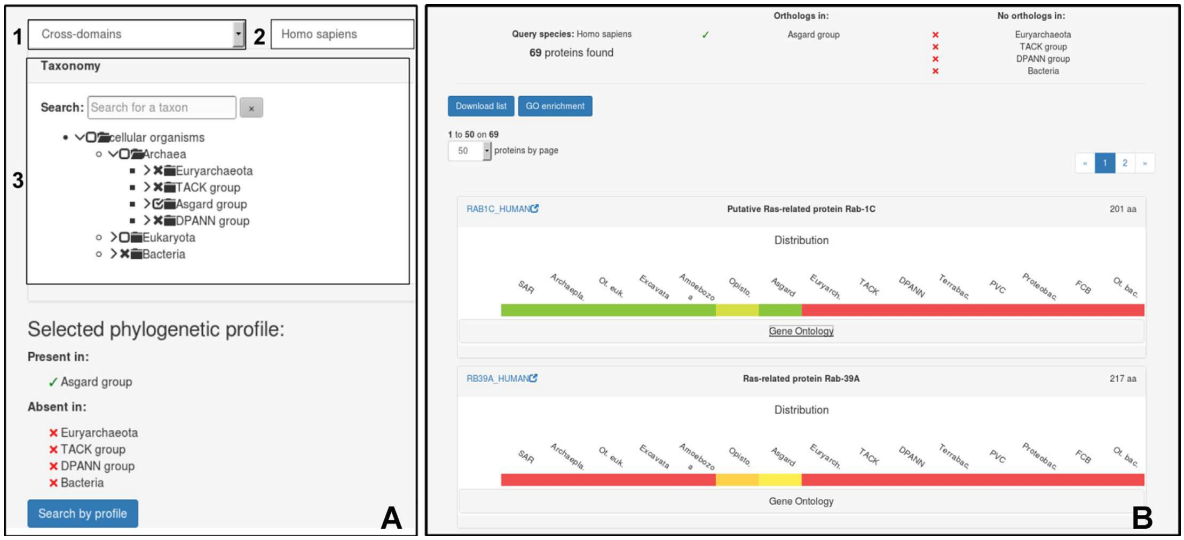

Figure 3. Phylogenetic profile search interface. (**A**) Definition of the phylogenetic profile. User selects: (1) the database, (2) the query species in the drop-down menu and (3) the presence/absence constraints using the phylogenetic tree. A summary of constraints is shown below the tree. Here, human proteins absent in Prokaryotes except the archaeal Asgard group are selected. (**B**) Output of the profile search. Constraints are included on the top with the number of proteins found. Proteins are displayed in panels, showing their distributions and functional information. Gene Ontology enrichment can be performed on the protein list.

The new databases boast a massive increase in the number of species across the three domains of life and provide the most comprehensive ortholog relations resource in terms of species coverage. Nevertheless, this increase did not involve simply adding a substantial number of species. Special attention was paid to both quality of proteomes and taxonomic coverage. With the increasing rate of genome sequencing, our scheduled strategy to ensure scalability will include regular updates of the current proteome content and the addition of new species while maintaining our standard of proteome quality. This will come with an updating procedure directly added to the software suite to allow any user to easily update their local databases with the latest data.

In terms of accessibility, the installation process of local databases using the software suite has been simplified and more importantly, the web interface of the OrthoInspector precomputed databases has been significantly reorganized. The new design offers improved access to orthologous data in the three domains of life. In addition, we believe that the implementation of original and user-friendly comparative genomics tools will be useful for anyone interested in comparative genomics and evolutionary studies of protein families. The next step for OrthoInspector will be the automated definition and analysis of orthologous families among 'model species' by exploiting our experience in multiple sequence alignment construction (25,26) (Kress, in prep). This will allow the exploration of protein evolution through the three life domains at different levels of resolution, from presence/absence of orthologs to subtler patterns of differential conservation at domain or block levels.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank the Bio-statistics, Informatics and Complex System platform (BICS) and BISTRO bioinformatics platforms for informatics support and the European Grid Infrastructure for cloud computing facilities. We also thank our users for their feedback that helped to improve our suite and website.

FUNDING

Agence Nationale de la Recherche [BIPBIP: ANR-10-BINF-03–02, ReNaBi-IFB: ANR-11-INBS-0013, Labex Agro: ANR-10-LABX-0001–01 to B.L., Labex CeMEB: ANR-10-LABX-0004 to B.L., Labex NUMEV: ANR-10-LABX-20 to B.L.]; Institute funds from the Centre National de la Recherche Scientifique and the Université de Strasbourg. Funding for open access charge: Centre National de la Recherche Scientifique.
Conflict of interest statement. None declared.

REFERENCES

1. Fitch,W.M. (1970) Distinguishing homologous from analogous proteins. *Syst. Zool.*, **19**, 99–113.
2. Tatusov,R.L., Koonin,E.V. and Lipman,D.J. (1997) A Genomic Perspective on Protein Families. *Science*, **278**, 631–637.
3. Nehrt,N.L., Clark,W.T., Radivojac,P. and Hahn,M.W. (2011) Testing the Ortholog Conjecture with Comparative Functional Genomic Data from Mammals. *PLOS Comput. Biol.*, **7**, e1002073.
4. Altenhoff,A.M., Studer,R.A., Robinson-Rechavi,M. and Dessimoz,C. (2012) Resolving the ortholog conjecture: orthologs tend to be weakly, but significantly, more similar in function than paralogs. *PLoS Comput. Biol.*, **8**, e1002514.
5. Koonin,E.V. (2005) Orthologs, paralogs, and evolutionary genomics. *Annu. Rev. Genet.*, **39**, 309–338.
6. Pellegrini,M., Marcotte,E.M., Thompson,M.J., Eisenberg,D. and Yeates,T.O. (1999) Assigning protein functions by comparative

