Université de Strasbourg

UNIVERSITÉ DE STRASBOURG

École Doctorale des Sciences de la Vie et de la Santé STRASBOURG

ICube – UMR 7357

THÈSE présentée par :

Yannis NEVERS

soutenue publiquement le : **14 décembre 2018**

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

Discipline/ Spécialité : Sciences de la vie et de la santé, spécialité Bioinformatique

Exploitation de marqueurs évolutifs pour l'étude des relations génotype/phénotype. Application aux ciliopathies.

THÈSE dirigée par :
Mme LECOMPTE Odile — Professeur, Université de Strasbourg

RAPPORTEURS EXTERNES :
M BARBRY Pascal — Directeur de recherche, CNRS
M CHALMEL Frédéric — Chargé de recherche, INSERM Université de Rennes I

EXAMINATEURS EXTERNES :
M BAIROCH Amos — Professeur, Université de Genève, SIB
M DESSIMOZ Christophe — Professeur, Université de Lausanne, University College of London

EXAMINATRICE INTERNE :
Mme FRIEDRICH Anne — Maître de conférences, Université de Strasbourg

Remerciements

Je tiens, en premier lieu, à remercier vivement le Dr Pascal Barbry, le Dr Frédéric Chalmel, le le Pr. Amos Bairoch, le Pr Christophe Dessimoz et le Dr Anne Friedrich pour avoir accepté de participer à mon jury et pour l'honneur qu'ils me font d'examiner mon travail.

Ce manuscrit est l'aboutissement de plus de trois années de recherches, enrichissantes tant du point de vue scientifique qu'humain. Tout au long de cette aventure, j'ai été accompagné par de nombreuses personnes sans qui ce travail ne serait pas ce qu'il est aujourd'hui et je souhaite leur témoigner ma profonde reconnaissance.

Merci à toute l'équipe CSTB, à tous ses membres, actuels ou anciens. Dès mon premier stage, je me suis rendu compte que l'ambiance y est toujours chaleureuse et accueillante et illustre bien que l'on peut faire de la science sérieuse sans trop se prendre au sérieux. Il est impossible de vous rendre justice en quelques lignes tant vous m'avez apporté. Merci Hélène pour m'avoir aidé, dès le départ, à m'intégrer dans l'équipe et m'avoir appris l'art de servir le café. Merci Pierre C, Christian, Pierre P, Anne J et Jean-Sébastien pour ce que vous m'avez appris de vos domaines respectifs, au-delà de la biologie. Merci Wolfgang, Laetitia et Luc pour votre disponibilité et tous les conseils que vous m'avez donnés, des prémisses de ma thèse jusqu'au dernier jour de rédaction. Merci Anne N pour toutes les fois où tu m'as sauvé du purgatoire administratif. Merci Raymond, notre McGiver, pour toutes les fois où tu m'as aidé à m'en sortir avec les machines et pour m'avoir fait profiter de l'immense boîte à outil bio-informatique qu'est Gscope. Merci Carlos, Julio, Pierre, François et les nouveaux venus Romain et Thomas qui ont partagé les tourments et les joies de la vie de thésard. Une petite pensée également à nos stagiaires, Nicolas et Camille, qui m'ont encouragé dans cette dernière ligne droite. Merci Arnaud, Kirsley, Gopal et Audrey, mes compagnons de bureau, vous avez fait du 555, un haut lieu de culture musicale (François Juno RPZ) et d'expertise des mèmes en tout genre. Merci à toi Alexis, pour le passage de flambeau et pour m'avoir permis de contribuer à MyGeneFriends, un projet unique qui te correspond bien. Merci finalement à Benjamin pour ton aide sur OrthoInspector et ton enthousiasme pour mon projet.

Merci aussi à toute l'équipe du LGM, alias « le 9eme » et en particulier Mégana, pour m'avoir fait bénéficier de leur expertise sur le cil et les ciliopathies et pour le précieux travail de manipulation expérimentale, essentiel pour compléter les analyses bio-informatiques.

Je souhaite remercie tout particulièrement trois personnes qui ont joué un rôle majeur dans mes travaux de thèses et qui sont pour beaucoup dans la bonne humeur généralisée du CSTB. Merci donc à Julie, pour toutes les occasions où tu as subi mon anglais et pour ton calme légendaire qui contrebalance bien le tempérament explosif de certains. Sans transition, merci à toi, Olivier, non seulement pour avoir toujours été disponible pour répondre à mes interrogations, mais aussi pour la façon dont tu fais partager ton enthousiasme pour la science et la culture qui va avec. Je ne mentirai pas, tu es parfois difficile à suivre pour le pauvre cornichon que je suis, mais tu restes un exemple à suivre pour moi. Enfin, un immense merci à Odile pour la confiance que tu

m’as accordée, ta patience(s) et pour ton encadrement sans faille et sans reproche. C’est toi qui m’as fait en grande partie découvrir le monde merveilleux de la génomique comparative, ouvert les portes de l’équipe et conseillé et encouragé dans la voie de l’enseignement. C’est déjà beaucoup, mais tu as aussi réussi l’équilibre délicat de me guider dans mes projets en me laissant toute mon autonomie. J’ai eu beaucoup de chance d’avoir des « chefs » comme vous et mes mots ne sauront exprimer toute ma reconnaissance. J’espère bien que l’on aura de nombreuses occasions à l’avenir, de travailler à nouveau ensemble.

Je veux aussi remercier les personnes qui m’ont accompagné sur le plan personnel pendant toute la thèse, m’ont aidé à sortir la tête de l’eau quand il le fallait et m’ont encouragé pendant les *rushs*. Merci donc, à Agathe et Max, Adeline et Greg (et au petit Clément), souvent loin des yeux, mais jamais loin du cœur. Merci à Alexia, pour avoir tenté de me garder en bonne santé avec du sport et du quinoa, c’était bien essayé. Merci à Audrey, Jean-Michel, Remi, Kévin, Cindy, Anja, Paul le Grand, Hélène, Lucile, Laura, et Paul le Roux pour les innombrables soirées strasbourgeoises, autour d’une bière ou d’un jeu (sérieux) et pour votre humour à géométrie variable. Merci à Raphaël et à Camille, enfin, pour votre soutien indéfectible depuis presque 15 ans.

Pour finir, j’adresse mes remerciements à ma famille, en particulier, Aymeric qui me précède de quelques années dans cette aventure, Flora et enfin toi, maman. Tu m’as toujours accompagné dans mes choix et tu as beaucoup donné pour que je réussisse, si j’en suis ici aujourd’hui, c’est en grande partie grâce à toi et je ne te remercierai jamais assez.

Table des matières

Liste des figures

Liste des tableaux

Abréviations

ADN	Acide désoxyribonucléique
API	*Application Programming Interface*
ARN	Acide ribonucléique
ARNnc	Acide ribonucléique non codant
ARNm	Acide ribonucléique messager
ARNr	Acide ribonucléique ribosomal
ARNt	Acide ribonucléique de transfert
BBS	*Bardet Biedl Syndrome*
BLAST	*Basic Local Alignment Search Tool*
CDS	*Coding sequence*
GO	*Gene Ontology*
GWAS	*Genome Wide Association Studies*
HGT	*Horizontal Gene Transfer*
HPO	*Human Phenotype Ontology*
IFT	*IntraFlagellar Transport*
MKS	*Meckel Syndrome*
NPHP	*Nephronophthisis*
OMIM	*Online Mendelian Inheritance in Man*
PCD	*Primary Ciliary Dyskinesia*
QFO	*Quest for Orthologs*
RBH	*Reciprocal Best Hit*
RDF	*Ressource Descriptions Framework*
SPARQL	*SPARQL Protocol and RDF Query Language*
SQL	*Structured Query Language*
WGD	*Whole Genome Duplication*

Avant-propos

L'objectif de ces travaux de thèse, est l'étude des relations génotype-phénotype à travers le prisme de l'évolution, par des approches de génomique comparative. Ils ont abouti, *in fine*, au développement d'outils complémentaires permettant d'explorer sous plusieurs angles le rapport entre l'histoire évolutive des gènes et leurs fonctions. Ces outils ont été intégrés à deux ressources publiques, OrthoInspector 3.0 et MyGeneFriends, et appliqués à l'étude d'une classe de pathologies génétiques, les ciliopathies.

Les chapitres d'introduction sont au nombre de trois, un pour chacun des axes essentiels sur lesquels s'appuient mes travaux.

Le premier pose le contexte général dans lequel s'inscrit cette thèse, l'étude des relations génotype-phénotype. Je commencerai par donner le contexte historique en exposant comment la compréhension des relations génotype-phénotype a évolué au cours du temps puis aborderai la façon dont l'émergence des technologies à haut-débit en a révolutionné l'étude. J'illustrerai ces progrès en décrivant les principales technologies 'omiques' et la manière dont chacune permet d'étudier un aspect distinct du passage du génotype au phénotype. J'aborderai ensuite les défis associés aux données massives issues de ces technologies et surtout les opportunités ouvertes par l'intégration des données omiques, qui permettent de reconstituer une image complète des systèmes biologiques à partir de données parcellaires mais complémentaires. La notion d'intégration est en effet essentielle pour comprendre mes travaux, dont l'objectif est de concevoir des marqueurs évolutifs à même d'être exploités dans cette logique.

Le second chapitre décrit le cadre conceptuel dans lequel je me suis placé pour l'élaboration de marqueurs évolutifs à savoir, le champ de la génomique comparative qui permet l'analyse de données massives sous l'angle de l'évolution. J'y définirai les notions de base : le concept d'homologie et ce qu'il permet d'apprendre sur la fonction des éléments génomiques. Surtout, je montrerai les multiples possibilités d'exploitation des données génomiques massives offertes par la génomique comparative pour en apprendre plus sur le Vivant. Je détaillerai, en particulier, les exploitations ayant trait à l'étude des relations génotype-phénotype, avec le profilage phylogénétique, représentation de l'histoire évolutive des gènes que j'utiliserai dans mes travaux comme marqueur évolutif. J'illustrerai concrètement le potentiel de ce dernier en présentant le cil eucaryote, dont l'étude a grandement bénéficié de la génomique comparative et qui est emblématique de la complexité des relations génotype-phénotype.

Finalement, le dernier chapitre s'intéresse en détail à l'inférence des relations d'orthologie, préalable à toute étude de génomique comparative et donc, au profilage phylogénétique. Je décrirai les méthodes existantes permettant de réaliser cette inférence, et les problématiques qui y sont liées dans le contexte des données massives. Dans ce chapitre, j'attribuerai une place importante aux ressources publiques dédiées aux relations d'orthologie et surtout, à la façon dont ces ressources permettent de représenter synthétiquement et de contextualiser ces données pour en guider l'interprétation. Ces deux aspects ont orienté en grande partie mes développements autour des marqueurs évolutifs.

La section Contributions s'articule elle aussi en trois chapitres ; les deux premiers décrivent la conception et l'exploitation de marqueurs évolutifs tandis que le dernier montre son application concrète à l'étude des ciliopathies.

Le premier chapitre décrit la façon dont j'ai exploité les données génomiques massives pour en extraire une information synthétique à travers la plateforme d'orthologie d'OrthoInspector 3.0, qui sert de socle à l'ensemble des outils d'exploitation développés par la suite. Je décrirai la façon dont j'ai répondu aux différentes problématiques des données massives, à savoir la sélection de données pertinentes et de qualité, l'organisation des données et la gestion technique du flux de données, dans l'optique d'une plateforme de génomique comparative complète. Finalement, je mettrai en avant l'importance de la représentation synthétique des relations d'orthologie pour en faciliter l'exploration et faire de l'histoire évolutive un marqueur aisément interprétable.

Le second chapitre détaille les méthodes et outils que j'ai développés pour exploiter ces marqueurs évolutifs selon trois grands axes : passer de l'histoire évolutive d'un gène au phénotype qu'il induit, éclairer un phénotype par l'histoire évolutive des gènes qui y sont associés et finalement, relier les gènes partageant une histoire similaire. Je décrirai l'implémentation des outils correspondants dans OrthoInspector 3.0 et illustrerai ces outils et leur complémentarité par des applications à l'étude des relations génotype-phénotype. Avec la dernière partie de ce chapitre, je montrerai comment le marqueur évolutif peut être exploité de concert avec d'autres types de données, dans le contexte de MyGeneFriends, un réseau social reliant gènes, maladies et chercheurs.

Les deux plateformes dont il est question dans ces deux premiers chapitres, OrthoInspector 3.0 et MyGeneFriends ont fait l'objet de publications dans des revues avec comité de lectures. Ces publications sont disponibles en annexe de la thèse.

Le troisième et dernier chapitre est consacré à l'exploitation multi-niveaux des profils phylogénétiques dans le cadre de l'étude du cil et des ciliopathies. Ces travaux ayant fait l'objet d'une publication détaillée, celle-ci constitue l'essentiel du chapitre. Je reviendrai sur ces travaux en illustrant concrètement l'importance de la représentation des données et la complémentarité des divers outils d'exploitation, thèmes vus aux chapitres précédents.

Un chapitre Matériel et méthodes est disponible à la suite de la section contribution, détaillant les sources de données utilisées au cours de cette thèse, les détails techniques des plateformes et outils auxquels j'ai contribué ainsi que les protocoles de manipulation de données que j'ai suivis.

A la lumière de l'expérience acquise au cours de cette thèse, je conclurai sur l'importance de la représentation et la contextualisation des données omiques. J'y ajouterai une réflexion sur les multiples possibilités d'exploitation de l'histoire évolutive pour comprendre les systèmes biologiques, en esquissant des perspectives d'extension des marqueurs évolutifs pour l'étude du Vivant.

Introduction

1 Les relations génotype-phénotype : des premiers pas à la révolution des *omics*

La notion de génotype désigne l'ensemble de l'information héréditaire d'un individu, porté par le génome. Pendant longtemps, son contenu ainsi que son support, sont restés inaccessibles. A l'opposé, tous les traits observables ou qu'il est possible de mesurer à l'échelle macroscopique comme microscopique chez un individu constituent son phénotype. Le phénotype résulte de l'interaction de deux facteurs, le génotype d'un individu et son environnement. Les relations génotype-phénotype décrivent les associations qui existent entre une partie de l'information génétique, les gènes ou d'autres éléments génétiques, et l'apparition d'un trait phénotypique donné. En s'intéressant aux relations génotype-phénotype, on cherche donc à comprendre le rôle de chaque gène dans les mécanismes biologiques et dans le cadre médical, comment certaines variations de ces gènes peuvent conduire à une pathologie.

Dans ce chapitre, je retrace la façon dont la compréhension des mécanismes liant génotype et phénotype a évolué avec notre compréhension du vivant. Je montre notamment comment les technologies à haut débit, dites *omics,* ont constitué une révolution pour l'étude de ces mécanismes tout en en révélant toute la complexité. Je termine par un tour d'horizon des méthodes modernes conçues pour évaluer cette complexité, basées sur l'intégration de l'ensemble des informations à disposition.

1.1 L'émergence des relations génotype-phénotype

1.1.1 La transmission des caractères : naissance du concept de gène

Dès les premiers temps de l'histoire, l'idée qu'une information cachée ordonne l'apparition des caractères physiques et observables était intimement liée au concept d'hérédité. Il était évident qu'un enfant ressemble le plus souvent à ses deux parents, voire à ses grands-parents. Dans un autre contexte, l'élevage montrait qu'il est possible d'obtenir des races d'animaux douées d'un caractère voulu en faisant les croisements judicieux. Les principales théories de l'hérédité qui ont imprégné pendant longtemps l'histoire de la biologie remontent à l'Antiquité. Selon Hippocrate (460-377 av. J-C.), chaque partie du corps d'un individu adulte produit des fluides qui convergent dans les organes reproducteurs et se mélangent lors de la reproduction pour déterminer les caractéristiques de l'enfant. Ultérieurement, Aristote (384-322 av. J-C) formulait l'idée selon laquelle la « graine », la contribution de l'homme, générait la forme permettant de mouler la « matière », la contribution de la femme. Ces deux modèles cherchaient avant tout à expliquer les mécanismes de la reproduction, mais aussi comment les caractéristiques des deux parents étaient observables chez l'enfant. Pour autant, les traits héritables dérivaient seulement du phénotype de parents et la notion de génotype n'étant pas présente, il ne s'agissait pas encore de théories de l'hérédité au sens où nous l'entendons de nos jours. Ces deux modèles, bien que

ne reposant sur aucune base expérimentale, furent les vues dominantes sur la question pendant près de 2000 ans (Cobb, 2006). L'observation directe des cellules germinales par Leeuwenhoek (1632-1723) pour le spermatozoïde, puis par Von Baer (1792-1876) pour l'ovule, fournit une connaissance plus complète de la reproduction, mais apporta peu d'éléments sur la façon dont ces gamètes transmettent l'information à la descendance.

Les bases théoriques des mécanismes de l'hérédité furent véritablement posées au XIXe siècle, avec les célèbres travaux de Gregor Mendel (1822-1884). Son corpus d'expérimentations repose sur le croisement de lignées pures de pois de Valence sur plusieurs générations et l'étude de la ségrégation de caractères discrets au cours des générations. Le résultat de ces expériences permit d'affirmer les premiers concepts clés en génétique : les caractères observables des êtres vivants sont déterminés par des facteurs discrets et chaque organisme possède une paire de ces facteurs, qui peuvent exister sous différentes formes, les allèles. Ces facteurs se transmettent selon trois lois, appelées lois de Mendel : 1. Ils sont transmis à la descendance par les gamètes, portant seulement un allèle de chaque facteur. 2. Ils ségrègent de façon indépendante lors de la formation du gamète. 3. Certains allèles ont un mode d'expression dominant sur les autres. L'allèle dominant détermine l'effet du caractère.

En séparant les traits observés et les facteurs porteurs de l'information, Mendel marque donc un tournant dans la compréhension de l'hérédité en séparant l'information du caractère observé. Ses travaux n'auront cependant un impact qu'un siècle plus tard. En effet, c'est en 1909, en s'inspirant, entre autres, des travaux de Mendel, que Wilhelm Johansen définit pour la première fois l'idée moderne de gène, équivalent des 'facteurs' mendéliens. De là, il définit le génotype qui comprend l'ensemble des gènes dans un gamète ou un zygote. Il oppose le génotype au phénotype, qui est l'ensemble des caractères observables d'un individu. Dans cette définition, les traits ne se transmettent jamais d'un individu à l'autre, ce sont uniquement les gènes qui sont transmis et qui définissent ensuite le phénotype. Pour autant, Johansen prenait déjà note que l'environnement joue un rôle important dans cette relation et suggérait l'idée qu'un trait phénotypique résulte non seulement de l'effet de plusieurs gènes, mais aussi de l'interaction entre ces gènes et l'environnement.