genome analysis: protein phylogenetic profiles. *Proc. Natl. Acad. Sci. U.S.A.*, **96**, 4285–4288.
7. Pellegrini,M. (2012) Using phylogenetic profiles to predict functional relationships. *Methods Mol. Biol.*, **804**, 167–177.
8. Nevers,Y., Prasad,M.K., Poidevin,L., Chennen,K., Allot,A., Kress,A., Ripp,R., Thompson,J.D., Dollfus,H., Poch,O. *et al.* (2017) Insights into Ciliary Genes and Evolution from Multi-Level Phylogenetic Profiling. *Mol. Biol. Evol.*, **34**, 2016–2034.
9. Forslund,K., Pereira,C., Capella-Gutierrez,S., Silva,D., Sousa,A., Altenhoff,A., Huerta-Cepas,J., Muffato,M., Patricio,M., Vandepoele,K. *et al.* (2018) Gearing up to handle the mosaic nature of life in the quest for orthologs. *Bioinformatics*, **34**, 323–329.
10. Linard,B., Thompson,J.D., Poch,O. and Lecompte,O. (2011) OrthoInspector: comprehensive orthology analysis and visual exploration. *BMC Bioinformatics*, **12**, 11.
11. Linard,B., Allot,A., Schneider,R., Morel,C., Ripp,R., Bigler,M., Thompson,J.D., Poch,O. and Lecompte,O. (2015) OrthoInspector 2.0: Software and database updates. *Bioinformatics*, **31**, 447–448.
12. Altenhoff,A.M., Boeckmann,B., Capella-Gutierrez,S., Dalquen,D.A., DeLuca,T., Forslund,K., Huerta-Cepas,J., Linard,B., Pereira,C., Pryszcz,L.P. *et al.* (2016) Standardized benchmarking in the quest for orthologs. *Nat. Methods*, **13**, 425–430.
13. Liebeskind,B.J., McWhite,C.D. and Marcotte,E.M. (2016) Towards Consensus Gene Ages. *Genome Biol. Evol.*, **8**, 1812–1823.
14. UniProt: the universal protein knowledgebase (2017) *Nucleic Acids Res.*, **45**, D158–D169.
15. Spang,A., Saw,J.H., Jørgensen,S.L., Zaremba-Niedzwiedzka,K., Martijn,J., Lind,A.E., van Eijk,R., Schleper,C., Guy,L. and Ettema,T.J.G. (2015) Complex archaea that bridge the gap between prokaryotes and eukaryotes. *Nature*, **521**, 173–179.
16. Zaremba-Niedzwiedzka,K., Caceres,E.F., Saw,J.H., Bäckström,D., Juzokaite,L., Vancaester,E., Seitz,K.W., Anantharaman,K., Starnawski,P., Kjeldsen,K.U. *et al.* (2017) Asgard archaea illuminate the origin of eukaryotic cellular complexity. *Nature*, **541**, 353–358.
17. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.
18. Camacho,C., Coulouris,G., Avagyan,V., Ma,N., Papadopoulos,J., Bealer,K. and Madden,T.L. (2009) BLAST+: architecture and applications. *BMC Bioinformatics*, **10**, 421.
19. Gene Ontology Consortium (2015) Gene Ontology Consortium: going forward. *Nucleic Acids Res.*, **43**, D1049–D1056.
20. Finn,R.D., Attwood,T.K., Babbitt,P.C., Bateman,A., Bork,P., Bridge,A.J., Chang,H.-Y., Dosztányi,Z., El-Gebali,S., Fraser,M. *et al.* (2017) InterPro in 2017—beyond protein family and domain annotations. *Nucleic Acids Res.*, **45**, D190–D199.
21. Dessimoz,C., Gabaldón,T., Roos,D.S., Sonnhammer,E.L.L., Herrero,J., Altenhoff,A., Apweiler,R., Ashburner,M., Blake,J., Boeckmann,B. *et al.* (2012) Toward community standards in the quest for orthologs. *Bioinformatics*, **28**, 900–904.
22. Federhen,S. (2012) The NCBI Taxonomy database. *Nucleic Acids Res.*, **40**, D136–D143.
23. Lechtreck,K.F. (2015) IFT-Cargo Interactions and Protein Transport in Cilia. *Trends Biochem. Sci.*, **40**, 765–778.
24. Mi,H., Poudel,S., Muruganujan,A., Casagrande,J.T. and Thomas,P.D. (2016) PANTHER version 10: expanded protein families and functions, and analysis tools. *Nucleic Acids Res.*, **44**, D336–D342.
25. Vanhoutreve,R., Kress,A., Legrand,B., Gass,H., Poch,O. and Thompson,J.D. (2016) LEON-BIS: multiple alignment evaluation of sequence neighbours using a Bayesian inference system. *BMC Bioinformatics*, **17**, 271.
26. Kress,A., Lecompte,O., Poch,O. and Thompson,J.D. (2018) PROBE: analysis and visualization of protein block-level evolution. *Bioinformatics*, **34**, 3390–3392.

Annexe 3

Gènes candidats issus de l'application de BLUR à la multiciliation

A3LT2_HUMAN	Haute priorité
AKAP9_HUMAN	Haute priorité
AL5AP_HUMAN	Haute priorité
ANF_HUMAN	Haute priorité
ANFC_HUMAN	Priorité moyenne
APOM_HUMAN	Haute priorité
ARP5_HUMAN	Haute priorité
ASB6_HUMAN	Priorité moyenne
CACB2_HUMAN	Priorité moyenne
CC126_HUMAN	Haute priorité
CCD92_HUMAN	Haute priorité
CD11A_HUMAN	Priorité moyenne
CE164_HUMAN	Haute priorité
CEP72_HUMAN	Haute priorité
CHCH2_HUMAN	Haute priorité
CHCH9_HUMAN	Haute priorité
CHUR_HUMAN	Priorité moyenne
CLD14_HUMAN	Haute priorité
CN37_HUMAN	Haute priorité
CNO11_HUMAN	Haute priorité
COIA1_HUMAN	Haute priorité
CP110_HUMAN	Haute priorité
CR025_HUMAN	Haute priorité
CRBA2_HUMAN	Haute priorité
EMSY_HUMAN	Haute priorité
ENDOU_HUMAN	Haute priorité
F181A_HUMAN	Haute priorité
FA53C_HUMAN	Haute priorité
GOLM1_HUMAN	Haute priorité
HAUS1_HUMAN	Haute priorité
HAUS3_HUMAN	Priorité moyenne
HS3S1_HUMAN	Haute priorité
I6L893_HUMAN	Haute priorité
IFM2_HUMAN	Haute priorité
MANS4_HUMAN	Haute priorité
MBD4_HUMAN	Haute priorité
MCIN_HUMAN	Haute priorité
MELPH_HUMAN	Haute priorité
MEOX2_HUMAN	Haute priorité
MSX1_HUMAN	Haute priorité
NFM_HUMAN	Haute priorité
NGB_HUMAN	Haute priorité
OC90_HUMAN	Haute priorité
PDLI5_HUMAN	Haute priorité
PHF20_HUMAN	Haute priorité
PHLP_HUMAN	Haute priorité
PIGM_HUMAN	Haute priorité
PLAC8_HUMAN	Haute priorité
PLAG1_HUMAN	Haute priorité
PPR26_HUMAN	Haute priorité
RN152_HUMAN	Haute priorité
SIX4_HUMAN	Haute priorité
SOX17_HUMAN	Haute priorité
SRFB1_HUMAN	Haute priorité
T22D2_HUMAN	Haute priorité
TAF3_HUMAN	Haute priorité
TAGAP_HUMAN	Haute priorité
TM6S1_HUMAN	Haute priorité
TRI11_HUMAN	Haute priorité
ZBT7B_HUMAN	Haute priorité

Annexe 4

Gènes surexprimés dans au moins deux expériences de transcriptomique sur la multiciliation

Gène	Cluster
A4gnt	3
abcc5	3
abcg2	3
actr1a	3
ahrr	3
ankdd1a	3
ankrd28	3
ankrd35	3
anxa1	3
ap3b1	3
araf	3
arhgef18	3
arhgef28	3
arhgef37	3
arhgef7	3
arl6ip6	3
astl	3
atat1	3
ate1	3
atg2a	3
atp10b	3
atp7a	3
auts2	3
b3gnt3	3
bcar3	3
bcat1	3
bdnf	3
bmp1	3
bmt2	3
btbd9	3
c1orf112	3
c4orf33	3
c5orf24	3
ca4	3
cadm4	3
camk2g	3
capn1	3
capn7	3
capn9	3
cask	3
cast	3
cc2d1a	3
cd69	3
cdc42ep1	3
cep68	3
clec16a	3
clock	3
cluh	3
cog6	3
cog7	3
copg1	3
cux1	3
cyb5a	3
cyb5b	3
dcun1d2	3
dgcr2	3
dhdds	3
dlx6	3
dnase1l3	3
dock5	3
dsc2	3
dsg2	3
dusp22	3
eml4	3
emp3	3
endod1	3
entpd4	3
epb41l4a	3
eps8l2	3
etv3	3
exoc6b	3
faah	3
fam129b	3
fam160a1	3
fam20a	3
fam241a	3
fam72a	3
fam83b	3
fbxo31	3
foxo1	3
foxp1	3
fry	3
gale	3
gbf1	3
gbx2	3
gclc	3
glipr2	3
Gmppb	3
gosr1	3
gpr75	3
gpr87	3
gramd1c	3
gse1	3
gstp1	3
gtpbp1	3
harbi1	3
hbg2	3
hexb	3
hid1	3
hoxa1	3
hoxa2	3
hpgds	3
ibtk	3
il17re	3
inpp5a	3
itga6	3
kcna2	3
kcnk16	3
kctd2	3
kiaa1324	3
kidins220	3
klhl9	3
krt222	3
ksr1	3
l1td1	3
lgals3	3
lhfp	3
lrriq3	3
madd	3
magi1	3
man1a1	3
map2k4	3
mapk12	3
mark4	3
mef2d	3
mgat3	3
mgat5	3
mgll	3
mon2	3
msantd3	3
muc4	3
mycbp2	3
myo18a	3
myo19	3

Clusters cibles de FOXJ1 et FOXN4

myzap	3
nadk	3
nbl1	3
nceh1	3
nedd1	3
nedd4	3
nol4l	3
nqo1	3
nrg1	3
nsf	3
nxn	3
ormdl2	3
osbpl11	3
osgin1	3
oxct1	3
pde4d	3
pdhx	3
pdlim3	3
pgm3	3
pkp3	3
plekha6	3
plekha8	3
plp2	3
por	3
ppl	3
ppp2r5b	3
proser2	3
prr15	3
prtg	3
ptprh	3
ptprz1	3
r3hdm4	3
rab27b	3
rab3gap2	3
rad51d	3
rasef	3
rassf2	3
rb1cc1	3
reps1	3
retsat	3
rhbdf1	3
rnf5	3
rps27	3
s100a11	3
s100a2	3

scara3	3
scel	3
scin	3
scube2	3
sec31b	3
sel1l	3
sema3f	3
sfmbt1	3
sh2d3c	3
sh3bp5	3
sh3gl1	3
slc25a1	3
slc35a1	3
slc35e4	3
slc38a10	3
slc41a1	3
slit2	3
slmap	3
snx33	3
snx9	3
spg11	3
spg21	3
sppl2c	3
spsb3	3
sptlc2	3
srebf2	3
steap1	3
stk11	3
stk11ip	3
stom	3
strip2	3
sult6b1	3
syap1	3
syne2	3
synpo2	3
syt7	3
sytl1	3
sytl4	3
sytl5	3
tax1bp1	3
tbc1d14	3
tbc1d22b	3
tep1	3
tiam1	3
tlr2	3

tmem164	3
tmem180	3
tmem30b	3
tmprss2	3
tmsb15a	3
tnk2	3
tssc1	3
tstd1	3
ttc13	3
ttc37	3
ttc5	3
uap1	3
ubtd1	3
usp32	3
usp8	3
vgll1	3
vil1	3
vti1a	3
wdfy4	3
wwc1	3
yipf3	3
zdhhc24	3
zdhhc3	3
zfyve28	3
znf593	3
ablim1	6
acsl1	6
adamts4	6
agap2	6
ago3	6
agpat2	6
akap9	6
amfr	6
amotl2	6
ano10	6
aoc3	6
appl1	6
arf6	6
arhgap1	6
arhgef17	6
arhgef6	6
asb18	6
atf6	6
atl3	6
atp11a	6