1.1.2 Hérédité et évolution

Au XIXème siècle, une autre avancée considérable fit date dans l'histoire de la compréhension des relations génotype-phénotype : la théorie de l'évolution des espèces par sélection naturelle. Cette avancée ne répondait pas à la question de la transmission des caractères d'un individu à l'autre, mais à celle de l'apparition des espèces. A cette époque, la notion essentialiste selon laquelle toutes les espèces d'êtres vivants ont été créées en l'état par Dieu et ne subissent pas de transformation était remise en cause par l'étude des fossiles par des naturalistes, dont Georges Cuvier (1769-1832), qui remarqua que les espèces que l'on trouve à l'état de fossile n'ont pas d'équivalents modernes. Cela donna naissance au mouvement du catastrophisme, qui postulait que des catastrophes majeures mènent à l'extinction comme à la création d'espèces. Le catastrophisme restait toutefois fixiste, c'est-à-dire qu'il prévoyait que les espèces ainsi créées ne subissent pas de modification au cours du temps.

C'est dans ce contexte qu'en 1803, Jean-Baptiste Pierre Antoine de Monet, chevalier de Lamarck, publia sa *Philosophie zoologique* qui lui vaudra d'être considéré comme le fondateur du transformisme. Dans ce système, il formula l'idée que les caractères acquis et conservés par un individu au cours de sa vie, sont transmis aux générations suivantes et affectent leur descendance. Selon cette idée, un organe dont l'usage est fréquent au cours de la vie d'un individu se trouvera renforcé chez ses descendants, quand un organe peu utilisé finira par disparaître. Il s'agit ici de la première théorie impliquant un 'changement' des espèces au cours du temps, bien que dans cette conception, c'est en premier lieu le phénotype des individus qui influe sur leur génotype.

En 1859, Charles Darwin publia son ouvrage *De l'origine des espèces par le moyen de la sélection naturelle* dans lequel il développa sa théorie de l'évolution. Celle-ci repose sur l'observation qu'il existe des variations de caractères entre individus d'une même espèce et que ces variations peuvent être transmises à leurs descendants. En faisant le parallèle avec la façon dont ces caractères sont artificiellement sélectionnés par les éleveurs pour modeler les lignées selon leurs besoins, Darwin fit l'hypothèse que des variations peuvent être naturellement sélectionnées et que c'est la fixation de multiples variations de ce genre qui donne naissance aux espèces. L'hypothèse de la sélection naturelle implique que les individus dont les variations sont avantageuses dans leur environnement au sens large, se reproduiront plus efficacement et donneront naissance à plus de descendants que d'autres individus non-avantagés. Si les conditions se maintiennent, le nombre d'individus avantagés croitra et la variation se transmettra donc à toute la population.

La théorie de l'évolution de Darwin, contrairement au transformisme, n'impliquait pas que le phénotype influe directement sur le génotype. Ce sont des variations héritables qui sont sélectionnées au cours des générations, mais l'on n'en connaissait pas encore le support physique.

1.1.3 Le support physique des gènes

C'est au début du XXe siècle que la question du support des gènes fut réellement abordée. En 1902, Walter S. Sutton (Sutton, 1903) observa que les chromosomes, dont l'existence était connue depuis le siècle précédent, se conduisent de façon similaire aux facteurs de Mendel. Notamment, les paires de chromosomes se séparent pour se répartir parmi les gamètes. Il fit alors l'hypothèse que ces structures sont les porteuses des gènes, à raison de plusieurs gènes par chromosome. Cette théorie permettait également d'expliquer pourquoi les gènes donnant lieu à certains caractères semblent ne pas ségréger de manière indépendante et elle fut étayée dans les années suivantes. La découverte du déterminisme sexuel par les chromosomes sexuels Y et X chez le coléoptère *Tenebrio molitor* par Nettie Stevens en 1902 suivie par des découvertes similaires d'Edmund Wilson (Wilson, 1905), posa le fondement des célèbres travaux de Thomas Hunt Morgan, reconnu comme un des pères de la génétique moderne. Ses croisements de drosophiles lui permirent de démontrer l'existence de caractères morphologiques dont la ségrégation est liée au sexe et ainsi attester de la présence de ces gènes

sur les chromosomes sexuels. Ses travaux aboutirent également à la première carte chromosomique, plaçant pour la première fois les gènes en tant qu'unité physique sur le chromosome.

Le XXème siècle finit d'apporter une compréhension sur la nature réelle de l'information génétique dans le chromosome. Les expériences de Avery-MacLeod-McCarty (Avery et al., 1944), puis de Hersey-Chase (Hershey et Chase, 1952) conclurent définitivement que le support de l'information génétique chez les êtres vivants est la molécule d'ADN et non pas l'autre composant des chromosomes, les protéines pourtant plus complexes chimiquement et longtemps pressenties. En 1953, James Watson et Francis Crick (Watson et Crick, 1953) permirent définitivement de conclure la recherche du support de l'information génétique en proposant la structure de la double hélice d'ADN, composée de deux brins complémentaires, sur la base des clichés de diffraction électronique de Rosalind Franklin.

1.1.4 Du gène au phénotype ?

Ces informations en main, comprendre les relations génotype-phénotype revient à comprendre quels sont les mécanismes permettant de passer de la molécule d'ADN à l'expression d'un phénotype donné. Exprimé autrement, comment l'on passe de l'ensemble des gènes à un organisme complet et fonctionnel.

En 1958, Francis Crick (Crick, 1958) fera progresser la compréhension de ces mécanismes en formulant pour la première fois une ébauche de la théorie centrale de la biologie. La séquence d'ADN, linéaire, porte le code nécessaire pour synthétiser les protéines, qui sont les réelles effectrices de la cellule. Il proposa notamment que le code génétique s'exprime sous forme de triplets. L'existence de l'ARN messager (ARNm), permettant de porter l'information génétique de l'ADN dans le cytoplasme, fut théorisée trois ans plus tard par Jacob et Monod (Jacob et Monod, 1961) et démontrée la même année par Gros (Gros et al., 1961). Le code génétique fut intégralement déchiffré quelques années plus tard par Nirenberg (Martin et al., 1962), venant compléter le schéma du passage du gène à la protéine : l'ADN, porteur du gène et capable de s'auto-répliquer, sert de patron pour l'ARNm, qui est ensuite traduit par le ribosome en protéine, selon un code de correspondance entre triplets d'acides nucléiques et acides aminés. A partir de là, il serait tentant de penser que la relation entre les gènes, les protéines et en définitive le phénotype, est linéaire (Un gène, une protéine, une fonction) mais les mécanismes sont bien sûr plus complexes que cela.

Les avancées majeures en génétique, notamment les découvertes de Mendel et Morgan, reposent sur l'étude de caractères liés à un gène seul. Il s'agit cependant de cas particulier, un trait phénotypique étant le plus souvent associé à plusieurs gènes interagissant dans un système. Cette intuition était d'ores et déjà centrale pour Johannsen (Johannsen, 1923):

"*But however far we may proceed in analysing the genotypes into separable genes or factors, it must always be borne in mind, that the characters of the organisms - their phenotypical features - are the reaction of the genotype* in toto*. The Mendelian units as such, taken* per se *are powerless."*

Une meilleure compréhension du mécanisme d'expression des gènes n'a pas permis de surmonter cet obstacle, principalement à cause de la complexité des systèmes biologiques, que l'on retrouve à tous les niveaux de la théorie centrale, des mécanismes régulant l'expression des gènes à ceux assurant les modifications des protéines en passant par l'épissage des ARNm. Cette complexité s'est révélée progressivement au cours des 60 années de recherche qui ont suivi la découverte de l'ADN et a été pleinement mise en lumière par l'essor des omiques.

1.2 La révolution des omiques

Le suffixe '-omique' désigne les méthodologies dites à haut-débit, dont le développement a été très rapide durant les 20 dernières années. Fondés sur le mot génome, historiquement l''ensemble des gènes', les omiques définissent 'l'étude de l'ensemble'. De cette manière, la transcriptomique correspond à l'étude de l'ensemble des transcrits et la protéomique à l'étude de l'ensemble des protéines. Les méthodes omiques permettent l'étude poussée des systèmes biologiques à chaque niveau, depuis le contenu du génome d'une cellule jusqu'à l'ensemble des protéines qui y sont produites.

Dans cette section, j'aborde les principales méthodologies et technologies que l'on réunit sous l'appellation omique ainsi que leurs apports respectifs dans l'étude des relations génotype-phénotype. Ces descriptions n'ont pas vocation à être exhaustives, ainsi je n'aborderai pas, par exemple, la métabolique ni la lipidomique, mais elles cherchent à illustrer par des exemples concrets la grande variété des mécanismes impliqués dans les relations génotype-phénotype que l'on peut aujourd'hui explorer à l'aide de ces méthodes (Figure 1-1).

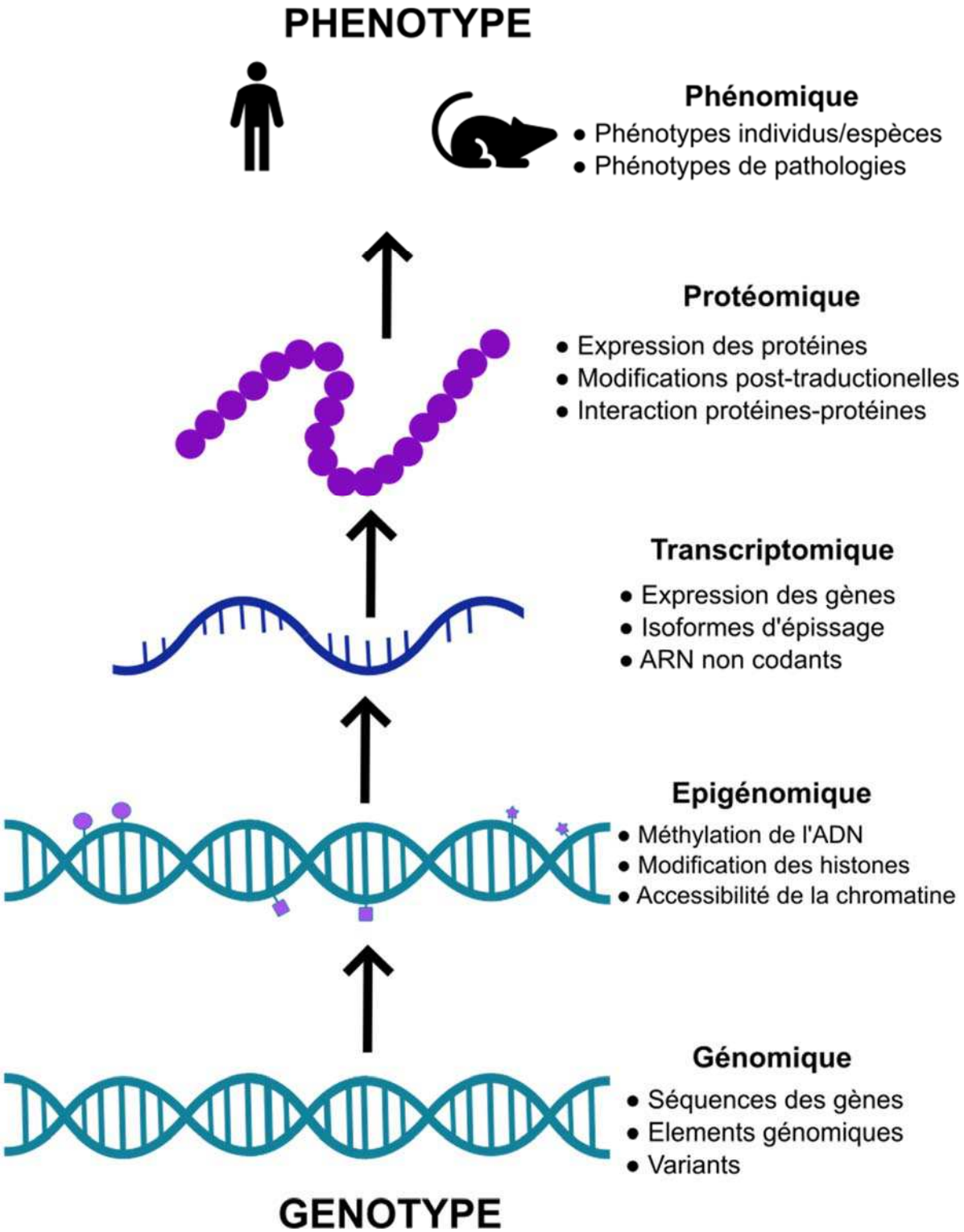


Figure 1-1 **Schéma du flux d'informations modulant la relation génotype-phénotype et les omiques associées**. Les différents niveaux de ce flux sont associés à des omiques et aux mécanismes qu'elles permettent d'étudier. La diversité de ces mécanismes reflète la complexité des processus impliqués.

1.2.1 La génomique

La génomique est parfois considérée comme « la mère de toutes les omiques » : la première à s'être développée et à avoir atteint la maturité, mais aussi celle qui s'intéresse directement au génotype (Figure 1-1). Ici, je présente brièvement les avancées en termes de séquençage ADN ayant permis son émergence et les diverses applications auxquelles elle a donné naissance.

1.2.1.1 Brève histoire du séquençage

Les premières techniques de séquençage, visant à retrouver l'enchainement des bases nucléotidiques constituant un gène ou un génome, furent développées dans les années 1960, soit dix ans après la découverte de l'ADN. Le développement des méthodes permit le séquençage d'objet de taille de plus en plus importante. Ainsi, la première séquence d'acide nucléique complètement déterminée fut celle de l'ARN de transfert (ARNt) de *Saccharomyces cerevisiae* en 1965 (Holley et al., 1965) ; la première séquence d'ARN ribosomale, l'ARN 5S de *Escherichia coli* en 1967 (Brownlee et al., 1967) ; le premier génome à ARN, celui du bactériophage MS2 en 1976 (Fiers et al., 1976) et le premier génome à ADN, celui du bactériophage ϕX174 (Sanger et al., 1977). Le séquençage par terminaison de chaine, ou séquençage de Sanger fut la première technique à être largement répandue et donna lieu aux premières machines de séquençage. Ce sont des variantes de ce protocole de séquençage, dit de première génération, qui permirent ensuite le séquençage du premier génome complet bactérien en 1995 (*Haemophilus influenzae*) (Fleischmann et al., 1995) et du premier génome eucaryote (*Saccharomyces cerevisiae*) en 1996 (Goffeau et al., 1996). En 2001, la première version du génome humain fut finalement obtenue par séquençage de première génération, au terme de plusieurs années de travail (Lander et al., 2001).

Malgré ces réussites, les techniques de première génération restaient sujettes à des limitations techniques rendant difficile le séquençage de génome à grande échelle, notamment pour des raisons de coût. Pour cette raison, les séquenceurs de seconde génération, illustrés par les techniques de pyroséquençage proposées par 454 (Margulies et al., 2005) et la méthode Solexa d'Illumina (Bentley et al., 2008) déclenchèrent un changement de paradigme dans l'étude des génomes. Basés sur le séquençage parallèle de millions de lectures courtes (400-500 paires de bases pour 455 et 35 paires de bases pour Illumina), ils permirent d'augmenter considérablement la quantité d'ADN pouvant être séquencé en une étape, réduisant dans le même temps le coût des projets de séquençage. A titre d'exemple, le séquençage du génome du Dr. Watson avec des techniques de première génération s'étendit sur plusieurs années pour un coût de 100M$ (Levy et al., 2007), le même génome fut séquencé en deux mois seulement pour 1M$ (Wheeler et al., 2008) avec des technologies de seconde génération. La troisième génération de séquenceurs, menée par Pacific Bioscience (Eid et al., 2009) et Oxford Nanopore Technologies (Clarke et al., 2009), plus récente, permet l'obtention de lectures longues (en moyenne 10 kpb, et jusqu'à 882 kpb (Jain et al., 2018)) dont la taille facilite l'assemblage des génomes. Si elles sont encore sujettes à un taux important d'erreurs dans la caractérisation de la séquence (5 à 15%), elles sont utilisées en complément des techniques de seconde génération pour combiner les avantages des lectures longues à la précision plus élevée des lectures courtes

(Jain et al., 2018; Johnson et al., 2018). Avec l'avènement de ces technologies, le nombre de projets de séquençage a véritablement explosé durant la dernière décennie, ouvrant la voie à des exploitations du génome impossible auparavant ainsi qu'au phénomène des données massives en biologie (Stephens et al., 2015) et par là, au champ des omiques.

1.2.1.2 La génomique pour les relations génotype-phénotype

Le séquençage d'un génome complet consiste, très schématiquement à fractionner des échantillons d'ADN prélevés d'un tissu ou d'une culture d'organismes, éventuellement en amplifier les fragments, puis à les séquencer. Les séquences ainsi obtenues sont ensuite assemblées informatiquement de façon à reconstruire le génome complet.

La possibilité d'obtenir la séquence du génome diploïde d'un individu, soit son génotype, est évidemment d'un intérêt considérable dans l'étude de son influence sur le phénotype. Une des principales répercussions de la démocratisation des techniques de séquençage est l'essor considérable de leur utilisation dans le domaine du diagnostic où il s'agit d'identifier le ou les variants génétiques responsables des pathologies héréditaires. Le génome d'un individu humain comptant, en moyenne, plus de 3 millions de variants par rapport au génome de référence (Shen et al., 2013), cette recherche se concentre sur les variants peu fréquents dans les populations humaines. On les compare pour cela à des données de référence, comme celles obtenues dans le cadre du projet 1000 génomes (1000 Genomes Project Consortium et al., 2015), recensant des dizaines de milliers de variants observés sur des milliers d'individus. Plus de 85% des variants à l'origine de maladies étant contenus dans les exons (Rabbani et al., 2014) qui représentent moins de 1,5% du génome, le séquençage d'exomes (l'ensemble des exons) est préférée aux génomes complets pour ce type d'analyse.

Avec les réductions de coût, cependant, de plus en plus d'initiatives visent à l'acquisition de données de génomes complets à grande échelle. Pour une illustration parlante, (Telenti et al., 2016) ont récemment séquencé plus de 10 000 génomes humains et des initiatives plus ambitieuses encore ont été lancées telles que le Projet 100 000 génomes au Royaume-Uni ou le Plan France Médecine génomique en France, visant à obtenir à terme plusieurs centaines de milliers de séquences génomiques par an. Ces résultats sont donc amenés à se multiplier dans le futur, rendent imaginables les comparaisons de génomes à grande échelle du type des *genome wide association studies* (GWAS). Brièvement, ces études ont pour principe de corréler des informations génotypiques à l'existence d'un phénotype sur un échantillon important d'individus. Si ces études sont classiquement réalisées sur des marqueurs de *loci* chromosomiques et non pas sur le génome entier, l'accessibilité croissante des techniques de séquençage à haut-débit les rendent à présent envisageables au niveau génomique (van Rheenen et al., 2016).

1.2.1.3 Séquençage de cellule unique

Une des avancées technologiques les plus marquantes de ces dernières années en termes de séquençage est probablement l'apparition de séquençage de cellules uniques. Comme leur nom

l'indique, ces techniques prennent avantage des technologies d'isolation des cellules, notamment la micro fluidique, pour extraire l'ADN d'une seule cellule, puis l'amplifier avant de la séquencer (Gawad et al., 2016). Bien que la qualité des génomes obtenus de cette façon ne soit pas au niveau de ce que l'on peut obtenir par séquençage classique, du fait de la quantité réduite de matériel, le séquençage de cellules uniques offre des opportunités inédites. Il est notamment possible d'évaluer les changements introduits dans le matériel génétique dans les cellules d'un organisme multicellulaire au cours de sa vie et donc, les variations que l'on retrouve au sein d'un même organisme. Un des exemples iconiques est la possibilité de suivre la trajectoire génétique des cellules cancéreuses pendant le développement des tumeurs (Kim et Simon, 2014). Il s'agit, à ce titre, d'une technologie unique pour étudier les marqueurs génotypiques de la prolifération cancéreuse.