atp6v1g2	6
atp6v1g3	6
atp8b2	6
b4galt1	6
bivm	6
c1galt1c1	6
c2orf80	6
cacng7	6
calcoco1	6
capns1	6
casd1	6
casp6	6
cbr4	6
ccdc102a	6
ccdc62	6
ccng2	6
cd82	6
cdh6	6
cerk	6
chmp1b	6
chmp5	6
cnga3	6
col4a3bp	6
cxadr	6
cyb561d2	6
cyp2c8	6
dhrsx	6
dld	6
dlg1	6
dnajc3	6
dnm2	6
dse	6
dvl3	6
dyrk2	6
efcab4a	6
ei24	6
eps8l1	6
etnk1	6
evc	6
evc2	6
eya2	6
ezr	6
fa2h	6
fam101b	6
fam13b	6

fam188b2	6
fam3a	6
farp1	6
fbp1	6
fez1	6
fnip2	6
foxp4	6
frk	6
gadd45b	6
ggt5	6
gk	6
golga3	6
gprc5c	6
grk6	6
hadha	6
hadhb	6
hdac4	6
hdac7	6
hectd3	6
hexdc	6
hhla2	6
hopx	6
hyal2	6
ifnlr1	6
ikbip	6
ikzf2	6
jak1	6
jak2	6
kctd14	6
kiaa0922	6
kiaa1468	6
kiaa1522	6
klc1	6
krt12	6
lmo7	6
lnx1	6
lnx2	6
lrit3	6
macf1	6
map3k2	6
mark2	6
mcoln3	6
mid1	6
misp	6
mnt	6

mpp7	6
mut	6
mvp	6
mylk4	6
myo5c	6
myo6	6
naaa	6
nbr1	6
nck2	6
ncoa1	6
nek7	6
net1	6
nfe2l2	6
nrip2	6
olfm4	6
osbpl2	6
pam	6
paox	6
pccb	6
pde1a	6
pdk3	6
pex11g	6
pi4ka	6
pigk	6
pkib	6
plekha2	6
pls3	6
plscr3	6
plxna2	6
pm20d1	6
ppp1r15b	6
ppp1r2	6
ppp2cb	6
prkab2	6
prkag2	6
prkcz	6
ptpn13	6
ptpn6	6
ptprd	6
ranbp9	6
rapgef3	6
rasal2	6
rassf6	6
rilpl1	6
scnn1g	6

Clusters cibles de FOXJ1 et FOXN4

sec14l2	6
serinc3	6
sgk3	6
siae	6
slc10a3	6
slc26a4	6
slc2a12	6
slc35a3	6
slc35b4	6
slc37a4	6
slc4a4	6
slc5a3	6
slurp1	6
snap23	6
sppl2a	6
sptlc1	6
ssuh2	6
stard13	6
stx3	6
stxbp5l	6
syt16	6
sytl2	6
tank	6
tmc7	6
tmem131	6
tmem222	6
tmem38b	6
tmem45b	6
tmf1	6
tmod3	6
tnfaip8	6
tnip1	6
trpm6	6
trps1	6
trpv6	6
ttc7b	6
tubd1	6
twf1	6
txn2	6
ugcg	6
uprt	6
vtcn1	6
wdtc1	6
ypel2	6
zer1	6

sept5	7
ache	7
adipor2	7
agfg1	7
aifm2	7
akap2	7
aldh3a2	7
angptl5	7
anks1a	7
ap1g1	7
apobec2	7
arfgef1	7
arfip1	7
asb12	7
asb9	7
atp6v0a1	7
atp6v1a	7
atp6v1h	7
atp9a	7
b3galt2	7
bcap31	7
bcr	7
bhlhe40	7
bin3	7
cacnb4	7
cand1	7
cbr1	7
cdc42se2	7
cdk14	7
cmpk1	7
cnga1	7
copa	7
csdc2	7
csgalnact1	7
csk	7
csnk1g3	7
ctif	7
cxcl9	7
cyp27a1	7
dennd1b	7
derl1	7
desi1	7
dhrs13	7
dsg3	7
edem3	7

efnb3	7
ern1	7
exoc7	7
fam135a	7
fam43a	7
fam46a	7
fbxo41	7
fcn1	7
fig4	7
fign	7
fkbp1b	7
flnb	7
folr4	7
fzd5	7
gabarapl3	7
galnt1	7
gipc1	7
gli3	7
Gmds	7
gpr116	7
gps1	7
gpt2	7
herc1	7
higd1a	7
hk2	7
hm13	7
hsd17b3	7
hsd17b6	7
id2	7
igf2	7
inadl	7
ist1	7
junb	7
kcnv2	7
kctd1	7
kif16b	7
kif5b	7
lin54	7
lipt1	7
lrat	7
lyn	7
lzic	7
magi3	7
meis1	7
mfsd1	7

mfsd6	7
mgat4a	7
micu2	7
mroh1	7
mrpl35	7
nes	7
nmral1	7
nomo2	7
nos1	7
nt5m	7
nudt4	7
nxpe4	7
palm	7
parva	7
pc	7
pcbp3	7
pcyt1a	7
pdia5	7
phex	7
phtf2	7
picalm	7
pigc	7
pigg	7
pigv	7
pik3ap1	7
pik3r5	7
pikfyve	7
pitpnc1	7
pkmyt1	7
plcg2	7
plch2	7
plekha7	7
plekhm1	7
pmm2	7
pnpla2	7
ppa1	7
ppp1r12a	7
ppp3ca	7
preb	7
prkcb	7
psen1	7
ptbp3	7
ptpn23	7
rab14	7
rap1gds1	7
rhpn2	7
rnaset2	7
rpl24	7
rpl27	7
rpl28	7
rps6kc1	7
ryr3	7
samd13	7
sds	7
sec11c	7
serinc2	7
serpinc1	7
sik2	7
slc12a9	7
slc18a2	7
slc25a16	7
slc35c1	7
slc39a10	7
slc39a11	7
slc39a7	7
slc39a8	7
slc41a2	7
smap1	7
snx12	7
spag9	7
spred1	7
stap2	7
stx8	7
sult2b1	7
syvn1	7
tbc1d5	7
tbx3	7
tgds	7
tirap	7
tmed10	7
tmem101	7
tmem115	7
tmem181	7
tmem229b	7
tmem62	7
tmem72	7
tnfaip3	7
tpcn2	7
trafd1	7
trappc10	7
trappc12	7
trib1	7
ube3c	7
ufm1	7
ugt8	7
upk2	7
vmp1	7
vps35	7
vps51	7
wdfy3	7
wdr7	7
wfikkn2	7
xiap	7
zdhhc5	7
zdhhc7	7
znf300	7
znf462	7
znf703	7
zw10	7
ac133919.6	14
ac244163.2	14
aco2	14
acsl5	14
adam11	14
adck2	14
agmo	14
akip1	14
apba3	14
appbp2	14
armc6	14
arnt	14
b3gat2	14
bap1	14
c15orf41	14
c1orf50	14
camk4	14
ccdc101	14
ccdc42	14
cdh15	14
cdhr5	14
cog2	14
coq4	14
cox5a	14
cpe	14
cryz	14

Clusters cibles de FOXJ1 et FOXN4

daglb	14
dapk2	14
dclk1	14
desi2	14
ell	14
exoc2	14
exosc8	14
fli1	14
foxd3	14
frmpd4	14
fxyd6	14
gba2	14
glrx	14
gprc5b	14
has2	14
herpud1	14
higd1b	14
hsd11b1l	14
ifitm10	14
il21r	14
ipo4	14
jkamp	14
kcng1	14
loxhd1	14
lppr5	14
lyrm1	14
map11	14
mcm9	14
metrnl	14
mettl21c	14
mettl5	14
mrps18c	14
mtmr9	14
mvk	14
ndufa5	14
ndufaf3	14
nfkb1	14
nostrin	14
nrg4	14
ocln	14
p2rx7	14
paaf1	14
pde1b	14
pdgfra	14
pex10	14

phka2	14
pkdcc	14
pm20d2	14
pmch	14
prdx5	14
psph	14
ptpn18	14
rad9b	14
rasgrp2	14
rgp1	14
rpl31	14
rpl34	14
rpl37a	14
rps15a	14
rps29	14
s100a10	14
scxa	14
slc44a5	14
slc5a1	14
slco3a1	14
snx14	14
spidr	14
tac3	14
tbc1d7	14
tbk1	14
tcf21	14
tmem258	14
tmem26	14
tmem55a	14
tmem88	14
trappc6b	14
trappc8	14
trpv5	14
tufm	14
vps33b	14
vps8	14
vsx2	14
xkr6	14
xrcc4	14
znf585a	14

Gènes	Cluster
ac079354.1	17
agbl3	17
alkbh7	17
ankar	17
ankrd60	17
arl2	17
azi1	17
bbs1	17
bbs4	17
c10orf67	17
c20orf201	17
c8orf37	17
ccdc150	17
ccdc160	17
ccdc166	17
ccdc169	17
ccdc175	17
ccdc63	17
ccdc92	17
cep104	17
cep19	17
cep89	17
cerkl	17
clhc1	17
crocc	17
dopey1	17
eny2	17
fam154a	17
fam221a	17
fam227a	17
fbn1	17
fbxl2	17
flacc1	17
flj27352	17
fopnl	17
fuz	17
gins3	17
heatr2	17
ift140	17
ift27	17
ift74	17
ift88	17
intu	17
kbtbd8	17
kiaa0556	17
kiaa1430	17
kif17	17
klhdc8a	17
lpcat4	17
lrp2bp	17
lrrc27	17
lrrc61	17
lrrc63	17
lrrc69	17
lrrc9	17
lztfl1	17
mapkbp1	17
mapre3	17
mycbp	17
mypop	17
nek1	17
nin	17
nit2	17
nme8	17
noxred1	17
nrde2	17
pak1ip1	17
pank2	17
pcnxl4	17
pdhb	17
pih1d3	17
rp11-503n18.3	17
rp11-579d7.1	17
spata4	17
stpg4	17
stx10	17
surf2	17
tctn2	17
tctn3	17
tepp	17
tmem17	17
tpgs1	17
trim13	17
ttbk2	17
ttc23l	17
ttc26	17
ttc39a	17
ttc6	17
ttll5	17
tubb2b	17
tubb4b	17
tube1	17
usp21	17
vstm5	17
wars	17
wdr47	17
wdr5	17
wdr5b	17
wdr60	17
wdr88	17
yes1	17
znf474	17
ac005841.1	18
adh1b	18
ank2	18
arntl	18
atp2c2	18
axdnd1	18
b3gnt5	18
bdh1	18
bend5	18
c11orf1	18
c20orf85	18
ccdc27	18
ccdc64b	18
cdkl2	18
cdkn1a	18
ces1	18
chl1	18
clic3	18
dgkd	18
dnah7	18
dync2li1	18
dynlt1	18
elmo3	18
fam169a	18
fam20c	18
fam214a	18
fam227b	18
fam228b	18
fdxr	18
gas8	18
Gmppa	18
hes1	18

Clusters cibles des couples MCIDAS/E2F4 et FOXJ1/FOXN4

ift122	18
irf4	18
katnal1	18
kiaa1549l	18
kiaa1875	18
kifap3	18
lgalsl	18
lpin2	18
lrrc49	18
lrriq4	18
mtus2	18
ncs1	18
nme7	18
nudt9	18
pex11a	18
pfn4	18
pofut2	18
ptplb	18
ptprn	18
rbl2	18
rgs12	18
rp11-723o4.6	18
sclt1	18
slc16a12	18
slc26a5	18
slc2a10	18
socs5	18
sp7	18
spata6	18
ssbp2	18
styk1	18
sun2	18
surf4	18
tbc1d32	18
tcp11l2	18
tex33	18
tex43	18
tm4sf18	18
tmbim1	18
tollip	18
trim32	18
ttc40	18
ttll10	18
txn	18
unc119b	18