1.2.1.4 Métagénomique

La métagénomique, une variante des méthodes d'analyse génomique, consiste à séquencer non pas le génome d'un organisme donné, mais de l'ensemble des organismes provenant d'un échantillon environnemental (Quince et al., 2017). Brièvement, les protocoles de métagénomique nécessitent de prélever un échantillon de l'environnement d'intérêt, d'en extraire l'ADN, de le fragmenter afin de générer les bibliothèques de séquençage. Une fois le séquençage réalisé, les séquences sont assemblées et attribuées aux différentes espèces présentes dans l'échantillon, par exemple en les comparant à des bases de données d'espèces connues ou en se basant sur la composition en bases nucléiques de l'échantillon.

Les techniques de métagénomique permettent l'acquisition de séquences d'un nombre important d'organismes microscopiques comme les bactéries, les archées, les protistes et même les virus. Elles rendent possible d'interroger des biosphères d'une rare diversité : l'ensemble des océans accessibles à l'homme (Sunagawa et al., 2015), les sols, les environnements extrêmes, mais aussi des environnements associés à des espèces tels que la peau et le système digestif des grands vertébrés.

De ce point de vue, la métagénomique apporte un moyen d'étudier les relations génotype-phénotype en santé humaine. En effet, on estime que chez l'homme, les communautés bactériennes représentent au minimum autant de cellules que les cellules proprement humaines (Gilbert et al., 2018). A ce titre, la faune bactérienne d'un individu, ou microbiome, est un facteur environnemental considérable affectant des traits phénotypiques tels que l'absorption de nutriments ou le système immunitaire (Schirmer et al., 2016). L'analyse des métagénomes intestinaux peut ainsi permettre d'identifier des marqueurs de pathologies non-transmissibles comme l'obésité (Ley et al., 2006).

1.2.1.5 Epigénomique

Le terme épigénétique rend compte de modifications héritables de l'expression du génotype, qui n'en modifient pas le contenu. Ces modifications correspondent à des modifications de l'état de la chromatine, induites par des altérations chimiques (méthylations, acétylations…) de

l'ADN ou des histones. Les changements épigénétiques peuvent avoir lieu à des stades précis du développement (ce qui permet la différenciation cellulaire) ou en réponse à des conditions environnementales. Du fait des modifications épigénétiques, deux individus ou cellules portant le même génotype peuvent développer des phénotypes différents, à l'image des types cellulaires chez les métazoaires.

L'épigénomique est l'omique dédiée à l'étude des modifications épigénétiques sur l'ensemble du génome. Schématiquement, les méthodes d'épigénomique reposent sur le séquençage ciblé d'échantillons du génome traités préalablement pour en isoler les régions de chromatine ouverte, accessibles à la transcription (DNAse-seq) ou les régions portant des modifications ciblées (ChipSeq) (Friedman et Rando, 2015). Ces techniques sont également utilisées en dehors du contexte de l'épigénomique pour identifier les régions où se fixent les facteurs de transcription régulant l'expression des gènes.

L'état de l'épigénome est spécifique d'un type cellulaire ou d'un tissu donné, et la détermination d'épigénomes de référence pour l'ensemble des tissus humains est actuellement un objectif majeur d'efforts internationaux, dans le cadre d'initiatives comme le *International Human Epigenome Consortium* (Stunnenberg et al., 2016). Ces références sont bien sûr importantes pour comprendre comment un même génotype donne lieu à des phénotypes différents au niveau cellulaire.

Les modifications épigénétiques, en définitive, régulent la façon dont les gènes sont exprimés dans une cellule donnée. Ils se situent de cette façon à un niveau plus élevé que la génomique sur l'échelle passant du génotype au phénotype (Figure 1-1). Les autres omiques que nous allons voir par la suite s'élèvent encore sur cette échelle en ne s'intéressant plus aux gènes mais à leurs produits.

1.2.2 La transcriptomique

La transcriptomique consiste en l'étude de l'ensemble des transcrits, les molécules d'ARN, présents dans une ou plusieurs cellules à un moment donné. Le domaine de la transcriptomique a émergé dans le début des années 2000 avec l'apparition des puces à ADN ((Lowe et al., 2017) pour une revue sur les technologies transcriptomiques). Ces puces présentent à leur surface des milliers de séquences nucléotidiques de taille réduite, appelées sondes, et permettent par hybridation aux transcrits étiquetés avec un fluorophore d'identifier les transcrits présents dans un échantillon et de les quantifier dans une certaine mesure. Depuis une dizaine d'années cependant, la transcriptomique utilise également les technologies de séquençage de deuxième génération, avec l'avènement du *RNA-seq*. Brièvement, ces techniques sont basées sur le séquençage des transcrits présents dans un échantillon qui sont ensuite alignés à un génome de référence.

Contrairement à la génomique où l'information est stable dans le temps, la transcriptomique permet d'obtenir une vision de la dynamique des processus cellulaires. Par exemple, des mesures à différents stades du développement offrent une vision des variations de l'expression

des gènes, à la fois qualitativement et quantitativement. Il est ainsi possible de suivre les changements d'expression d'un gène dans le temps, en réponse à des changements de conditions comme l'apparition d'un pathogène ou l'administration d'un médicament, ou dans l'espace par les différences d'expression d'un même gène d'un type cellulaire ou d'un tissu à l'autre.

Un apport important de la transcriptomique est l'identification et la quantification de l'expression des ARN non codants (ARNnc) qui ne sont pas traduits en protéines mais interviennent à différents niveaux des processus cellulaires. Sans entrer dans les détails, on compte dans cette catégorie les ARN ribosomaux et ARN de transfert nécessaires à la traduction en protéines mais aussi les long ARN non codants, les micro ARN, les petits ARN interférant impliqués dans la régulation de l'expression des gènes, ou encore les petits ARN nucléaires impliqués dans la maturation des ARNm dont leur épissage. Les mécanismes d'épissage des ARNm contrôlent la façon dont un gène s'exprime en protéine par le biais de l'épissage alternatif, mécanisme par lequel un même gène peut donner plusieurs transcrits, et potentiellement plusieurs protéines différentes, par l'inclusion ou non de certains exons. Là encore, la transcriptomique apporte des informations additionnelles par rapport à la génomique, en permettant l'identification et la quantification des transcrits alternatifs (Trapnell et al., 2010).

De bien des façons, la transcriptomique en s'intéressant à la façon dont les éléments génétiques s'expriment au sein des cellules, donne donc une vision complémentaire à la génomique sur la façon dont le génotype est traduit en phénotype.

1.2.3 La protéomique

La protéomique est l'omique qui s'intéresse à un des niveaux le plus proche du phénotype en identifiant les protéines présentes dans un échantillon. Elle repose sur des principes techniques très différents des deux premières omiques que nous avons vues plus tôt, car elle s'intéresse à une molécule fondamentalement différente des acides nucléiques. Un des défis majeurs étant que, contrairement à l'ADN, il est impossible d'amplifier les protéines présentes dans un échantillon ; il est donc nécessaire de travailler directement avec la quantité disponible. Néanmoins, les technologies impliquées évoluent aussi et avec elles le potentiel ouvert par l'analyse directe des protéines seules ou en interactions dans la cellule (Altelaar et al., 2013). Les méthodes de protéomique modernes reposent principalement sur la spectrométrie de masse. Les protocoles standards comprennent l'extraction des protéines de l'échantillon, le traitement par une protéase pour les réduire en peptides, puis la fragmentation de l'échantillon par chromatographie liquide. Les peptides de chaque phase sont ensuite fragmentés par spectrométrie de masse, ce qui permet d'en identifier la séquence. L'identification de la protéine correspondante peut ensuite se faire par comparaison à une base de données.

A l'instar de la transcriptomique, la protéomique informe sur les processus biologiques inaccessibles par l'étude du seul génome. Elle permet notamment l'identification précise des sites de modifications post-traductionnelles, la glycosylation, la phosphorylation, la nitrosylation ou l'ubiquitination, dont le rôle est central au bon fonctionnement de la cellule.

En s'intéressant directement aux réels effecteurs de la cellule, la protéomique rend possible la caractérisation des interactions protéines-protéines ayant lieu dans la cellule. Ces réseaux d'interactions sont caractérisés par une organisation en modules hiérarchiques, chacun correspondant à une fonction donnée ou à un phénotype donné, comme une pathologie. Ces particularités sont utilisées pour reconstruire les réseaux fonctionnels permettant de comprendre comment la dysfonction d'un seul gène influence le phénotype final et identifier d'autres gènes potentiellement associés à ce phénotype (Barabási et al., 2011).

1.2.4 Phénomique

La phénomique est la discipline s'intéressant au dernier niveau de l'échelle menant du génotype au phénotype (Figure 1-1) l'ensemble des traits phénotypiques d'un individu. En principe, le phénome décrit l'ensemble des informations microscopiques comme macroscopiques, que l'on peut mesurer pour un individu et qui peuvent varier en fonction du temps et d'une cellule à l'autre, ce qui le rend impossible à caractériser intégralement. Dans les faits, le terme désigne donc l'acquisition de multiples données phénotypiques à l'échelle de l'organisme (Houle et al., 2010).

Les mesures phénotypiques les plus complètes possibles sont essentielles à l'examen des relations génotype-phénotype car elles permettent de tirer profit des données à haut-débit générées par les autres omiques, notamment en réalisant des études d'associations entre les variants observés et les phénotypes, de type *GWAS*. Le besoin de centraliser les informations phénotypiques pour réaliser ces analyses a, par exemple, permis le développement de la *Mouse Phenome Database* (Grubb et al., 2014), qui regroupe de nombreux enregistrements phénotypiques sur les différents souches de souris utilisées en laboratoire.

Une autre approche rentrant dans le cadre de la phénomique est la caractérisation détaillée des différences phénotypiques dues à l'atteinte d'un ou plusieurs gènes notamment dans le cadre des maladies génétiques. Une définition précise des phénotypes associés aux pathologies connues permet d'orienter la recherche de gènes impliqués dans d'autres pathologies avec des syndromes proches. En santé humaine, ces informations phénotypiques sont structurées dans des bases de données comme *Online Mendelian Inheritance in Man* (OMIM) et *Human Phenotype Ontology* (HPO). Ces annotations standardisées permettent de classifier les pathologies en modules en fonction du nombre de phénotypes en commun, qui superposés aux réseaux de gènes issus des autres omiques, peut montrer leur associations à des modules fonctionnels communs (Wu et al., 2009).

1.3 Défis et opportunités des omiques

Les technologies à haut-débit, tout en permettant d'éclairer la complexité des processus biologiques, ont fait passer le vivant dans l'ère des données massives, les fameuses *big data* (Stephens et al., 2015). La gestion et l'exploitation de ces données pour la compréhension des relations génotype-phénotype sont donc soumises aux défis propres aux données massives, à

savoir les problèmes de qualité, d'hétérogénéité et de flux des données. Plutôt que d'explorer la forme que prennent ces problèmes au niveau de chaque omique, je les illustrerai ici avec la génomique dont l'exploitation est au cœur de cette thèse.

1.3.1 La qualité des données génomiques

Un des défis posés par les données massives est le problème de leur qualité et de leur hétérogénéité. Toutes les données accessibles ne sont pas pertinentes, à plus forte raison lorsqu'elles sont issues de protocoles expérimentaux différents. En ce qui concerne les données génomiques, les sources d'erreurs sont nombreuses et les contrôles qualité y sont nécessaires à plusieurs étapes.

Malgré les nombreux développements et améliorations des protocoles et des machines de séquençage ces dernières années, toutes les technologies de séquençage sont sujettes à des erreurs lors de la détection des bases nucléiques. La plupart des protocoles d'analyse des données de séquençage intègrent des programmes permettant de filtrer les lectures sur la base de leur qualité afin de limiter l'impact des erreurs de séquençage sur l'assemblage du génome (Pabinger et al., 2014). Le risque potentiel d'erreur dans la détermination de la séquence finale dépend également de la profondeur de séquençage. Ce paramètre correspond au nombre de fois où chaque base du génome est lue en moyenne, plus la profondeur est élevée, plus on peut être sûr de la bonne attribution des bases dans la séquence génomique.

La profondeur de séquençage est également déterminante pour obtenir une séquence génomique complète, ou tout au moins la plus complète possible. Lors de l'assemblage *de novo* (sans séquence de référence déjà connue), les lectures chevauchantes sont dans un premier temps regroupées les unes aux autres en séquences consensus continues, les *contigs* qui peuvent elles-mêmes être regroupées en *scaffolds*. Ainsi, deux indicateurs de la contiguïté d'une séquence génomique sont le nombre de contigs nécessaires pour couvrir la moitié du génome et la taille en paires de bases du plus petit contig à ajouter pour couvrir 50% du génome (N50).

Les problèmes de qualité au niveau de l'acquisition des données se répercutent dans leur interprétation. J'illustre ce problème par les données d'annotations que l'on retrouve dans les bases de données publiques. La séquence de gènes présents aux extrémités de *contig*, quand ils ne sont pas complètement absents de la base, sont fragmentaires. Les annotations automatiques, même de génomes complets, font régulièrement des erreurs dans les prédictions des bornes des séquences codant pour les protéines, notamment dans l'identification du codon stop ou dans la détermination des bornes exons/introns. Finalement, les annotations fonctionnelles des gènes par des protocoles automatisés sont elles aussi sujettes à erreur, ce qui mène à une proportion considérable de mauvaises annotations dans les bases de données publiques, de 25% à 60% des séquences protéiques selon les familles de protéines considérées (Schnoes et al., 2009). Ces difficultés tendent à augmenter avec le temps : les annotations erronées pouvant être réutilisées pour annoter d'autres séquences, ce qui mène à des propagations d'erreurs d'un génome à l'autre (Gilks et al., 2005).

Les erreurs de prédiction de gènes ou d'annotation fonctionnelle peuvent être résolues par l'intervention d'experts dans la curation de données, permettant d'augmenter de façon considérable la fiabilité de celles-ci, mais le flux des données rend impossible l'application de ces interventions à grande échelle. Cet état de fait est derrière l'asymétrie des deux sections de la base de connaissances UniprotKB (The UniProt Consortium, 2017), dédiée aux séquences et à l'annotation de protéines. Ainsi en novembre 2018, TrEMBL, la section qui contient les entrées de protéines issues de procédures automatiques, en totalise plus de 126 millions, quand Swissprot regroupe 558 590 entrées ayant fait l'objet d'une évaluation manuelle, soit moins d'1% des données totales.

1.3.2 Flux des données en génomique

Le séquençage à haut-débit produit *a minima* des dizaines de petaoctets de données par an, des projections récentes estiment que l'ensemble des données génomiques approchera les zettaoctets (10^{21}) de données d'ici une dizaine d'années, et que les données génomiques humaines représenteraient à elles seules de 2 à 40 exaoctets (10^{18}) d'ici 2025 (Stephens et al., 2015) (Figure 1-2). Au regard de la dernière décennie, le volume de données génomiques double tous les 7 mois et les prévisions pour les années qui viennent estiment, au plus bas, qu'elles continueront à être multipliées par deux tous les 12 mois. Ces chiffres illustrent bien l'arrivée de la biologie dans l'ère des données massives, d'autant plus qu'ils ne prennent pas en compte les données générées par les autres 'omiques', dont la croissance est similaire.

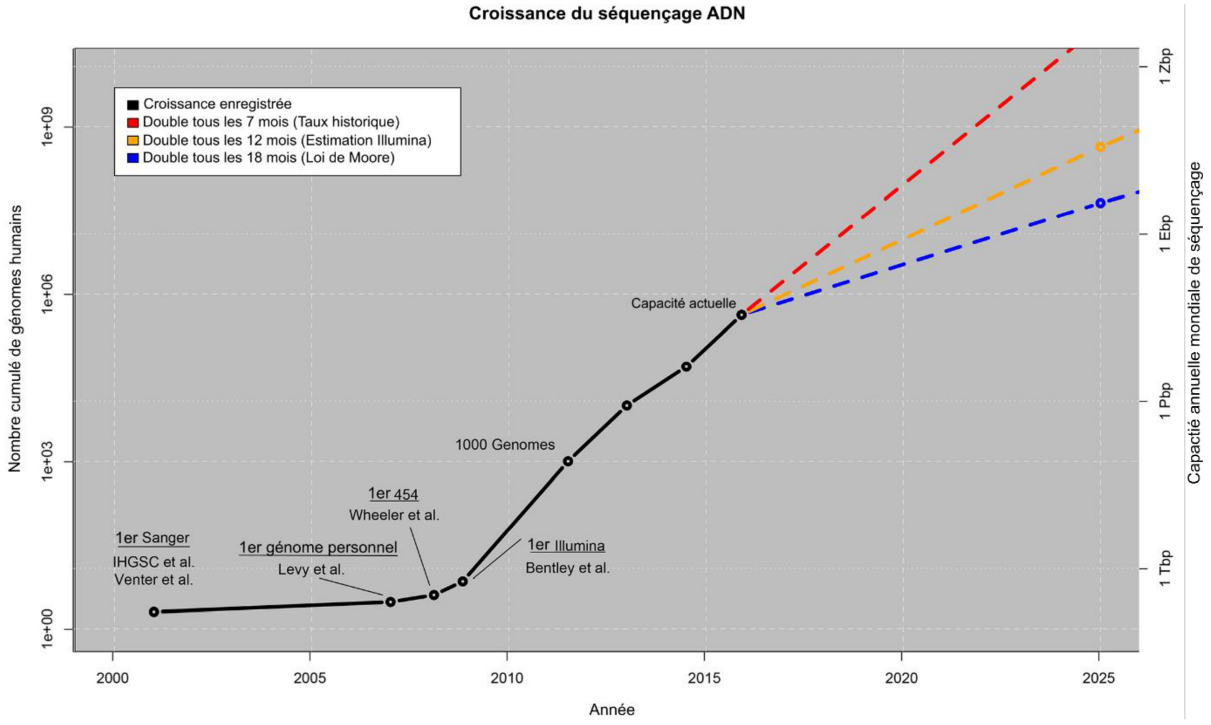


Figure 1-2 **Croissance du séquençage ADN**. Volume des données génomiques humaines en fonction du temps en termes de nombre de génomes (axe gauche) et croissance de la capacité annuelle de séquençage (axe droit). Quelques jalons de l'histoire du séquençage y sont indiqués pour référence, ainsi que les projections de croissance jusqu'en 2025. Adaptée de (Stephens et al., 2015).

L'augmentation exponentielle des données disponibles pose des problèmes nouveaux en biologie dans le sens où ces données doivent être stockées de manière efficace, être aisément mises à disposition malgré leurs tailles et pouvoir être analysées dans des délais raisonnables. Cela nécessite l'accès à du matériel performant, mais surtout, d'être en mesure d'extraire la connaissance des données disponibles et de la représenter synthétiquement pour en faciliter l'analyse et l'accès au plus grand nombre.

1.3.3 Vers l'intégration des données ?

Chacun des types de données 'omiques' permet de décrire une partie des mécanismes en jeu dans le passage du génotype au phénotype, qui ne sont pas accessibles aux autres. Une description plus complète de ces relations passe donc par la considération en parallèle des données issues de l'ensemble des technologies à haut débit. L'accumulation massive de données 'omiques' décrivant de nombreux contextes biologiques dans les bases de données publiques est une opportunité considérable dans cette optique. La perspective de prendre en compte l'ensemble des niveaux de l'échelle entre le génotype et le phénotype pousse au développement de méthodes susceptibles de comparer des données très hétérogènes (tailles, formats, dimensionnalités), plus ou moins bruitées, et plus ou moins informatives.

Une des méthodes traditionnelles d'analyse de données issues de différentes sources (Gligorijević et Pržulj, 2015) est la création de graphes, ou réseaux, où les gènes sont les nœuds. Les arrêtes, c'est-à-dire les relations entre gènes, sont issues de données expérimentales (par exemple, les données d'interactions protéines-protéines, mentionnées plus tôt) ou extraites statistiquement d'un ensemble de données (co-expressions de gènes) (Jansen et al., 2003). Ce genre de représentation permet d'analyser simultanément les différents types d'informations, et d'identifier les gènes fortement interconnectés entre eux : les modules. Ces modules peuvent être mis en correspondance avec des entités phénotypiques, comme des processus biologiques ou moléculaires (Dutkowski et al., 2013), ou encore des pathologies (Vanunu et al., 2010). La construction des graphes peut suivre plusieurs types de méthodologie allant de la simple projection d'associations qui combine les arrêtes issues de différents jeux de données en un graphe unique (Dutkowski et al., 2013) à la construction de réseaux bayésiens, intégrant le pouvoir prédictif de chaque source pour estimer la vraisemblance des interactions (Jansen et al., 2003) à partir d'ensembles de référence.

Les réseaux bayésiens rentrent également dans une autre vaste catégorie de méthodes d'intégration des données biologiques, les méthodes d'apprentissage automatique (*machine learning*). Ces méthodes ont pour avantage de permettre à la fois de mesurer l'apport de chaque type de données pour répondre à une question biologique et de les combiner pour en faire un prédicteur efficace. Un tel classifieur est utilisés par la base de données MitoCarta (Calvo et al., 2016), inventoriant les gènes humains impliqués dans la mitochondrie. Le classifieur intègre des données diverses issues d'analyses génomiques (domaines protéiques, homologies), transcriptomiques (co-expression de gènes dans plusieurs tissus) et protéomiques (MS/MS), et les combine pour prédire des gènes associés à l'organelle (Calvo et al., 2006, 2016). Cette application, comme d'autres du même type (van der Lee et al., 2015), montre le potentiel

supérieur de l'intégration des données pour prédire la fonction de gènes comparé à l'analyse d'un seul type de données.

Dans ce but, les méthodes d'intégration prennent avantage des mines d'or que représente l'ensemble des données omiques enregistrées dans les bases de données publiques. On y retrouve notamment les résultats d'expériences générées dans une grande variété de contextes biologiques, que ce soit chez l'homme ou les organismes modèles. Il s'agit donc de choisir les données pertinentes relatives aux phénotypes d'intérêt pour réaliser ces analyses multi-niveaux.

Au-delà des données relatives à l'homme ou aux organismes modèles, nous disposons d'un autre gisement de données pour apporter une dimension supplémentaire aux analyses multi-omiques : les séquences génomiques annotées d'une grande diversité d'espèces vivantes. Ces données couvrent jusqu'à 15 000 (Mukherjee et al., 2017) organismes, dont la variété phénotypique est plus que considérable. Leur exploitation dans des analyses intégratives implique de pouvoir extraire de cette masse de données, des informations synthétiques et pertinentes et donc, de les considérer dans un cadre commun. L'évolution, qui permet de faire le lien entre toutes les espèces vivantes, peut constituer ce cadre. Comparer les données génomiques de différentes espèces dans ce cadre, permet d'en repérer les points communs et les différences, et de comprendre comme ces différences se retranscrivent au niveau fonctionnel et donc phénotypique.

2 La génomique comparative à l'ère des *big data*

Une des révolutions apportées par la théorie de la sélection naturelle est l'idée que les différentes espèces descendent d'un ancêtre commun et que l'ensemble des espèces résultent de variations successives des descendants de cette espèce ancestrale. Les progrès en biologie moléculaire des 150 dernières années ont complété cette théorie, en montrant que l'ADN est le porteur de l'information génétique commun à tous les êtres vivants et que ce sont des variations de cette molécule qui sont affectées par la sélection naturelle. Le terrain de l'évolution est donc remarquable pour l'étude des relations génotype-phénotype car il offre un cadre permettant de comparer dans le même temps les traits phénotypiques et le contenu génomique de plusieurs êtres vivants.

La génomique comparative est la discipline dédiée à l'étude comparée de la structure et du contenu des génomes de différentes espèces. Elle participe ainsi à la compréhension des dynamiques évolutives du génome et de la fonction de l'ensemble des éléments génomiques. A titre d'exemple sur les apports de la génomique comparative, la première comparaison de génomes de mammifères, l'homme et la souris, permit de déterminer que 5% de l'ensemble du génome est soumis à une pression sélective à l'échelle des mammifères (Mouse Genome Sequencing Consortium et al., 2002), soit une portion bien plus considérable que les régions codantes du génome qui en constituent seulement 1%. L'essor considérable des techniques de séquençage explique que ces études comparatives se soient considérablement multipliées et diversifiées ces dernières années.

Dans cette partie, je définirai les principes fondateurs de la génomique comparative autour du concept d'homologie et illustrerai leurs utilisations pour étendre notre compréhension du vivant. J'aborderai ensuite la façon dont l'explosion des données génomiques a contribué à l'essor de cette discipline et ce faisant, à affiner notre compréhension du génome et des mécanismes de l'évolution. Je terminerai sur son potentiel pour la description des relations génotype-phénotype à travers les techniques du profilage phylogénétique en prenant pour référence le cil eucaryote.

2.1 L'homologie : un concept pour la comparaison des génomes

L'homologie est la notion centrale en génomique comparative qui est essentielle pour toute comparaison entre espèces. Conçue en premier lieu pour comparer des caractères phénotypiques, elle est principalement utilisée aujourd'hui pour définir les relations entre gènes et par extension, entre protéines.

2.1.1 Homologie de caractères

Assez ironiquement connaissant son importance en biologie évolutive aujourd'hui, la première définition de l'homologie est antérieure à la théorie de l'évolution proprement dite. Richard Owen la définissait en ces termes : '*The same organ in different animals under every variety of form and function*' (Owen, 1843). Il s'agit ici d'un critère purement morphologique dont le but était de faciliter la comparaison des espèces pour permettre leur classification. Owen oppose ce concept à l'analogie qui décrit des structures avec les mêmes fonctions, qui ne correspondent pas au même organe. Ses critères pour distinguer les organes homologues sont d'ailleurs leurs 'positions relatives et connections'. A titre d'exemple, les membres antérieurs des Vertébrés tétrapodes ont des fonctions très différentes : la marche chez la plupart d'entre eux, la nage pour les cétacés, le vol pour les oiseaux et les chauves-souris mais l'organisation similaire de leurs os permet de les identifier comme homologues (Figure 2-1).

Avec l'acceptation de la théorie de l'évolution, la définition du terme change pour devenir celle qui est acceptée actuellement : les caractères homologues sont les caractères de deux espèces qui dérivent d'une même structure chez leur ancêtre commun. Sous la nouvelle définition, les analogues sont des structures apparues indépendamment chez les ancêtres de chaque espèce par évolution convergente.

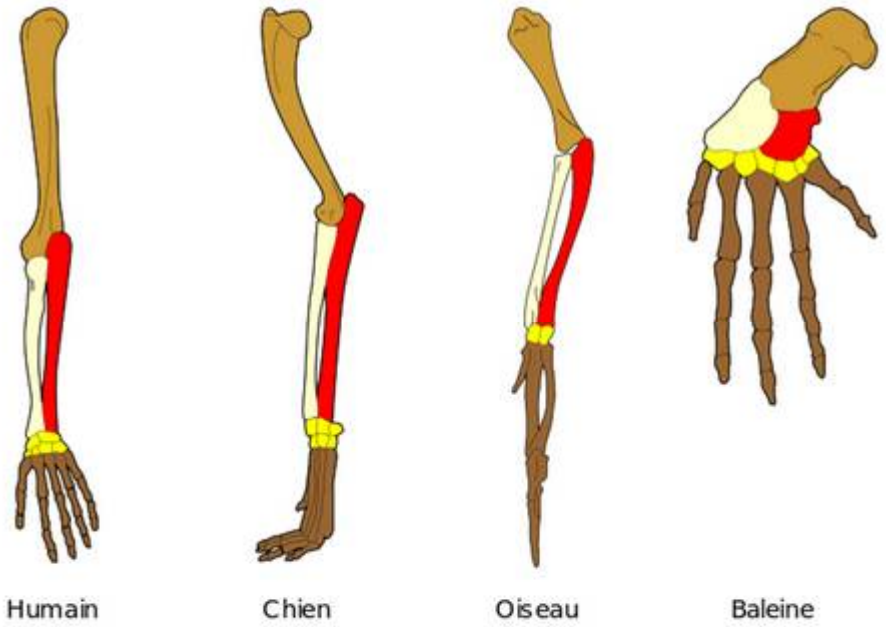


Figure 2-1 **Exemple de relation d'homologie**. Les membres antérieurs des vertébrés, malgré leurs morphologies différentes, descendent d'une même structure ancestrale. Les structures homologues sont désignées par couleur. Source : Wikipédia

Pour établir des relations d'homologie, on identifie les caractères présentant des similarités de structure entre certaines espèces et on cherche ensuite des preuves que ces structures descendent

bien d'un ancêtre commun. Pour cela, on replace l'hypothèse dans le cadre phylogénétique en comparant la distribution de ce caractère avec un arbre représentant les relations de parenté entre ces espèces, si les caractères identifiés comme similaires sont présents seulement chez les descendants d'un ancêtre commun et chez la plupart de ceux-ci, on peut alors les confirmer comme homologues.

Le principe d'homologie de caractères s'est cantonné jusqu'à la moitié du XXème siècle à décrire les relations entre caractères morphologiques observables et a constitué le cadre de référence pour la construction des classifications phylogénétiques du vivant, c'est-à-dire la classification des espèces par rapport à l'ancêtre duquel ils descendent (voir section 2.3.2 La taxonomie du Vivant). Avec l'identification de l'ADN comme support de l'information génétique et l'arrivée des techniques de séquençage, la définition de l'homologie est passée du phénotype au génotype par son application aux objets macromoléculaires.

2.1.2 Homologie en génétique moléculaire

En génétique moléculaire, l'homologie décrit la relation entre deux gènes ou protéines qui descendent d'un même gène ancestral. De la même façon que l'on considère comme homologues des structures avec des positions similaires dans le plan d'organisation, on pose l'hypothèse que les gènes sont homologues en se basant sur la similarité de leurs séquences, mesurée en alignant celles-ci l'une à l'autre. En effet, on suppose que deux gènes descendant d'un ancêtre commun tendent à conserver l'organisation globale et les résidus clés de la séquence ancestrale, donc à avoir une similarité plus élevée que ce que l'on attendrait par chance.

Il est toutefois important de noter qu'homologie et similarité ne sont pas interchangeables : la première est une notion qualitative – on est homologue ou on ne l'est pas -, la seconde une notion quantitative. En outre, deux séquences homologues n'ont pas forcément une forte similarité. Conceptuellement, il est également impossible de démontrer que deux gènes avec une forte similarité sont bien homologues et qu'ils ne dérivent pas d'une convergence évolutive : un processus par lesquelles des séquences sans ascendance commune adopteraient la même séquence. On estime cependant que le cas est suffisamment rare pour que l'hypothèse tienne.

2.1.3 Orthologie et paralogie

On s'intéresse le plus souvent au concept d'homologie pour étudier l'histoire évolutive d'une famille de gènes et obtenir des informations sur la fonction des produits de ces gènes. Pour ce genre d'étude, la notion d'homologue peut malheureusement manquer de précision. En effet, contrairement aux traits phénotypiques, les gènes sont sujets à des duplications qui entrainent l'apparition de deux descendants d'un même gène chez une même espèce. Pour prendre en compte ce paramètre, Fitch introduisit en 1970 (Fitch, 1970) deux catégories d'homologue (Figure 2-2, Figure 2-3):

- **Les orthologues** dérivent d'un même ancêtre commun par un évènement de spéciation ;
- **Les paralogues** dérivent d'un ancêtre commun par un évènement de duplication.

Ces deux catégories n'ont pas intrinsèquement d'implications fonctionnelles ; elles se réfèrent uniquement à l'histoire évolutive des gènes. Cependant, il est communément admis que deux orthologues tendent à conserver une fonction similaire dans les organismes où ils sont présents. Par contraste, les paralogues ont des fonctions redondantes après duplication et la présence d'une copie peut contribuer à relâcher la pression de sélection sur l'autre. Ceux-ci seraient donc plus enclins à avoir des destins différents au cours de l'évolution. Si un résultat fréquent est la perte de fonction d'un des gènes qui dégénère en pseudogène, les copies peuvent développer de nouvelles fonctions (néofonctionalisation) ou une fonction plus spécialisée par rapport au gènes ancestral (spécialisation) (Force et al., 1999). Je reviendrai sur l'impact des duplications dans la section 2.3.3.3.

L'hypothèse voulant que les orthologues conservent plus souvent la fonction ancestrale, alors que les paralogues se diversifient, est appelée la 'conjecture d'orthologie'. Ce principe est communément admis dans la communauté, bien que peu d'études se soient intéressées aux différences fonctionnelles entre orthologues et paralogues (Studer et Robinson-Rechavi, 2009). En 2011, une étude de Nehrt *et al*. (Nehrt et al., 2011) jeta un pavé dans la mare par une étude comparant les équivalences fonctionnelles entre des homologues de l'homme et de la souris, en fonction de leur classe. Ce travail mis notamment en évidence un meilleur pouvoir prédictif des paralogues pour déterminer la fonction des gènes, notamment dans le cas des séquences à forte similarité. Les auteurs mirent alors en garde contre l'utilisation trop commune de la conjecture dans les études comparatives.

Des études postérieures constatèrent des biais dans l'utilisation de *Gene Ontology*, un des points majeurs de comparaison fonctionnelle utilisée par Nehrt pour étudier la correspondance fonctionnelle (Altenhoff et al., 2012; Chen et Zhang, 2012). En corrigeant en partie ces biais, Altenhoff *et al.* ont montré que les orthologues étaient bien, en général, plus similaires en terme de fonction que les paralogues. La différence est cependant faible et inégale selon les fonctions considérées.

La conjecture d'orthologie reste une hypothèse applicable à la plupart des cas et on note que, si le lien direct à la fonction a pu être remis en cause, il existe d'autres critères objectifs plus conservés entre orthologues qu'entre paralogues. Il a été notamment montré que l'organisation des introns (Henricson et al., 2010), la structure tridimensionnelle des protéines (Peterson et al., 2009), l'architecture en domaines (Forslund et al., 2011) tendent à être plus conservées entre orthologues ou encore que ceux-ci s'expriment en général dans les mêmes tissus (Kryuchkova-Mostacci et Robinson-Rechavi, 2016).

2.1.4 Inparalogie et outparalogie

Lorsque l'on compare des gènes homologues dans deux espèces, il est important de déterminer l'ordre des évènements de duplication ayant donné lieu à des paralogues, par rapport à l'évènement de spéciation séparant les deux espèces (Sonnhammer et Koonin, 2002). On appelle **outparalogues,** les paralogues dérivant d'un événement de duplication antérieur à l'évènement de spéciation. Une paire de gènes paralogues était donc présente chez le dernier ancêtre commun des espèces considérées et, en l'absence d'un événement de perte, chacune des espèces filles possède les paralogues issus de ces gènes ancestraux (Figure 2-2). Les paralogues ayant pour origine un événement de duplication postérieur à l'événement de spéciation sont eux appelés **inparalogues**. On retrouve donc les paralogues issus de cette duplication uniquement dans les espèces d'une des branches issues de la spéciation. Les inparalogues sont considérés co-orthologues des autres protéines descendant de l'évènement de spéciation considéré **(**par exemple α2' et α2'' sont co-orthologues de α2 dans l'exemple de la Figure 2-2).

On le voit, les définitions d'inparalogues et d'outparalogues se réfèrent à un événement de spéciation donné, les mêmes séquences paralogues peuvent en conséquence être considérées inparalogues ou outparalogues en fonction de la spéciation à laquelle on se réfère.

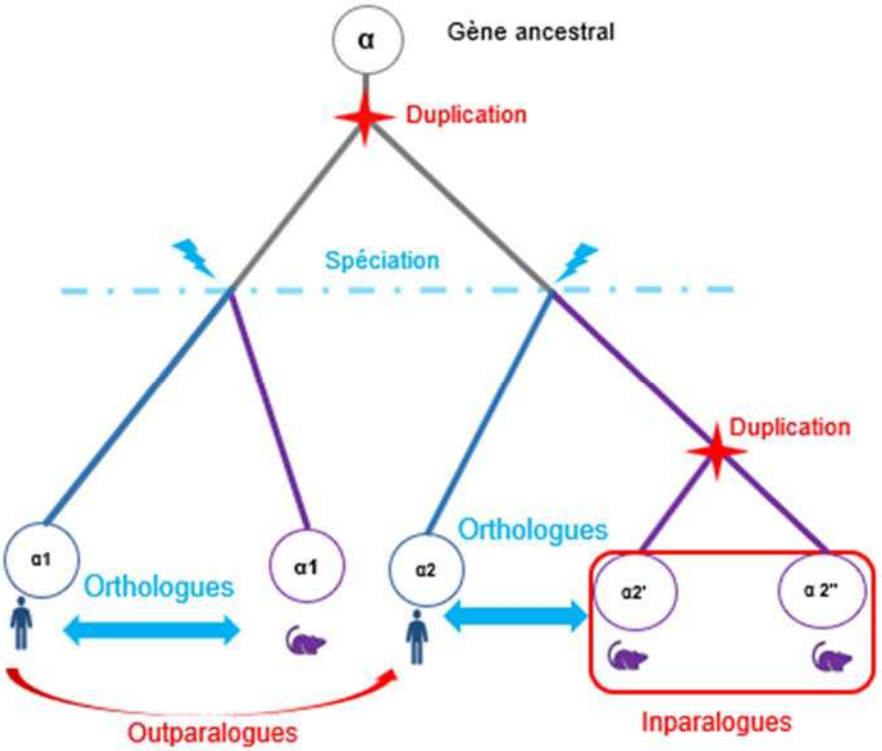


Figure 2-2 **Représentation schématique des relations de paralogie**. Les deux catégories de paralogues (outparalogues et inparalogues) sont définies par rapport à un événement de spéciation, ici entre l'homme et la souris.