ush1g	18
ac002365.1	19
adra2a	19
ak5	19
ankrd26	19
arhgap27	19
arl11	19
b4galt3	19
c17orf98	19
c20orf194	19
cav1	19
cdkl5	19
cebpb	19
celf5	19
cep164	19
cep350	19
cfap97d1	19
cfc1b	19
chst8	19
clba1	19
cmya5	19
commd1	19
crisp2	19
dlgap4	19
eml6	19
ephx1	19
fhdc1	19
frmpd2	19
gdap1l1	19
gjb2	19
gpr156	19
heatr4	19
kcnmb3	19
kiaa0586	19
kiaa1841	19
kif19	19
kif3a	19
krt36	19
map2	19
mapk4	19
meis2	19
mep1b	19
mtmr11	19
paqr5	19
pax7	19

pfkp	19
pip5kl1	19
rgs9bp	19
rnf223	19
rp11-694i15.6	19
smpx	19
sox21	19
spata45	19
spats1	19
sstr5	19
tctex1d1	19
tecr	19
tmem238	19
tpgs2	19
ttc24	19
ypel1	19
zdhhc17	19
zdhhc2	19
znf605	19
acap3	26
adprm	26
anxa2	26
ap3s1	26
bmp3	26
canx	26
cbln1	26
cd4	26
cldn1	26
dner	26
ecm2	26
ehbp1	26
fam83g	26
gata1	26
Gm2a	26
gnpat	26
icam5	26
il20ra	26
irx2	26
itsn2	26
kcnj5	26
kif5c	26
leprot	26
msmo1	26
myh10	26
nedd4l	26

nhsl2	26
prss8	26
psca	26
slc7a5	26
sos1	26
spire1	26
stard9	26
syt1	26
tubgcp5	26
upf3a	26
vwc2l	26
akt1	27
angel1	27

bri3bp	27
camsap1	27
capn2	27
cpt1a	27
dnajc16	27
dnajc27	27
dnase1l2	27
efcab4b	27
gab3	27
gde1	27
gpd1	27
gpr151	27
kcnn2	27

naa60	27
ncr3lg1	27
nt5c2	27
rab15	27
reps2	27
slc39a9	27
tal1	27
tesk1	27
tmem198	27
tmem246	27
tmem65	27
unc93b1	27
zic3	27

Clusters murins

Gènes	Cluster
acox2	22
armc4	22
au021034	22
bcas1	22
c20orf96	22
c2orf81	22
c5orf49	22
c5orf52	22
c8orf89	22
ccdc13	22
ccdc187	22
ccdc3	22
ccdc81	22
cckar	22
cdhr3	22
cmbl	22
cxcl15	22
cyp2s1	22
dcdc2a	22
ddo	22
dnah5	22
elmod1	22
erich2	22
fam81a	22
fsd2	22
gas2l2	22
gck	22
kcnmb2	22
kiaa2012	22
kndc1	22
liat1	22
lrrc10b	22
lrrc43	22
lrrc46	22
lrrc51	22
lyz1	22
myl4	22
odf3b	22
pacrg	22
pon1	22
rasl10b	22
riiad1	22
scgb3a2	22
smim24	22
smrp1	22
sntn	22
spag16	22
spef1l	22
stk36	22
tcp11	22
tsnaxip1	22
ttll6	22
zfp474	22
1700086L19Rik	24
acot1	24
aoc1	24
arhgdig	24
calml4	24
ccdc114	24
ccdc74a	24
cckbr	24
cdh26	24
cfap299	24
cldn10	24
ctxn1	24
cyp2a12	24
dio1	24
dmbt1	24
dmkn	24
dnah12	24
drc3	24
drc7	24
ephx4	24
fam167a	24
fam216b	24
fgl1	24
fhad1	24
fndc7	24
Gm281	24
gp2	24
lyz2	24
march4	24
nhlrc4	24
prr29	24
scgb1a1	24
scgb3a1	24
sec14l3	24
serpinb12	24
sftpc	24
slc25a18	24
spag6l	24
syt5	24
tctex1d4	24
timp4	24
ugt1a7c	24
vpreb3	24
1600029I14Rik	25
3300002A11Rik	25
6330403K07Rik	25
adam8	25
agrp	25
aldh1a1	25
aldh3b1	25
arhgap18	25
c16orf89	25
bicc1	25
c030048h21rik	25
cacna1h	25
cdhr1	25
cfap46	25
cfap58	25
Cfap97d2	25
clic6	25
diras2	25
klhl3	25
klk11	25
ly6c1	25
map1b	25
neurl1a	25
plin5	25
plvap	25
ppfia3	25
reg3g	25
rgl1	25
rnase4	25
speer4c	25
synm	25
treml1	25
trf	25
tspoap1	25
tulp2	25
uox	25
wdr20b	25
zp3	25

ankfn1	28
bpifb1	28
c1orf87	28
calcoco2	28
cfap77	28
chad	28
dnah3	28
ect2l	28
Elapor1	28
fhl1	28
fsd1l	28
Gm5431	28
iqcg	28
jhy	28
kcnh3	28
lca5l	28
map1a	28
melk	28
nek10	28
p2rx6	28
pigr	28
psg20	28
scrn1	28
sult1d1	28
tm6sf2	28
xylb	28

Cluster 20

Gènes	Cluster
alms1	20
anapc10	20
bora	20
btg3	20
c19orf44	20
cbarp	20
ccdc14	20
ccdc171	20
ccdc50	20
ccdc77	20
cct4	20
cenpj	20
cep295	20
cep44	20
cep57l1	20
cep85	20
cetn2	20
cetn3	20
cntln	20
cntrob	20
cplane1	20
cyp2j2	20
dip2a	20
e2f7	20
ece2	20
eif4enif1	20
farp2	20
fbf1	20
flnc	20
fmr1	20
fos	20
h1f0	20
hesx1	20
hsp90aa1	20
iqcc	20
kcna4	20
kdm7a	20
kiaa0753	20
lrrc45	20
mcm4	20
meioc	20
miip	20
msrb2	20
nfe2l3	20
npnt	20
nptx2	20
pcnt	20
rttn	20
rxfp1	20
sertad4	20
sfi1	20
sgsm1	20
srpk2	20
ssx2ip	20
tp73	20
tuba1b	20
wdr90	20
zbtb18	20