2.1.5 Xénologie

Les définitions des orthologues et des paralogues s'inscrivent dans une hypothèse de transmission verticale de l'information génétique : les gènes sont transférés de l'ancêtre à ses descendants. Elles ne couvrent donc pas le transfert horizontal de gène, c'est-à-dire d'une espèce à l'autre (Voir section 2.3.3.2 Les transferts de gènes).
Le terme de xénologue (Gray et Fitch, 1983) permet de décrire les homologues dérivant d'un tel événement (Figure 2-3). La xénologie a majoritairement été observée chez les procaryotes où elle est considérée comme un moteur considérable d'évolution, bien qu'elle soit parfois difficile à distinguer de l'orthologie lorsqu'il s'agit d'un transfert entre espèces proches. Le rôle chez les eucaryotes est moins prédominant, notamment à cause de la séquestration du matériel génétique dans le noyau, mais il apparait que les événements de transfert sont relativement communs avec des procaryotes ou d'autres eucaryotes, par exemple entre hôte et endosymbionte ou par phagotrophie chez les eucaryotes unicellulaires (Soucy et al., 2015). A titre d'exemple, une partie des gènes de la mitochondrie, dont l'origine est probablement phagotrophique, sont passés dans le génome nucléaire des eucaryotes. Avec une définition aussi simple, la relation peut englober des histoires évolutives différentes, plusieurs sous-catégories ont donc été proposées pour les classifier de façon plus nuancée (Darby et al., 2017).

2.1.6 Ohnologues et homéologues

Les ohnologues et homéologues désignent deux catégories d'homologues dus à des événements évolutifs bien particuliers : les évènements de polyploïdies qui se traduisent par l'obtention d'une copie supplémentaire du génome. Ce qui distingue ces deux catégories est la nature exacte de cet événement et la provenance de la copie supplémentaire, de la même espèce ou d'une autre.

Le terme ohnologue décrit une relation de paralogie ayant pour cause un évènement de duplication complète du génome (*Whole Genome Duplication*, WGD) ou événement d'autopolyploïdie (Figure 2-3). Le terme a été proposé par Ken Wolfe (Wolfe, 2000) en hommage à Susumu Ohno (1928-2000), dont les travaux ont permis de mettre en évidence l'importance des WGD dans l'évolution des vertébrés. Etant issus d'un événement de duplication, les ohnologues sont une classe particulière de paralogues. Cette distinction est cependant importante pour étudier les effets des duplications de génomes sur l'ensemble de l'évolution et pour leur propriété propre : les ohnologues, issus d'un même événement, ont divergé entre eux sur une même durée et leur contexte génomique est originellement identique.

Les homéologues sont une classe d'homologues dérivés d'un événement de spéciation, à l'instar des orthologues, mais que l'on retrouve dans la même espèce à la suite d'un événement d'allopolyploïdie (Figure 2-3) : la création d'une espèce hybride par le croisement de deux espèces proches (Glover et al., 2016 pour une définition désambigüisée). Les homéologues ont des caractéristiques uniques parce qu'ils ont été soumis aux mêmes types de pression évolutive que des orthologues durant le temps entre la spéciation et l'hybridation, mais se retrouvent ensuite en état de polyploïdie de la même façon que les ohnologues. De manière générale,

l'étude des homéologues est importante pour mieux comprendre les implications évolutives de l'allopolyploïdie, un évènement relativement commun chez les Plantes et dont on retrouve la trace chez de nombreuses espèces cultivées.

Les différentes catégories d'homologie que nous venons de voir sont des notions nécessaires pour comprendre l'histoire évolutive des gènes et ainsi appréhender les comparaisons entre génomes. La Figure 2-3, ci-dessous, synthétise les caractéristiques de chacune.

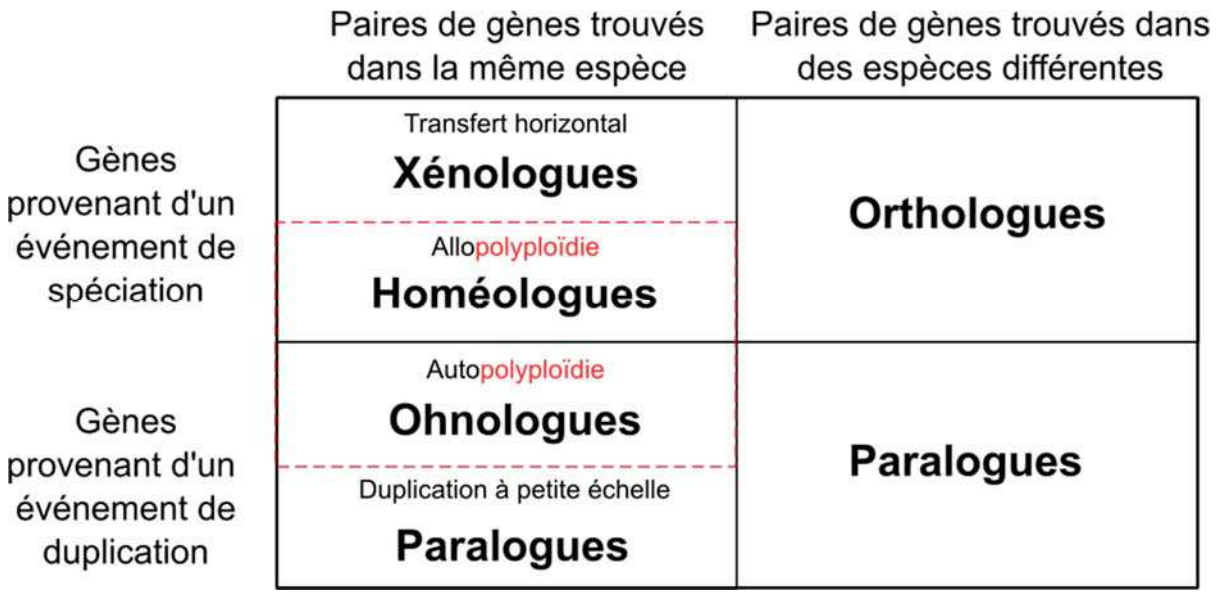


Figure 2-3 **Résumé schématique des catégories d'homologues**. Ces classes sont définies en fonction de l'évènement qui les sépare et le contexte dans lequel on les retrouve. Les homologues issus de polyploïdie forment des catégories à part, indiquées en rouge. Adaptée de (Glover et al., 2016).

2.1.7 Homologies de domaines

Bien que la notion d'homologie soit souvent utilisée pour les gènes ou les protéines, elle peut se référer à d'autres entités biologiques notamment les domaines protéiques. En effet, un gène peut résulter d'un événement de fusion de deux gènes, résultant en une protéine multidomaine. Cette architecture modulaire peut réunir des unités aux histoires évolutives différentes. Dans ce cas, une partie de ce gène fusionné peut être homologue à un gène A et la seconde à un gène B sans qu'A et B ne soient homologues entre eux. Ceci illustre la limite de l'utilisation du concept de l'orthologie entre gènes. Les événements tels que les fusions, les insertions, duplications ou les pertes de domaines qui peuvent altérer la structure du gène, rendent « l'équivalence » entre orthologues plus incertaine : entre deux espèces données, on estime qu'une proportion non négligeable d'orthologues (de 10 à 50%) présentent une architecture en domaine variable (Sonnhammer et al., 2014).

Lors d'études évolutives complètes, cet état de fait nécessite parfois de travailler non plus au niveau du gène protéique, mais au niveau des domaines (Forslund et al., 2018) qui en sont souvent les unités fonctionnelles.

2.2 Exploiter l'homologie pour renseigner sur la fonction

Les relations d'homologie, et à plus forte raison celles d'orthologie, définissent un cadre théorique nécessaire à la comparaison des données génomiques entre plusieurs espèces. Dans cette section, j'aborde la façon dont elles facilitent l'étude du Vivant, à la fois en permettant l'étude inter-espèces des processus biologiques et en éclairant la fonction des séquences moléculaires à la lumière de l'évolution.

2.2.1 Le transfert d'annotations

Comme nous l'avons vu, les gènes orthologues ont tendance à conserver des fonctions équivalentes d'une espèce à l'autre. Cette particularité est abondamment utilisée dans l'annotation des génomes ou pour poser des hypothèses fonctionnelles préalables à des expérimentations.

2.2.1.1 Annotation de génome

Une séquence génomique, même complète, apporte peu d'informations en soi si on ne sait pas y placer les différents éléments génétiques, ni interpréter, même superficiellement, le rôle de ceux-ci dans l'organisme complet. Dans ce chapitre, je n'entrerai pas dans les détails d'annotations de génomes proprement dites (voir (Yandell et Ence, 2012) pour une revue) ; mais je détaillerai les deux étapes où l'orthologie est utilisée.

- **Prédictions de gènes protéiques** : La plupart des stratégies d'annotation modernes utilisent des informations 'externes' pour identifier correctement l'emplacement des gènes, complétant ainsi les prédictions de séquences codantes (*coding DNA sequence*, CDS) basées sur des modèles. Le placement de séquences protéiques homologues sur le génome, par le biais par exemple de TBLASTN (Camacho et al., 2009; Gertz et al., 2006) fait partie de ces méthodes, exploitant les informations disponibles dans d'autres espèces. Ces informations sont notamment importantes pour placer les éléments souvent sujets aux erreurs d'annotations : le codon initiateur et les bornes exons/introns. La comparaison avec des séquences homologues est également utilisée pour évaluer la qualité de l'annotation d'un gène et ainsi faciliter la curation par des experts (Drăgan et al., 2016).

- **Le transfert de fonctions** : Une fois l'ensemble des gènes positionnés sur le génome, la prochaine étape est de leur attribuer une fonction pour, par exemple, faire la lumière sur les particularités biologiques de l'organisme considéré. L'annotation fonctionnelle des gènes s'effectue en identifiant des orthologues dans d'autres espèces pour lesquelles il existe des indications fonctionnelles, issues en grande partie de caractérisations expérimentales, référencées dans des bases de données généralistes telles qu'UniProt (The UniProt Consortium, 2017) ou par le biais de *Gene Ontology* (Gene Ontology Consortium, 2015), puis en transférant ces annotations aux gènes de fonction inconnue.

Plusieurs protocoles automatiques permettent de réaliser cette annotation aujourd'hui (voir (Amar et al., 2014) pour une comparaison de ces protocoles).

Les méthodes de transfert d'annotations s'appuyant sur l'orthologie offrent un gain de temps considérable lorsqu'il s'agit d'annoter les génomes, il est cependant nécessaire d'en noter quelques biais. Premièrement, si la fonction d'un orthologue tend à être équivalente d'une espèce à l'autre, ce n'est pas une loi absolue comme nous l'avons déjà vu. Deuxièmement, si des erreurs existent dans les annotations des homologues auxquels l'on se réfère, cette erreur se répètera avec le transfert d'annotation. Il est donc conseillé de partir le plus possible d'homologues d'espèces proches à celle d'intérêt et de séquences ayant fait l'objet d'une curation par des experts.

2.2.1.2 Transfert de validation expérimentale

De la même façon, on peut utiliser l'information d'homologie entre gènes pour transférer les informations obtenues expérimentalement d'une espèce à une autre, à condition que ces organismes soient suffisamment proches pour permettre ce transfert.

Le consortium *Gene Ontology* (Ashburner et al., 2000; Gene Ontology Consortium, 2015), qui met à disposition une annotation des gènes à travers un vocabulaire standardisé, transfère des informations de fonction en se basant sur ce principe. En résumé, les annotations d'un gène, lorsqu'elles dérivent d'une source expérimentale, sont transférées à ses orthologues dans le même embranchement taxonomique. Les annotations issues de ce genre de transfert portent le code IEA (*Inferred from Electronic Annotation*). *Gene Ontology* intègre également le transfert d'information semi-automatisé prenant en compte les annotations d'homologues de plusieurs espèces et les relations phylogénétiques entre espèces (Gaudet et al., 2011). Ces annotations, plus sûres, portent le code IBA (*Inferred from Biological ancestry*).

2.2.2 Piloter l'utilisation des organismes modèles

La possibilité de transférer les informations obtenues sur les gènes entre deux espèces permet l'étude d'un processus biologique au niveau moléculaire à travers plusieurs espèces, et dans les faits, chez les organismes modèles. Les méthodes de diagnostic basées sur le séquençage d'exomes et génomes, répandues aujourd'hui permettent d'isoler des gènes potentiellement responsables de pathologies. Etudier la fonction des gènes expérimentalement chez l'homme étant impossible pour des raisons éthiques, *a fortiori* à l'échelle de l'organisme, elles nécessitent de passer par les organismes modèles. Travailler sur l'orthologue de ces gènes permet de tester et vérifier des hypothèses fonctionnelles et d'une façon générale de mieux comprendre les origines des pathologies (Spradling et al., 2006). En dehors des considérations pratiques, notamment le coût, plusieurs paramètres peuvent rentrer en compte pour le choix de l'organisme modèle :

- L'existence d'un orthologue 1-à-1, la présence d'inparalogues rendant moins fiable le transfert des conclusions.

- La similarité entre homologues. Pour des séquences conservées, il est possible de reproduire les variations considérées comme pathogènes et tester leur effet.
- Une homologie entre l'ensemble des gènes du processus étudié. De cette façon, les mécanismes d'interactions mis en évidences ont une équivalence chez l'homme.

2.2.3 Les interologues : transférer les informations aux niveaux des systèmes biologiques

Lorsque l'on étudie un processus biologique, on s'intéresse non seulement à la fonction d'un gène donné, mais également à leurs interactions avec d'autres gènes, en d'autres termes au système qui sous-tend ce processus. Comme nous l'avons vu au chapitre précédant, la représentation de ces systèmes sous forme de réseau est une forme d'intégration des données omiques. Ces informations systémiques peuvent être transférées d'une espèce à l'autre par la notion d'interologues.

Le terme d'interologues, proposé par (Walhout et al., 2000), décrit la relation entre deux couples de protéines A*h* et B*h* d'une première espèce et A*m* et B*m* d'une seconde espèce. On considère que A*h*-B*h* et Am-Bm sont interologues si : (a) A*h* est orthologues de A*m,* (b) B*h* est orthologue de B*m*, (c) les membres de chaque couple interagissent entre eux (Figure 2-4).

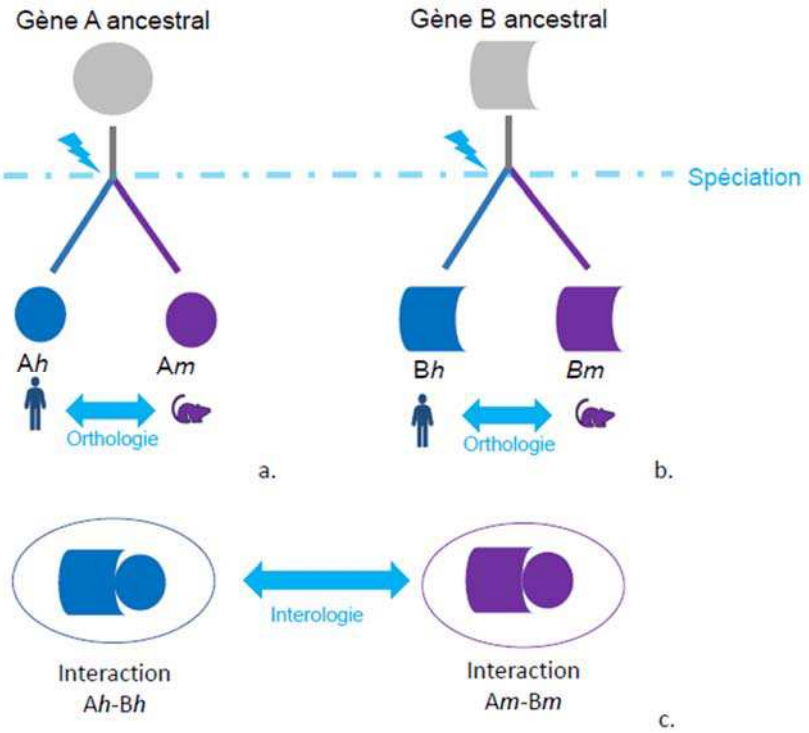


Figure 2-4 **Représentation schématique d'une relation d'interologie.** a. La protéine Ah humaine et la protéine A*m* murine sont orthologues. b. Les protéines B*h* et B*m* sont orthologues. c. Ah interagit avec A*m*, B*h* avec B*m*. A*h*-B*h* et A*m*-B*m* sont interologues.

De la même façon qu'on utilise la conjecture d'orthologie pour émettre des hypothèses sur la fonction de protéines à partir de la fonction des orthologues, on utilise le concept d'interologue pour inférer les interactions entre protéines d'une espèce donnée en se basant sur les interactions

connues entre leurs orthologues chez une autre espèce. Cela permet notamment de tirer parti des grands volumes de données obtenues par les expériences des analyses d'interactions à haut débit par protéomique. Les différentes méthodologies pour inférer des interactions obéissent à un principe simple : on identifie les orthologues de protéines en interaction et l'on transfert l'information au couple d'orthologues, voire aux co-orthologues le cas échéant, en intégrant parfois des paramètres comme la similarité de séquences. (Yu et al., 2004; Garcia-Garcia et al., 2012).

Comme on pourrait l'attendre de méthodes visant à transférer de nombreuses informations d'expériences à haut-débit, ces approches ont tendance à apporter des faux positifs. Pour cette raison, les inférences d'interologie sont utilisées de façon combinée (Huang et al., 2007), en les croisant avec des inférences issues de plusieurs espèces ou avec d'autres données (Rhodes et al., 2005; Wan et al., 2015). Une des meilleures illustrations de cela est probablement leur exploitation dans la base de données d'interaction protéine-protéine STRING (Szklarczyk et al., 2017) , où elles sont l'un des nombreux indicateurs d'interaction possibles.

2.2.4 Etudes des familles de protéines

Une autre façon de tirer des informations fonctionnelles de l'homologie des gènes ou des protéines est d'analyser la façon dont les pressions de sélection ont affecté leur séquence. Il s'agit de comparer les séquences macromoléculaires homologues issues de différentes espèces, ce qui permet ensuite d'inférer des informations fonctionnelles notamment sur les éléments de séquences importants fonctionnellement. Ces protocoles de comparaisons de séquences homologues reposent principalement sur les alignements multiples de séquences, un outil essentiel en génomique comparative.

2.2.4.1 Alignements de séquences

Après séparation, les séquences homologues dérivent de la séquence ancestrale par des événements de mutations ponctuelles (*i.e.*, changement d'un acide nucléique au niveau de la séquence génique, pouvant induire des changements d'acides aminés dans la séquence protéique), des insertions ou délétions de séquences (*indel*). Le but des alignements de séquences est d'aligner les résidus des homologues dérivant de la même position dans la séquence ancestrale. Cet objectif n'est pas trivial à atteindre pour la simple raison que cette séquence ancestrale est inaccessible et dans les faits, les alignements de séquences deux-à-deux se font par l'optimisation mathématique d'un score d'alignement. Brièvement, cela revient à maximiser le nombre de résidus identiques alignés et à insérer des *gaps* aux emplacements supposés des *indels*. Les méthodes d'alignements de séquences deux-à-deux permettent soit l'alignement de l'ensemble des résidus des séquences, on parle alors d'alignement global (Needleman et Wunsch, 1970), soit l'alignement du segment le plus conservé, on parle alors d'alignement local (Smith et Waterman, 1981). Les alignements locaux sont notamment utiles lorsque les séquences comparées ne sont pas co-linéaires ou si l'on s'intéresse à l'homologie d'une partie de la séquence seulement, par exemple un domaine.