Clusters ciliés

Gène	Cluster
acyp1	8
adrb1	8
ak8	8
ankrd42	8
ankrd45	8
arcn1	8
arl2bp	8
armc9	8
asb13	8
b3gnt7	8
bbs12	8
bbs9	8
c10orf53	8
c11orf65	8
c16orf46	8
c1orf158	8
c6orf120	8
cby1	8
ccdc103	8
ccdc113	8
ccdc18	8
ccdc61	8
ccp110	8
cdc20b	8
cdk20	8
cdk7	8
cep135	8
cep152	8
cep41	8
cep70	8
cep76	8
cep78	8
cep95	8
cetn4	8
cfap20	8
cfap36	8
chfr	8
cidec	8
ckm	8
clec3b	8
cluap1	8
cntn1	8
csf2rb	8
cspp1	8
cyb5d1	8
dcdc2	8
deup1	8
dmrta2	8
dnaaf1	8
dnal1	8
dnal4	8
dpcd	8
dvl1	8
dydc1	8
efcab7	8
egfr	8
elmod3	8
elovl2	8
enkd1	8
epas1	8
erich3	8
errfi1	8
esrra	8
fam177a1	8
fam184a	8
fbxo43	8
fbxo8	8
fbxw9	8
fhod1	8
foxj1	8
foxn4	8
gadd45a	8
galnt4	8
git2	8
gja3	8
grasp	8
heatr6	8
hspb11	8
hyls1	8
idh1	8
ift20	8
ift22	8
ift43	8
ift46	8
ift52	8
ift57	8
ift80	8
ikzf4	8
iqcb1	8
katnal2	8
katnb1	8
kiaa1456	8
kif2a	8
lasp1	8
leng9	8
lexm	8
lrg1	8
lrrc36	8
lrrcc1	8
maats1	8
map9	8
mcidas	8
mid1ip1	8
mink1	8
mks1	8
mok	8
morn2	8
mxd3	8
myb	8
myh11	8
nell1	8
nme5	8
nme9	8
nphp1	8
odf3	8
paqr8	8
pde9a	8
phospho2	8
plk4	8
pltp	8
poc1a	8
poc5	8
ppp1r36	8
prdx6	8
prph2	8
prxl2c	8
rabl2b	8
rcbtb2	8

Clusters ciliés

ribc1	8
rnf219	8
ropn1l	8
rp2	8
rsph1	8
rsph3	8
ruvbl1	8
ruvbl2	8
sccpdh	8
sftpb	8
six2	8
slc30a1	8
slx1a	8
sox7	8
spice1	8
sptlc3	8
sptssb	8
stil	8
stkld1	8
stoml3	8
styxl1	8
tbc1d31	8
tbcel	8
tchp	8
tctex1d2	8
tctn1	8
tekt3	8
tmem117	8
tmem63a	8
tppp3	8
ttc30b	8
ttll1	8
tubgcp3	8
uap1l1	8
usp3	8
was	8
wdpcp	8
wdr31	8
wdr34	8
wdr92	8
wrap73	8
zc2hc1a	8
zmynd10	8
acss2	9
acta2	9

adcy6	9
adprh	9
agbl5	9
agr2	9
agtpbp1	9
aifm3	9
al590867.1	9
aldoc	9
ank3	9
ankrd37	9
ankrd9	9
aqp4	9
arhgef38	9
arl3	9
arsg	9
atp13a4	9
atp2a3	9
atp6v1b1	9
b9d2	9
bbs5	9
bmper	9
bok	9
btbd17	9
c4orf45	9
cab39l	9
cabcoco1	9
cacna2d2	9
cdc25a	9
cep126	9
cfap410	9
cfh	9
chia	9
cib3	9
clmp	9
col28a1	9
cops5	9
cybrd1	9
cyld	9
degs2	9
dkfzp686j19100	9
dock8	9
dthd1	9
dtna	9
dusp14	9
dyrk3	9

efhd1	9
eif2ak2	9
esyt3	9
fam181a	9
fam81b	9
fcgbp	9
fetub	9
foxq1	9
gas6	9
gdpd2	9
ggt6	9
gjb4	9
glb1l2	9
Gmpr	9
gng13	9
gpr155	9
gpr37	9
hes2	9
hist1h2aj	9
hoxa3	9
htr3a	9
ift172	9
itga2b	9
itga4	9
itln1	9
itpkb	9
kcnj16	9
kiaa0825	9
lbh	9
ldlrad1	9
lipg	9
lpar3	9
lrrc31	9
mamdc2	9
map7	9
mdm1	9
morn1	9
mr1	9
msln	9
mtap	9
na	9
napepld	9
neurl	9
nipal1	9
nlrc3	9

nme6	9
nmnat2	9
nphs2	9
npy2r	9
nr2f1	9
nr2f2	9
olig3	9
osbpl3	9
p4htm	9
palm3	9
pcsk5	9
pdlim7	9
pdzd3	9
pir	9
pisd	9
pkhd1l1	9
plac8	9
plxnb3	9
ppp4r4	9
prickle2	9
prkar2b	9
prkg1	9
prom1	9
prr7	9
psmc3ip	9
ptpdc1	9
rab27a	9
rab28	9
rgs22	9
rhof	9
rita1	9
rln3	9
rpgrip1l	9
rsph4a	9
samd11	9
scn5a	9
sgpp2	9
shank2	9
slc16a14	9
slc24a3	9
slc25a22	9
slc27a2	9
slc34a2	9
slc6a1	9
smc1b	9

sox5	9
spata13	9
srcin1	9
st6gal1	9
steap4	9
sv2c	9
syne1	9
syt13	9
tbx1	9
tent5c	9
tgfbr3	9
tmc5	9
tmem11	9
tmem212	9
tmem232	9
tnni1	9
trib2	9
ttc16	9
ttll3	9
ttll7	9
ube2u	9
upk3a	9
usp18	9
wdr49	9
ybey	9
adam22	12
adgb	12
ak1	12
ak9	12
ankef1	12
ankmy1	12
ankrd66	12
armc2	12
bbof1	12
bbs7	12
c21orf58	12
c22orf23	12
c4orf47	12
c6orf118	12
c8orf34	12
caps2	12
casc1	12
cc2d2a	12
ccdc138	12
ccdc148	12

ccdc17	12
ccdc173	12
ccdc181	12
ccdc189	12
ccdc191	12
ccdc24	12
ccdc33	12
ccdc39	12
ccdc42b	12
ccdc57	12
ccdc89	12
cdc14a	12
cdhr4	12
cep128	12
cep83	12
cfap100	12
cfap157	12
cfap300	12
cfap43	12
cfap44	12
cfap47	12
cfap53	12
cfap54	12
cfap61	12
ckb	12
dcdc2b	12
dixdc1	12
dnaaf3	12
dnah6	12
dnali1	12
dydc2	12
dzank1	12
dzip1l	12
efcab12	12
efcab2	12
efcab5	12
efcab6	12
efhb	12
enkur	12
esrrg	12
fam149a	12
fam161b	12
fam166b	12
fam179a	12
fbxo15	12