Si l'alignement de séquences deux-à-deux peut être résolu grâce à des méthodes déterministes, les alignements impliquant plus de deux séquences aussi appelés alignements multiples de séquences, sont un problème NP-complet (Wang et Jiang, 1994), impossible à résoudre pour un grand nombre de séquences. Les programmes d'alignements multiples utilisent donc des méthodes heuristiques pour produire les alignements. Les méthodes d'alignements progressives en sont un exemple, les séquences sont d'abord soumises à une étape de *clustering* où elles sont ordonnées selon leur proximité. Cet arbre détermine dans quel ordre les séquences sont ajoutées progressivement à l'alignement, avec à chaque étape des alignements deux-à-deux entre séquences ou avec un profil représentatif des séquences déjà présentes dans l'alignement. Ces méthodes sont intégrées dans les programmes ClustalW (Thompson et al., 1994) et T-Coffee (Notredame et al., 2000), et servent aussi de base à des méthodes itératives telles que MAFFT (Katoh et al., 2002), Muscle (Edgar, 2004) et ClustalOmega (Sievers et al., 2011).
Les alignements multiples de séquences permettent de visualiser les séquences homologues dans un contexte ayant un sens biologique et d'en tirer des informations utiles à la compréhension de leur fonction :

- Comparer l'organisation en domaine des différents homologues,
- Identifier les différences entre paralogues,
- Identifier les résidus les plus conservés, potentiellement importants pour la fonction de la protéine,
- Inférer les structures secondaires ou tertiaires (tridimensionnelles) des protéines,
- Reconstituer l'histoire évolutive des gènes.

2.2.4.2 Phylogénétique moléculaire

La phylogénétique moléculaire est la discipline qui s'applique à reconstruire l'histoire évolutive des protéines à partir des alignements multiples de séquences. De la même manière que la phylogénie classique exploite des comparaisons de caractères morphologiques pour classifier les êtres vivants en fonction de leur parenté vis-à-vis d'un ancêtre commun, la phylogénie moléculaire se base sur les comparaisons des séquences alignées pour retrouver leur relation par rapport à la séquence ancestrale. Le résultat d'une telle comparaison prend la forme d'un arbre dont les nœuds correspondent aux évènements de spéciations et de duplications.

Les méthodes de phylogénie peuvent être divisées globalement en deux catégories (Yang et Rannala, 2012) :

- **Les méthodes basées sur les distances** qui évaluent une matrice de distances entre séquences et l'utilisent pour générer un arbre par des algorithmes de *clustering*. La méthode la plus couramment utilisée pour cette tâche est l'algorithme de *Neighbour Joining*.
- **Les méthodes basées sur les caractères :** Ces approches considèrent l'ensemble des séquences pour chaque position une à une, la nature du résidu à cette position étant considérés comme un caractère. A partir de l'ensemble des positions, on calcule un score pour chaque arbre phylogénétique possible et l'on sélectionne celui avec le

meilleur score. Le maximum de parcimonie, le maximum de vraisemblance et les méthodes d'inférences bayésiennes sont toutes des méthodes de cette catégorie.

Les méthodes de phylogénie moléculaire peuvent être utilisées pour reconstruire l'arbre phylogénétique des espèces en utilisant les séquences de gènes ou protéines variant relativement peu au cours de l'évolution et présents dans la plupart des espèces, principalement les gènes « domestiques » et les gènes des ARN ribosomiques. Lorsqu'elles sont appliquées à l'étude d'une famille de gènes, elles permettent de reconstruire exactement l'ordre d'occurrence des événements évolutifs impliqués dans cette famille : en comparant à un arbre des espèces fiables, on peut identifier à chaque nœud de l'arbre les événements de spéciations, de duplications et de transferts horizontaux de gènes.
A partir de l'étude phylogénétique d'une famille de protéines, il devient également possible de reconstituer une séquence macromoléculaire au plus proche de la séquence ancestrale en replaçant les évènements de substitutions et d'*indels* sur l'arbre phylogénétique (Thornton, 2004).

Comme nous l'avons vu dans cette section, l'étude des relations d'homologie, et plus particulièrement d'orthologie, entre gènes apporte des informations utiles pour l'étude de leurs fonctions, individuellement ou dans leurs relations les uns aux autres. La révolution génomique des deux dernières décennies a induit un changement d'échelle, en permettant d'étudier l'homologie non plus au niveau des gènes mais à celui du génome et à travers une diversité toujours croissante d'organismes vivants.

2.3 La génomique comparative pour connaître le vivant

Avec le séquençage de plus en plus d'espèces, la génomique comparative a contribué de façon importante à comprendre le Vivant : la structure des génomes, les nombreux mécanismes de l'évolution et donne même l'opportunité de reconstituer l'histoire de la vie depuis son apparition, il y a environ 4 milliards d'années. Dans cette section, je brosse un portrait de la quantité des données accessibles à la génomique comparative et survole les avancées permises par la génomique comparative, pour notre compréhension du Vivant dans son ensemble et pour faciliter l'interprétation des données génomiques.

2.3.1 L'apport des données massives

Comme nous l'avons vu dans le chapitre précédant, l'emergence des techniques de séquençage a permis de démocratiser l'accès aux génomes complets et ce pour une quantité vertigineuse d'espèces. Les approches métagénomiques ont accéléré cette tendance, en donnant accès aux séquences génomiques d'espèces inconnues, difficiles à isoler ou non cultivables en laboratoire.

La croissance de la base de données *Genomes OnLine Database* (GOLD) (Mukherjee et al., 2017), qui recense les projets de séquençage de génomes et métagénomes, est une illustration flagrante de cette accélération dans l'accumulation des données. La tendance n'a cessé de s'accélérer ces dix dernières années (Figure 2-5) pour arriver, rien que pour l'année 2017, à

l'ajout de plus de 23 000 'brouillons' permanents et presque 4500 génomes complets. Au total, on relève aujourd'hui ~15 000 organismes pour lesquels il existe un génome complet d'après cette base de données. Le séquençage complet d'un génome demandant toujours une quantité de ressources importante, relativement peu de projets vont jusqu'à l'assemblage du génome dans son intégralité, mais donnent lieu à des 'ébauches permanentes', qui concernent jusqu'à 95 000 espèces. Au niveau de la diversité, une grande partie de ces projets correspond à des bactéries (85 000 ébauches permanentes, ~10 000 génomes complets), suivis par les eucaryotes (4 500 ébauches permanentes, 424 génomes complets), les archées (800 brouillons permanents, 300 génomes complèts) et les virus (3 000 brouillons permanents, 5 000 génomes complets).

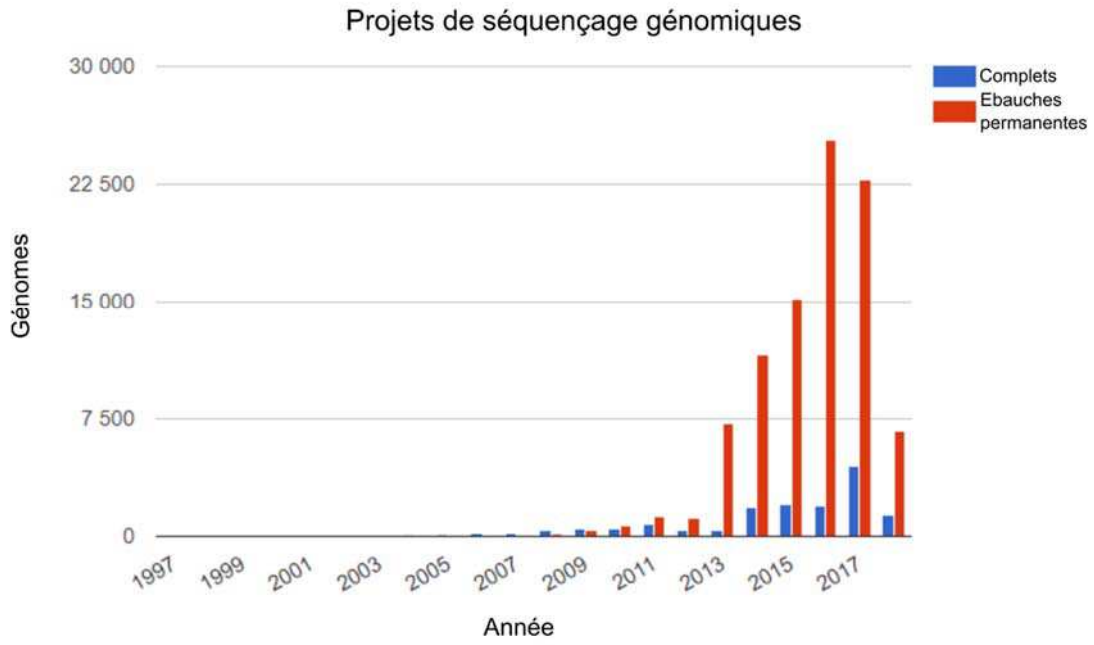


Figure 2-5 **Projets de séquençage terminés dans GOLD**. L'acquisition de données génomiques s'est considérablement accélérée ces dix dernières années. Tiré de https://gold.jgi.doe.gov.

Ce nombre ne correspond qu'à une fraction minime des 2,3 millions d'espèces dont on connait l'existence, mais on peut s'attendre à ce que la tendance à l'accélération continue dans les années qui viennent, avec l'émergence de projets ambitieux visant à séquencer le génome de l'ensemble des espèces connues (Lewin et al., 2018). Pour autant, la diversité disponible est d'ores et déjà une opportunité unique pour étudier le Vivant à l'aune de l'évolution.

2.3.2 La taxonomie du vivant

La théorie de l'évolution a révolutionné notre façon de penser le vivant, en théorisant l'existence d'un ancêtre commun à toutes les espèces vivantes. C'est ce changement de paradigme qui a permis l'émergence de la systématique phylogénétique, visant à regrouper les organismes vivants par les relations de parenté, devenue l'outil de classification des espèces vivantes majoritaire depuis son introduction par Hennig en 1950 (Hennig, 1950). Cette classification repose sur la comparaison de caractères pour déterminer les relations les plus

probables. Pour en servir de base, le génome est un outil de choix car il s'agit d'un des seuls éléments communs à l'ensemble des êtres vivants cellulaires que l'on peut comparer.

Comme évoqué dans la section précédente, les phylogénies d'espèces ont dérivé dans un premier temps des phylogénies de gènes conservés comme les ARN ribosomaux ou certains gènes 'domestiques', mitochondriaux comme nucléaires. L'avènement de ces méthodes, concomitantes à celle du séquençage, ont permis de changer considérablement notre vision du Vivant, jusque-là basée sur des comparaisons phénotypiques uniquement. Leur émergence révolutionna notre compréhension des relations de parenté dans de nombreux groupes : les Amphibiens (Frost et al., 2006), les Squamates (Townsend et al., 2004), les Amniotes (Hedges et Poling, 1999) et même les Métazoaires au sens large. Pour prendre ces derniers pour exemple, les phylogénies moléculaires se basant sur les ARNr 18S enterrèrent les conceptions traditionnelles d'un clade regroupant les animaux articulés, notamment les arthropodes et les annélides (Kim et al., 1996) et retrouvèrent une relation de proximité inattendue entre arthropodes et nématodes, formant avec les tardigrades et les onychophores le clade des Ecdysozoaires, animaux capables de muer (Aguinaldo et al., 1997).

Les phylogénies moléculaires sont surtout à l'origine d'un changement de paradigme important dans notre représentation du vivant, d'une hypothèse voulant l'existence de deux grandes Domaines du vivant, les procaryotes et les eucaryotes, elles nous ont fait passer à une conception à trois Domaines : les Bactéries, les Archées et les Eucaryotes (Woese et al., 1990; Woese et Fox, 1977), ces deux derniers constituant un clade monophylétique, c'est-à-dire regroupant tous les descendants d'un même ancêtre commun. Les phylogénies moléculaires ont également apporté des éléments importants sur l'origine de la mitochondrie, acquise par endosymbiose (Sagan, 1967), en l'identifiant comme une proteobacteria α, proche du genre *Rickettsia* (Gray et al., 1999). A ce titre, ces études comparatives sont un outil essentiel pour nous éclairer sur les évènements ayant eu lieu il y a plusieurs millions, voire milliards, d'années.

A l'ère du séquençage à haut débit, les phylogénies peuvent s'appuyer sur des comparaisons incluant des comparaisons à l'échelle du génome (Crawford et al., 2012) et un nombre toujours plus élevé d'espèces (Alexander Pyron et Wiens, 2011; Bininda-Emonds et al., 2007; Pyron et al., 2013) pour parfaire notre compréhension de l'arbre de la Vie. Il apparait toutefois que ces comparaisons, en théorie plus fiables, ne remettent pas en cause la plupart des hypothèses définies sur une quantité bien moindre de séquences (Pyron, 2015). L'état actuel de nos connaissances en taxonomie est accessible dans des bases de données publiques, telle que la base Taxonomy du NCBI (Federhen, 2012; Sayers et al., 2009) et initiative Open Tree of Life (Hinchliff et al., 2015; Rees et Cranston, 2017). Le contenu de ces bases de données, comme notre conception de la taxonomie, ne sont toutefois en aucun cas gravés dans le marbre, et peuvent être remises en causes, par exemple, suite à la découverte de nouvelles espèces.

Encore récemment, la découverte d'un nouvel embranchement d'Archées, nommés ASGARD en référence au panthéon nordique (Spang et al., 2015; Zaremba-Niedzwiedzka et al., 2017) , a saisi l'intérêt de la communauté. On retrouve chez ces Archées des homologues de protéines jusque-là considérées comme spécifiques des Eucaryotes, notamment des protéines de

cytosquelettes (Actin) et de trafic intracellulaire. Ces découvertes ont attisé le débat autour de la question contentieuse du placement des Eucaryotes dans l'arbre du Vivant, en tant que clade frère des Archées ou inclus en son sein.

Ainsi notre représentation de l'arbre du Vivant (la Figure 2-6 illustre une vision récente de celui-ci), éclairée par la diversité génomique à notre disposition est amenée à évoluer et à se raffiner par la suite, elle est toutefois suffisamment stable pour servir d'appui à des analyses comparatives qui sont l'objet du reste de la section.

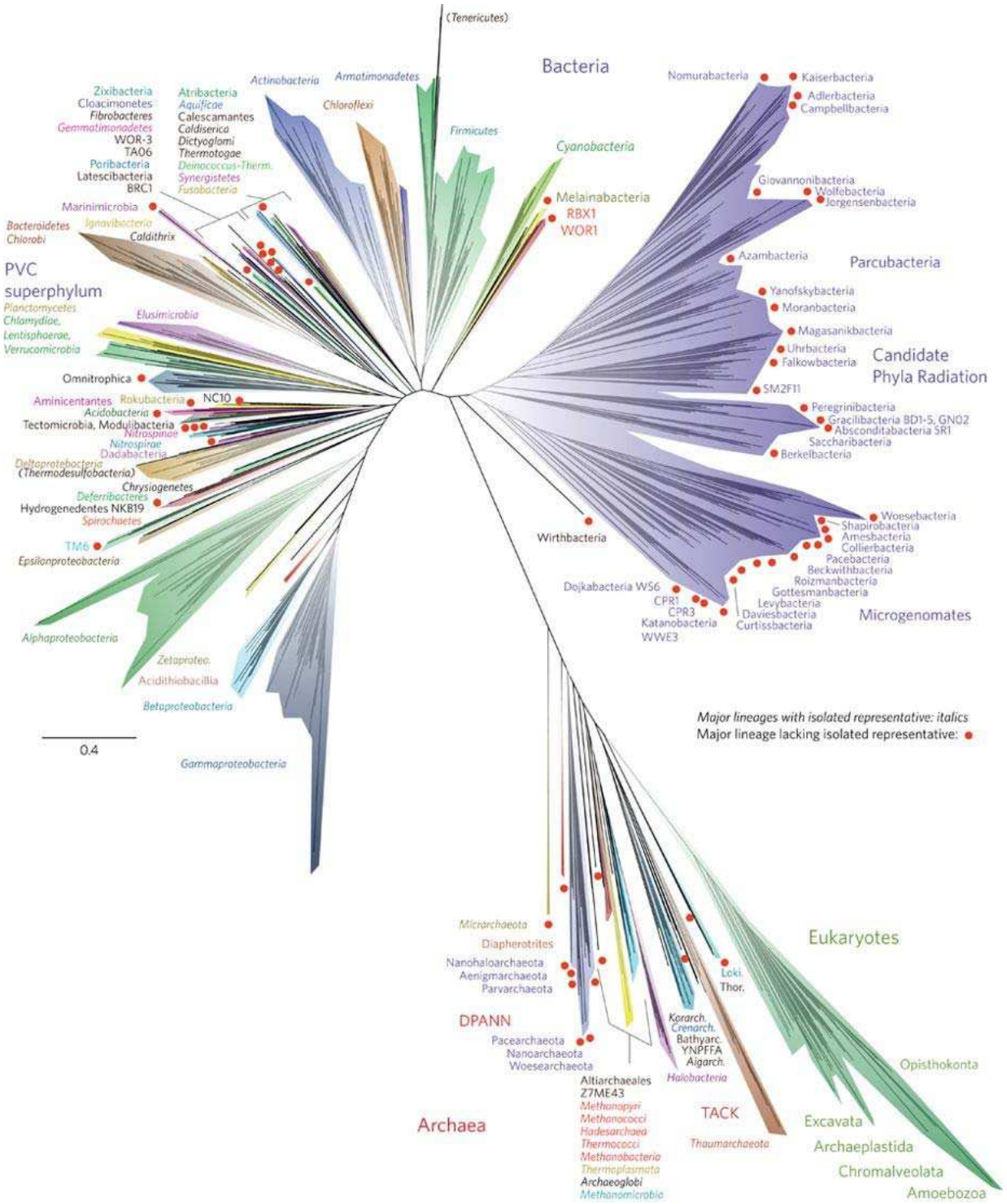


Figure 2-6 **Le diversité de l'arbre du Vivant**. Arbre phylogénétique récent reconstruit à partir des séquences de protéines ribosomales comprenant 3 083 génomes. Les embranchements majeurs sont représentés et colorés arbitrairement. Tiré de (Hug et al., 2016).

2.3.3 La plasticité génomique

Les génomes des organismes vivants sont sujets à des variations importantes, il suffit pour s'en convaincre de comparer le génome circulaire des bactéries et archées aux 23 paires de chromosomes du génome humain. Même entre organismes relativement proches, on trouve des différences dans la taille du génome, le nombre de gènes, et l'organisation des différents éléments génétiques. La génomique comparative a permis d'appréhender cette plasticité et d'envisager les mécanismes biologiques qui en sont responsables. Dans cette section, je décris quelques-uns des événements majeurs qui structurent l'évolution du génome en soulignant ce que les comparaisons à l'échelle du génome ont permis d'en apprendre. J'insiste notamment sur les concepts importants pour comprendre le contexte de mes travaux.