Clusters ciliés

fbxo16	12
fbxw10	12
ifltd1	12
ift81	12
iqcd	12
iqch	12
iqck	12
kif24	12
kif27	12
kif6	12
lca5	12
lhb	12
lrguk	12
lrrc56	12
lrrc71	12
lrrc73	12
lrriq1	12
mns1	12
nek2	12
nsun7	12
pih1d2	12
ppp1r32	12
ppp1r42	12
rab36	12
rec8	12
rfx2	12
rfx3	12
rhpn1	12
ribc2	12
rnf32	12
rsph10b2	12
rsph14	12
saxo2	12
six3	12
slc7a7	12
spa17	12
spaca9	12
spag17	12
spata17	12
spata6l	12
spef1	12
spef2	12
stpg1	12
tbata	12
tex26	12

tex9	12
thegl	12
tm4sf1	12
traf3ip1	12
trpm2	12
tsga10	12
ttc34	12
ttll13	12
tuba1a	12
usp2	12
vwa3b	12
wdr19	12
wdr38	12
wdr54	12
wdr66	12
wdr78	12
xrra1	12
zbbx	12
zdhhc1	12
agbl2	13
agr3	13
ak7	13
akap14	13
armc3	13
armh1	13
bbox1	13
c11orf16	13
c11orf88	13
c1orf189	13
c1orf194	13
c22orf15	13
c2orf50	13
c2orf73	13
c7orf57	13
c9orf116	13
c9orf135	13
capsl	13
catip	13
ccdc135	13
ccdc146	13
ccdc151	13
ccdc153	13
ccdc170	13
ccdc180	13
ccdc30	13

ccdc40	13
ccdc60	13
ccdc65	13
ccdc78	13
ccdc96	13
ccna1	13
ccno	13
cdkl4	13
cfap126	13
cfap161	13
cfap206	13
cfap221	13
cfap298	13
cfap45	13
cfap52	13
cfap57	13
cfap65	13
cfap69	13
cfap70	13
cfap74	13
daw1	13
dlec1	13
dnaaf4	13
dnah1	13
dnah10	13
dnah2	13
dnah9	13
dnai1	13
dnai2	13
dnajb13	13
drc1	13
dynlrb2	13
efcab1	13
efcab10	13
efhc1	13
efhc2	13
eno4	13
fam161a	13
Fam166c	13
fam183b	13
fam47e	13
fam92b	13
fank1	13
fbxl13	13
fbxo36	13

fsip1	13
hydin	13
iqca1	13
iqub	13
itpka	13
kcnrg	13
kiaa0895	13
kif9	13
klhdc9	13
lrrc18	13
lrrc23	13
lrrc34	13
lrrc48	13
lrrc6	13
lrrc74b	13
mak	13
map3k19	13
mapk15	13
mcf2l	13

mdh1b	13
meig1	13
mlf1	13
morn3	13
morn5	13
mycbpap	13
nek11	13
nek5	13
oscp1	13
pifo	13
ppil6	13
prr18	13
rsph9	13
spag6	13
spag8	13
spata18	13
sphk1	13
stk33	13
stmnd1	13

tcte1	13
tekt1	13
tekt2	13
tekt4	13
tmem107	13
ttc12	13
ttc21a	13
ttc25	13
ttc29	13
ttll9	13
ubxn10	13
ubxn11	13
ulk4	13
vwa3a	13
wdr63	13
wdr93	13
zmynd12	13

Audrey DEFOSSET

Etude évolutive multi-niveaux des gènes de la multiciliation chez les Métazoaires

Résumé

A l'ère des approches à haut-débit où les flux de données sont abondants, bon nombre de processus restent encore largement inexplorés. Les approches intégratives représentent de nouvelles méthodes de choix pour l'étude de tels systèmes biologiques en permettant leur caractérisation selon plusieurs angles. Dans ce contexte, les travaux de cette thèse ont porté sur l'étude de la multiciliation, processus complexe encore méconnu, par l'intégration de données évolutives et fonctionnelles à plusieurs niveaux de granularité. Après avoir établi un bilan évolutif de la multiciliation, ces travaux ont mené à la conception de BLUR, une nouvelle ressource de génomique comparative conçue pour détecter des divergences de séquence atypiques au sein de familles protéiques à l'échelle d'un protéome complet. Son application à la multiciliation au cours d'une approche intégrative couplant génomique comparative et génomique fonctionnelle a permis d'exploiter la signature évolutive particulière de la multiciliation pour identifier de nouveaux gènes candidats multiciliés.

Mots-clés : génomique comparative, évolution, multiciliation, relations génotype-phénotype

Summary

In the era of high-throughput technologies where data flow is abundant, many processes are still poorly characterized. Integrative approaches represent a whole new array of methods particularly well suited to the study of such biological systems by allowing their description through various perspectives. In that context, this thesis focuses on multiciliation, a complex process about which knowledge is still limited, through the integration of both evolutionary and functional data of different levels of granularity. After shedding light on the evolution of known multiciliation genes, this thesis work led to the development of BLUR, a new comparative genomics resource designed to detect atypical sequence divergences within protein families at the whole proteome scale. Its application to multiciliation in the course of an integrative approach combining comparative and functional genomics allowed for the exploitation of the peculiar evolutionary signature of multiciliation to identify new candidate genes linked to that process.

Keywords: comparative genomics, evolution, multiciliation, genotype-phenotype relations