2.3.3.1 Ordre des gènes et réarrangements chromosomiques

Les analyses comparatives à l'échelle du génome permettent tout d'abord d'étudier la conservation de l'ordre des gènes, ou synténie, entre espèces. Cela est réalisé traditionnellement en identifiant les relations d'orthologie entre les gènes de chacun des génomes et en comparant ensuite leur position sur celui-ci.

Les études de ce genre, menées chez les procaryotes ont montré que la synténie est généralement retrouvée entre espèces proches (échelle du genre), mais se perd rapidement avec la distance évolutive (Huynen et Bork, 1998). L'ordre des gènes est moins conservé que la séquence des orthologues et ceci chez les archées (Lecompte et al., 2001) comme chez les bactéries. Certains moteurs de ces différences sont les mécanismes de réarrangements (Hughes, 2000) au niveau du génome que l'on peut classer en deux grandes catégories :

- La translocation : un segment du génome est transféré d'une position chromosomique à l'autre,
- L'inversion : un segment du génome est replacé au même endroit, mais dans une position inverse.

Ces réarrangements, déjà fréquents entre espèces proches, peuvent expliquer les variations importantes dans l'ordre des gènes entre espèces éloignées (Hughes, 2000). Si la synténie est très peu conservée chez les procaryotes, certains éléments génomiques échappent à cette règle. Les opérons, groupes de gènes consécutifs sur la séquence génomique dont l'expression est concomitante et qui codent généralement pour des protéines impliquées dans la même voie, en sont un bon exemple. On retrouve ces gènes dans le même ordre chez des espèces relativement éloignées. D'une manière générale, la conservation de l'ordre des gènes peut donc être indicatrice de correspondance fonctionnelle (Tamames, 2001)

Les réarrangements sont également fréquents chez les Eucaryotes, bien qu'ils soient plus difficiles à étudier à cause de la taille des génomes. Les deux principaux types de réarrangements sont les translocations (échanges) de séquences intra et interchromosomiques, et les inversions de séquences au niveau du même chromosome (Eichler et Sankoff, 2003). La fréquence de réarrangements diffère selon les espèces ou les chromosomes considérés, mais

l'ordre des gènes, bien qu'encore moins conservés que les séquences d'orthologue, le reste plus que chez les procaryotes (Koonin, 2009; Zdobnov et al., 2005). On ne retrouve pas d'opérons chez les eucaryotes, mais il a été montré que les disruptions de synténie ne sont pas complètement aléatoires et que certains gènes fonctionnellement proches tendent à être colocalisés (Hurst et al., 2004). A titre d'exemple, une étude réalisée sur les génomes d'espèces modèles a montré que les gènes impliqués dans les mêmes voies métaboliques ont tendance à être plus colocalisés qu'attendu par chance (Lee et Sonnhammer, 2003). Cette propension à la proximité entre gènes fonctionnellement proches, que ce soit chez les Procaryotes ou les Eucaryotes, peut être utilisée pour des inférences fonctionnelles, nous reviendrons sur ce point dans la section 2.3.5.

2.3.3.2 Les transferts de gènes

Les comparaisons à l'échelle du génome ont également mis en lumière l'ampleur, dans l'évolution, des transferts horizontaux de gènes (*horizontal gene transfert,* HGT). Ceux-ci sont aujourd'hui considérés comme le moteur majeur de l'évolution chez les Procaryotes, Bactéries comme Archées, et tiennent également une place importante chez les Eucaryotes (Soucy et al., 2015). A partir de la séquence génomique, on peut les détecter selon deux grandes méthodes.

La première se base sur des biais de composition des séquences. En effet, la composition en nucléotides (A+T/G+C) varie entre espèces (Sueoka, 1988), comme la fréquence de dinucléotides et l'utilisation préférentielle des codons synonymes (codant pour un même acide aminé). En s'intéressant aux biais anormaux de composition au niveau du génome, on peut donc potentiellement identifier les transferts horizontaux (Lawrence et Ochman, 2002), à condition qu'une signature génomique différente existe entre les deux génomes (Koski et al., 2001).

Une autre méthode, plus utilisée, est celle dite du conflit phylogénétique (Soucy et al., 2015) : il s'agit de générer l'arbre phylogénétique du ou des gènes étudiés et de le comparer à un arbre des espèces. Une différence notoire entre ces deux arbres, rendant compte d'une proximité inattendue entre gènes d'espèces éloignés permet de supposer le transfert horizontal. Cette méthode est cependant moins efficace pour détecter les événements entre espèces proches (la différence entre les arbres étant minime ou nulle), ce qui peut être résolu par un échantillonnage plus fin, incluant plus de représentants d'espèces de forte proximité taxonomique.

Inclure des espèces de forte proximité taxonomique est d'autant plus important chez les procaryotes qu'une majorité des événements de transferts horizontaux a lieu entre espèces proches (Williams et al., 2012). Ainsi, par comparaison d'espèces ayant divergé récemment, Treangen et Rocha (Treangen et Rocha, 2011) ont montré que des gènes initialement attribués à des duplications sont attribuables à des événements de transfert, illustrant ainsi l'importance des transferts dans l'évolution des génomes bactériens. Les transferts de gènes y sont tellement fréquents que la notion de génome décrit parfois mal la réalité d'une espèce, ainsi seulement 6% des familles de gènes retrouvés chez les différentes souches d'*Escherichia coli* sont présentes chez l'ensemble de ses représentants connus (Lukjancenko et al., 2010). On parle alors de pan-génome pour décrire les gènes retrouvés chez tous les représentants d'une espèce.

Les transferts horizontaux de gènes étant en comparaison moins nombreux chez les Eucaryotes, principalement à cause de l'enveloppe nucléaire et de la distinction entre lignées germinales et somatiques chez les pluricellulaires, leur détection pose moins de difficultés que chez les Procaryotes d'autant plus qu'une proportion importante provient d'espèces éloignées, souvent de bactéries. Un exemple emblématique des HGT chez les Eucaryotes est le transfert des gènes de l'endosymbiote vers l'hôte, illustrés par le transfert des gènes mitochondriaux (et chloroplastiques chez les plantes), dans le génome nucléaire (Timmis et al., 2004). De plus en plus d'exemples de ces événements chez les Eucaryotes sont découverts avec les nouvelles séquences, et on note même que des transferts horizontaux ont été observés entre cellules somatiques humaines et bactéries du système digestif (Riley et al., 2013).

L'impact des transferts horizontaux est tel que l'on peut parler de réseau du Vivant plutôt que d'arbre du Vivant, tant ses branches sont interconnectées (Williams et al., 2011). Ces événements sont donc évidemment à prendre en compte lorsque l'on s'intéresse à l'histoire évolutive des gènes.

2.3.3.3 Les duplications

La duplication de gènes est un autre des moteurs essentiels de l'évolution, permettant l'apparition de nouvelle fonction. Cette idée, avancée pour la première fois par Susumu Ohno indiquait que les duplications permettaient de réduire la pression évolutive sur un des paralogues, dont le destin était soit de disparaître (pseudogénisation) ou de développer une nouvelle fonction (néofonctionnalisation). L'arrivée des séquences de génomes complets a depuis permis d'obtenir une vision plus complète de ces phénomènes.

Ainsi les comparaisons de génomes eucaryotes ont permis d'estimer la fréquence des duplications de gènes à environ 0.01 par gène par million d'années, un taux considérable à l'échelle de l'évolution (Lynch et Conery, 2000). Les duplications n'affectent pas tous les gènes de la même façon, et la comparaison de nombreux génomes a permis d'observer que les cas de duplication de gènes domestiques, dont la fonction est centrale à la survie de l'organisme, sont relativement rares quand on retrouve plus souvent des duplications des gènes dits environnementaux (Kondrashov et al., 2002), ceux impliqués dans la réaction aux signaux externes, les relations hôtes pathogènes ou la réponse au stress, entre autres (Ponting, 2008). La duplication contribue ainsi à l'expansion importante de certaines familles de gènes, à titre d'exemple, les récepteurs olfactifs des Mammifères, varient fortement en nombre selon les espèces considérées ; pour 780 gènes estimés chez l'ancêtre commun, on en retrouve un millier de copies chez les rongeurs et jusqu'à presque 2000 chez l'éléphant (Niimura et al., 2014).

En plus des cas concernant les segments de génome, une autre source de duplication de gènes est la duplication du génome complet, ou polyploïdie, mentionnée précédemment avec les notions d'ohnologie et d'homéologie. Ces duplications peuvent survenir lors du croisement fertile de deux espèces différentes (allopolyploïdie) ou par l'obtention de deux copies du génome de la même espèce (autopolyploïdie) (Van de Peer et al., 2009).

Des traces d'évènements de ce type ont été retrouvées dans le génome de divers Eucaryotes : l'alvéolés *Paramecium tetraurelia* (Aury et al., 2006), les levures (Wolfe et Shields, 1997), les Vertébrés (Dehal et Boore, 2005) et surtout les Plantes, chez qui ces événements ont été détectés dans plusieurs lignées (Tang et al., 2008). Ces événements ne sont pas toujours triviaux à détecter, à plus forte raison si l'événement est ancien et que de nombreux gènes issus de la duplication ont disparu. On peut néanmoins identifier des duplications individuelles par la tendance des gènes dupliqués à garder un ordre similaire sur leurs chromosomes respectifs (Dehal et Boore, 2005). La fixation des gènes dupliqués dans le génome est plus importante pour les événements de ce type et peut concerner des gènes dont la fixation est rarement observée dans le cas de duplications individuelles (Van de Peer et al., 2009). Les régulateurs de la transcription et du développement sont des exemples de gènes dont l'expansion, chez les Plantes comme chez les Vertébrées, est due en grande partie aux duplications de génomes complets. Quelles que soient leurs origines, il apparait toutefois que les duplications sont l'un des principaux mécanismes d'acquisition de gènes et sont donc propices au développement de nouvelles fonctions.

2.3.3.4 Les pertes de gènes

Le pendant de l'acquisition de nouveaux gènes, que ce soit par transferts horizontaux ou par duplications, est la perte de gènes. On ne peut d'ailleurs pas évoquer la duplication de gènes sans évoquer la nonfonctionalisation, qui est essentiellement la perte d'un des paralogues. Elle peut advenir suite à deux événements : la pseudogénisation qui correspond à une perte de fonction du gène et à l'accumulation de mutations au niveau de sa séquence ou sa disparition physique du génome par exemple par croisement inégal pendant la méiose (Albalat et Cañestro, 2016). Encore une fois, l'importance de la perte des gènes au cours de l'évolution des espèces a été réellement mise en évidence grâce à la comparaison de génomes complets.

Les premières séquences de Cnidaires (i.e. méduses et polypes), l'un des clades ayant divergé le plus anciennement des autres Métazoaires, permirent de comparer le répertoire de gènes communs aux eucaryotes. Ces comparaisons montrèrent que le répertoire total de gènes de l'ancêtre commun était plus étendu que ce qui était imaginé, et que seulement 72% de ce répertoire se retrouvait dans les grandes lignées eucaryotes (Putnam et al., 2007). La propension à la perte de gènes n'est pas identique selon les espèces considérées et les comparaisons de répertoires de gènes indiquent par exemple de nombreuses pertes spécifiques aux ecdysozoaires (Takahashi et al., 2009). Les pertes de gènes jouent également un rôle majeur chez les Procaryotes où le taux de pertes de gènes peut dépasser le taux de gains (Puigbò et al., 2014), et influent donc également sur la variation des répertoires de gènes. Les exemples les plus marquants de pertes de grande ampleur sont celles constatées chez les espèces parasites (Corradi et Slamovits, 2011) ou symbiotiques, connues pour avoir à la fois les génomes et le nombre de gènes les plus réduits (Moran, 2003) ; leurs pertes de gènes étant compensées par l'interaction avec l'hôte.

Comme pour les duplications, les pertes de gènes n'affectent pas toutes les catégories de gènes de la même façon, et les catégories affectées diffèrent selon les espèces. Principalement, elles dépendent de contraintes environnementales, différentes entre les espèces (Albalat et Cañestro, 2016).

Qu'il s'agisse d'architecture du génome, ou de répertoire de gènes, les comparaisons de génomes complets font état de la plasticité remarquable d'un objet pourtant universel. La suite de cette section montre de quelles façons on peut exploiter cette diversité pour extraire des connaissances fonctionnelles des séquences génomiques.

2.3.4 L'identification d'éléments fonctionnels

Comme mentionné au début de cette section, la fraction de régions codant pour des protéines est minoritaire dans les génomes eucaryotes (1% chez l'homme). Pour autant, le reste de la séquence génomique n'est pas inactif et on y retrouve des éléments fonctionnels. Ainsi, les études de génomiques comparatives menées sur plusieurs génomes de vertébrés estiment que 8 à 9% du génome humain montrent des signes de pressions de sélection et sont ainsi conservés entre les différentes espèces (Rands et al., 2014).
Ces éléments fonctionnels conservés sont pour la plupart des éléments régulateurs de l'expression des gènes :

1. Les régions régulatrices de la transcription : cette catégorie comprend les promoteurs, activateurs, inactivateurs et insulateurs. Ces régions sont des sites de fixation de protéines dont la présence influent sur la transcription des gènes (Maston et al., 2006).
2. Les ARN non codants : ARN de transfert, ARN ribosomiques mais aussi long ARN non codants, micro ARN, et petits ARN nucléaires et nucléolaires. Ces régions, transcrites en ARN mais non traduites, sont impliquées dans les différentes étapes menant à l'expression des gènes (Morris et Mattick, 2014).

Si on peut détecter ces éléments à l'aide de l'épigénomique et de la transcriptomique comme nous l'avons vu au premier chapitre, la génomique comparative permet d'accéder à une partie d'entre elles à moindre coût.

On peut détecter ce genre de régions conservées par la technique du *phylogenetic footprinting* : il s'agit d'aligner les régions orthologues du génome de plusieurs espèces et d'identifier des motifs dont la conservation est supérieure à ce qui est observé sur l'ensemble du génome (Blanchette et Tompa, 2002). Les premières applications de cette méthode se bornaient à l'étude de sites entourant des gènes protéiques, pour identifier les régulateurs (Tagle et al., 1988), mais avec les séquences de génomes complets, il est devenu possible d'aligner les régions orthologues de plusieurs génomes et de retrouver les sites conservés sur l'ensemble du génome (Margulies et al., 2003; Rands et al., 2014). Pour faciliter ces comparaisons, des mesures plus robustes de contraintes génomiques ont été développées, à l'image du score GERP (*Genomic Evolutionary Rate Profiling*) (Cooper et al., 2005) et ces informations sont publiquement accessibles sur les explorateurs de génomes comme Ensembl (Zerbino et al., 2018), afin de

faciliter l'analyse des génomes. La précision de ces méthodes dépend évidemment de la proximité, de la diversité et du nombre d'espèces utilisées pour la comparaison.

Chez les espèces ayant divergé récemment, l'ensemble de la séquence génomique est trop proche pour pouvoir déterminer clairement les zones plus conservées sans risquer de faux positifs et le *phylogenetic footprinting* est donc moins efficace. On peut cependant détecter les régions contraintes par une autre méthode, le *phylogenetic shadowing* (Boffelli et al., 2003) qui fonctionne selon le principe inverse. Au lieu de s'intéresser aux régions conservées à travers l'échantillon d'espèces, on s'intéresse à l'ensemble des sites où l'on retrouve des variations inter-espèces. Les sites où peu de variations sont observées correspondent aux éléments fonctionnels. Cette méthode est donc plus efficace pour retrouver les sites spécifiques à un clade ayant divergé relativement récemment, comme celui des primates (Boffelli et al., 2003).

Ces techniques d'identification illustrent la façon dont la génomique comparative peut extraire du sens des données génomiques, et ainsi apporter des informations complémentaires aux autres types de données omiques.

2.3.5 L'aide à l'inférence fonctionnelle

Nous avons vu comment les comparaisons de génomes permettaient d'identifier les éléments présents sur le génome, il nous reste donc à voir comment on peut les utiliser pour inférer leur rôle biologique en exploitant la plasticité génomique.

La synténie, ou la conservation de l'ordre des gènes sur le génome, est une première façon d'inférer la faction de gènes pour lesquelles aucune annotation n'existe (Osterman et Overbeek, 2003). En effet, nous l'avons vu dans une section précédente, la proximité génomique de gènes impliqués dans les mêmes fonctions biologiques tend à être conservée au cours de l'évolution, notamment chez les Procaryotes. Cette particularité permet des transferts d'annotations d'un gène à un autre, à la condition qu'ils soient proches sur le génome et que leurs orthologues soient retrouvés dans le même ordre chez d'autres espèces éloignées (von Mering et al., 2003; Yelton et al., 2011). Ces méthodes ont, par exemple, été utilisées avec succès pour l'identification de gènes manquants dans une voie métabolique (Osterman et Overbeek, 2003).

Plus significatif encore que la synténie, les liens fonctionnels peuvent être inférés à partir des événements de fusion, par la méthode de la Pierre de Rosette (Enright et al., 1999; Marcotte et al., 1999). La fusion des gènes est un contributeur considérable à l'évolution des protéines multidomaines (Buljan et al., 2010; Pasek et al., 2006). Le fait que ces gènes se retrouvent fusionnés chez certaines espèces suggère que les protéines qu'ils codent interagissent et sont donc impliquées dans les mêmes processus biologiques (Yanai et al., 2001), et peut donc aider à identifier la fonction d'un gène pour lequel aucune information existe.

La dernière méthode d'inférence fonctionnelle basée sur le génome que j'aborderai ici est non plus basée sur la position physique des gènes les uns par rapport aux autres, mais sur les événements de gains et de pertes de gènes. On fait l'hypothèse que deux gènes impliqués dans

les mêmes fonctions sont conservés et perdus de la même façon au cours de l'évolution, et donc qu'on les retrouve chez les mêmes espèces. On peut donc identifier la fonction d'un gène sur la base de sa présence et de son absence chez une variété d'espèces. Cette technique étant également une méthode de choix pour l'inférence de relations génotype-phénotype, je la traiterai en détail dans la section suivante.

Bien que moins fiables que les comparaisons de séquences et les relations d'orthologie pour inférer la fonction des gènes, les méthodes vues ici sont particulièrement utiles pour constituer des hypothèses fonctionnelles et complémenter les approches classiques. Ces différentes sources d'inférence sont utilisées à grande échelle pour prédire les interactions protéines-protéines dans la base de données STRING (Szklarczyk et al., 2015). Ces réseaux peuvent venir complémenter les informations d'autres sources, notamment de protéomique, dans la compréhension des systèmes biologiques.

Ce tour d'horizon des applications de la génomique comparative était volontairement large et relativement superficiel, à la fois pour introduire des notions sur lesquelles je me suis appuyé au cours de mes travaux de thèse et pour démontrer le potentiel de l'ensemble des données génomiques accessibles. La dernière section de ce chapitre traite concrètement et plus en détail la manière dont on peut utiliser la génomique comparative pour traiter les relations génotype-phénotype.

2.4 La génomique comparative et les inférences génotype-phénotype

L'ensemble du Vivant regroupe une diversité phénotypique extrêmement riche, de bactéries unicellulaires de moins d'un micromètre à l'arbre *Sequoia sempervirens* pouvant atteindre plus de 100 mètres de haut. Les particularités phénotypiques propres à chaque espèce étant, en général, connues, on peut aujourd'hui associer l'ensemble de ces traits à des différences au niveau du répertoire génétique. On exploite ainsi directement les effets de la plasticité du contenu en gène d'un génome.

Dans cette section, j'aborde les méthodes utilisées pour réaliser cette association, d'abord entre peu d'espèces avec l'association soustractive et sa généralisation adaptée à l'étude combinée de nombreux génomes, le profilage phylogénétique. J'explicite les méthodologies utilisées pour ces derniers, qui constituent un socle important de ces travaux de thèse. Pour expliciter le potentiel de ces techniques, je termine en présentant leur application à l'étude du cil eucaryote, un objet unique dans l'étude des relations génotype-phénotype.

2.4.1 Association soustractive

L'association soustractive a pour objectif d'identifier les gènes spécifiques aux espèces ayant un phénotype donné. Dans les faits, cela revient à comparer le répertoire de gènes d'une espèce d'intérêt A possédant le trait phénotypique d'intérêt à deux autres espèces : une espèce C proche taxonomiquement qui en est dépourvue et une espèce B plus éloignée le possédant. Pour cela, on fait l'intersection des répertoires de gènes des deux espèces A et B présentant le phénotype

(encadré sur la Figure 2-7) en identifiant toutes les relations d'orthologie entre ces deux espèces. On y soustrait ensuite les gènes ayant un orthologue dans l'espèce C, négative pour le phénotype d'intérêt (en rouge sur la Figure 2-7). On considère que les gènes restants sont associés à ce trait phénotypique (en bleu sur la Figure 2-7).

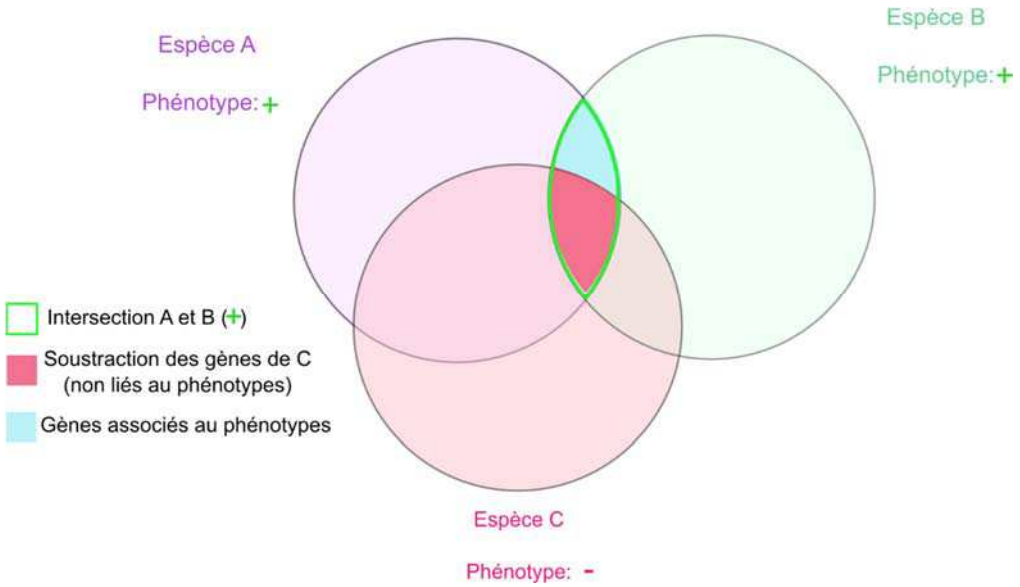


Figure 2-7 **Association soustractive**. Le diagramme de Venn montre les intersections des répertoires de gènes de trois espèces possédant (+) ou non (-) un phénotype d'intérêt. Les trois ensembles de gènes pertinents pour cette analyse sont identifiés.

Cette approche a été introduite par Huynen (Huynen et al., 1998) pour la comparaison du génome du pathogène *Helicobacter pylori* à celui d'un autre pathogène *Haemophilus influenzae* et d'une souche bénigne d'*Escherichia coli*. Elle a mis en évidence 17 gènes communs liés aux interactions hôtes-pathogènes. Elle fut ensuite appliquée (Li et al., 2004) à des génomes eucaryotes pour l'identification de gènes liés à un organelle eucaryote, le cil ou flagelle. Cette étude s'intéressait au répertoire de gènes de *Chlamydomonas reinhardtii*, une algue flagellée, comparé à celui d'une autre plante, *Arabidopsis thaliana* dépourvue de cil et à celui de l'Homme, qui possède des cellules ciliées. Elle permit d'identifier 688 protéines communes aux seules espèces ciliées, comprenant la majorité des gènes connus comme étant liés au cil (90%) et permettant la validation de nouveaux gènes associés à l'organite.

L'approche soustractive est applicable à la recherche de tous gènes liés à un trait phénotypique ou processus biologique ayant été perdu chez au moins une espèce au cours de l'évolution (Bedez et al., 2013). Impliquant peu d'espèces, sa mise en place est possible avec peu de données génomiques et nécessite de calculer relativement peu de relations d'orthologie. Revers de la médaille, la comparaison de peu d'espèces la rend sujette aux faux positifs.

Conceptuellement, il est possible de réduire ce risque en impliquant plus d'espèces dans la comparaison, le principe étant toujours de retrouver les gènes présents uniquement dans toutes les espèces présentant le phénotype et absents des autres (Avidor-Reiss et al., 2004). Si l'on

souhaite l'étendre à plus d'une dizaine de génomes, ce type d'analyse nécessite cependant de changer de cadre méthodologique, on passe alors dans le domaine du profilage phylogénétique.

2.4.2 Profilage phylogénétique

Le profil phylogénétique d'un gène représente la présence ou l'absence d'orthologues de ce gène dans les génomes de plusieurs espèces (Tatusov et al., 1997). Les profils phylogénétiques sont une représentation idéale pour la généralisation des associations soustractives car l'on peut considérer dans le même temps un nombre important d'espèces (jusqu'à plusieurs centaines). Avant d'aborder les applications aux relations génotype-phénotype, je présente rapidement le contexte dans lequel les méthodes de profilage phylogénétique ont été développées dans un premier temps, à savoir l'inférence fonctionnelle par co-occurrence de gènes.

2.4.2.1 Inférence par co-occurrence

L'hypothèse sur laquelle reposent les analyses fonctionnelles par profils phylogénétiques est la suivante : deux gènes liés fonctionnellement tendent à être préservés ou perdus de manière corrélée au cours de l'évolution (Pellegrini et al., 1999). Selon cette hypothèse, les gènes dont les orthologues sont présents et absents dans les mêmes génomes (*g1* et *g3* dans la Figure 2-8) ont plus de chances d'être liés fonctionnellement. Cette hypothèse permet d'inférer la fonction de gènes non-caractérisés ou des interactions physiques entre protéines. De fait, cette méthode a été appliquée avec succès à l'annotation de gènes, principalement procaryotes (voir (Kensche et al., 2008) pour une liste non-exhaustive). L'étude menée par Cunningham et al., 2000 en est un exemple type : les auteurs se sont intéressés à la voie métabolique MEP/DOXP de synthèse des isoproprénoïdes dont l'ensemble des réactions étaient connues, mais dont le catalyseur de l'une des étapes manquait. Le profilage phylogénétique des enzymes déjà connues dans 30 espèces leur permit de retrouver un gène partageant la même distribution. L'analyse expérimentale leur confirma ensuite qu'il s'agissait du catalyseur manquant.

La méthode d'occurrence exclusive est une variante de l'inférence par co-occurrence et fonctionne sur le principe inverse : au lieu de s'intéresser aux gènes dont les profils sont très similaires, on recherche ceux dont les profils sont mutuellement exclusifs. Lorsque le gène *g3* (Figure 2-8) est absent d'une espèce, le gène *g4* y est présent et inversement. On peut de cette façon identifier les remplacements non-orthologues, c'est-à-dire les cas où des gènes sans relation d'homologie remplissent un rôle similaire dans des espèces différentes (Galperin et Koonin, 2000).

Avec les autres méthodes d'inférence par contexte génomique, traitées en fin de section précédente et auxquels elle peut être associé (Gabaldón et Huynen, 2004; Lee et al., 2004), la prédiction par co-occurrence permet de construire des hypothèses solides sur la fonction potentielle des gènes pour guider les expérimentations. Au même titre que les autres méthodes, les informations de co-occurrence sont disponible dans la base de données d'interactions protéines-protéines STRING (von Mering et al., 2003).

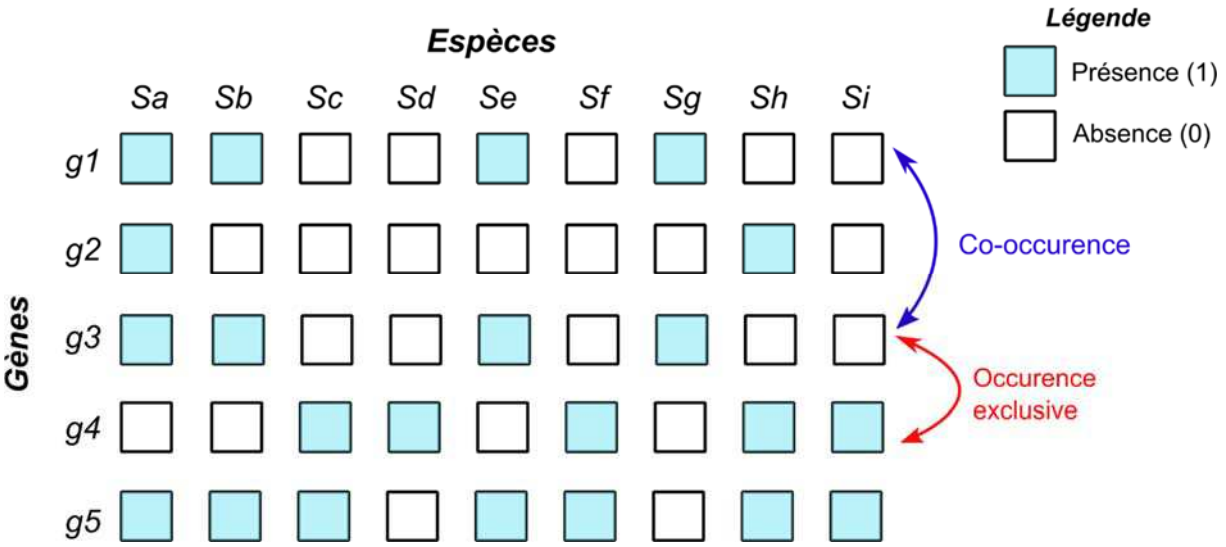


Figure 2-8 **Co-occurrence et occurrence exclusive.** Profils phylogénétiques de 5 gènes dans 9 espèces différentes (Sa-Si). Les gènes *g1* et *g3* sont présents et absents dans les mêmes espèces, ils sont potentiellement liés fonctionnellement. *g4* est uniquement présent dans les espèces où *g1* et *g3* sont absents, et inversement, il s'agit d'un cas potentiel de remplacement non orthologue.

2.4.2.2 *Prédiction de relations génotype-phénotype*

La force du profilage phylogénétique, et ce qui nous intéresse principalement ici, réside dans le fait que les profils peuvent non seulement être comparés entre eux mais également à tous types de distribution de présence-absence, y compris aux traits phénotypiques. En cela, il s'agit d'une extension de la méthode d'analyse soustractive adaptée à l'intégration des informations provenant d'un nombre important d'espèces.

De cette façon, l'une des premières études d'association phénotype-génotype de grande ampleur, réalisée avec 86 génomes procaryotes (Jim et al., 2004) permit d'identifier des gènes associés à la thermophilie, au *pilus* ou encore au flagelle procaryotes (malgré leur dénomination commune, le flagelle procaryote n'est pas homologue, mais analogue au flagelle eucaryote). Les relations génotype-phénotype sont particulièrement étudiées dans le domaine de la santé, et l'explosion du nombre de génomes eucaryotes permet maintenant d'explorer l'ensemble des gènes humains à la lumière de leurs distributions. Une telle tâche fut entreprise par Tabach et ses collaborateurs (Tabach et al., 2013a) et permit d'identifier une corrélation entre groupes de profils phylogénétiques similaires et des maladies génétiques humaines. Plus précisément, ils catégorisèrent les gènes humains en 1026 *clusters* ayant un profil, donc une histoire évolutive similaire. Une mesure de surreprésentation des gènes annotés par des termes *Human Phenotype Ontology* (HPO) leur permis d'identifier 54 *clusters* significativement associés à une catégorie de symptômes spécifiques.

2.4.3 Méthodologies de profilage phylogénétique

Les techniques de profilage phylogénétique reposent essentiellement sur la comparaison de profils de présence et d'absence d'un gène mais l'on peut implémenter ces analyses de façon multiple. J'aborde ici les différentes possibilités qui existent à chaque étape de ces analyses.

2.4.3.1 La construction

La première étape à réaliser dans les analyses de profilage phylogénétique est, bien sûr la génération des profils à partir des génomes des espèces considérées.

La détermination de la présence ou l'absence d'un gène dans des espèces différentes nécessite la recherche des orthologues de ce gène dans les génomes des espèces considérées. Il existe plusieurs façons d'inférer l'orthologie, mais cela sera abordé en détail dans le prochain chapitre). Une fois que l'on a connaissance des relations d'orthologie du gène, l'on créé un vecteur binaire de longueur égale au nombre d'espèces considérées, en attribuant une valeur de 1 aux espèces présentant un orthologue, et 0 dans le cas inverse.

Rien n'oblige, cependant, à se contenter de valeurs et l'on peut intégrer d'autres données aux profils en remplaçant l'indicateur de présence par une valeur correspondante à la similarité de séquences avec l'orthologue le plus proche. Généralement, ces scores sont tirés de l'E-value (Marcotte, 2000) ou du bit score (Enault et al., 2003; Tabach et al., 2013a) obtenus par recherche BLASTP ou BLASTN (Altschul et al., 1997), parfois normalisés pour tenir compte de la distances entre les espèces considérées et de la vitesse d'évolution de chacune. On parle alors de profils phylogénétiques continus, qui permettent notamment de prendre en compte les cas où les gènes ne sont pas perdus mais dont la séquence montre des signes de pression sélective relâchée.

Dans une dernière variante (Ranea et al., 2007), les scores de présence-absence sont remplacés par le nombre d'orthologues retrouvés dans chacune des espèces considérées de manière à prendre en compte les évènements de duplications dans ces comparaisons. Les profils phylogénétiques, quel que soit leur type, sont le plus souvent construits pour chaque gène protéique de l'espèce pour former une matrice de présence-absence de p (nombres de protéines) lignes sur n (nombre d'espèces considérés) colonnes : la matrice phylogénétique.

La sélection des espèces à prendre en compte est une étape très importante lors de la construction de profils phylogénétiques pour éviter certains biais lors de l'analyse. La précision des méthodes de profilage phylogénétique dépend du nombre, mais surtout de la diversité des espèces sélectionnées : pour le même nombre d'espèces, les mesures intégrant des représentants de plusieurs clades sont toujours plus efficaces que celles incluant uniquement les représentants d'un seul clade (Skunca et al., 2013; Sun et al., 2005). Le second facteur à considérer est la complétion et la qualité des prédictions de gènes des génomes sélectionnés, des génomes de faible qualité entrainant des erreurs aux niveaux des profils.

Les matrices phylogénétiques ainsi obtenues sont, en quelque sorte, une représentation synthétique de l'information contenue dans les génomes faisant partie de l'analyse. Reste à les analyser de différentes manières pour identifier les cas de co-occurrence ou d'association à un trait phénotypique, c'est-à-dire extraire du sens biologique de cette information condensée. Les prochaines sections décrivent la façon pour procéder à cette dernière étape.

2.4.3.2 Comparaisons de profils

Une première possibilité d'analyse des profils est la recherche de gènes dont le profil est similaire au profil du gène ou du phénotype d'intérêt. Si la tâche, une comparaison de vecteurs, semble conceptuellement simple, plusieurs catégories de méthodes ont été développées pour s'adapter aux particularités des profils phylogénétiques.

1) Les méthodes « naïves »

Les profils étant représentés par de simples vecteurs de valeurs (binaires ou non), les méthodes naïves utilisent des méthodes de comparaison conçue pour ce genre de données. La méthode de comparaison la plus évidente est la recherche de profils identiques (Pellegrini et al., 1999), mais cela est vite limité par la nature non-déterministe des processus biologiques et les erreurs possibles dans les profils de présence-absence. Pour les vecteurs binaires, les plus utilisées sont la distance de Hamming (Kensche et al., 2008; Pellegrini et al., 1999) égale aux nombres d'espèces pour lesquelles les valeurs de présence ou d'absence diffèrent, et le complément de l'indice de similarité de Jaccard, qui équivaut à l'intersection des deux vecteurs (co-présence) divisé par leur union (somme des présences distinctes) (Glazko et Mushegian, 2004; Yamada et al., 2006). Ce dernier est donc plus strict pour la co-occurrence de gènes observés dans peu d'espèces.

Alternativement aux mesures de distances, des mesures de corrélations statistiques sont applicables aux vecteurs binaires comme continus: le coefficient de corrélation de Pearson (Glazko et Mushegian, 2004) et l'Information Mutuelle (Wu et al., 2003). A l'inverse du coefficient de corrélation, cette dernière ne discrimine pas les profils ayant une corrélation exacte et ceux de distribution inverse (anti-corrélant), ce qui en fait une méthode idéale lorsqu'on s'intéresse à l'inférence par 'occurrence exclusive' (Glazko et Mushegian, 2004).

Finalement, une dernière classe de méthode estime la dépendance entre deux profils par des tests statistiques. Le test exact de Fisher (Barker et Pagel, 2005) évalue l'hypothèse que deux gènes soient indépendants : on considère que les profils de gènes sont liés si la *P-value* (probabilité de l'observation sous l'hypothèse nulle) est faible. Les méthodes basées sur la distribution hypergéométrique (Wu et al., 2003) mesurent la probabilité d'observer des correspondances de profils par chance en fonction du nombre d'observations par profil, et peuvent être adaptées pour prendre en compte des paramètres comme le nombre d'orthologues totaux retrouvés dans chaque espèce considérée (Kharchenko et al., 2006).

Toutes ces méthodes naïves ont l'avantage d'être simples à mettre en œuvre, mais ont le défaut de ne pas prendre en considération les particularités des profils phylogénétiques. En effet, elles font l'hypothèse que les variables de présence-absence sont indépendantes les unes des autres, or ce n'est pas le cas, elles sont le reflet d'histoires évolutives qui sont d'autant plus partagées que les espèces sont proches (Kensche et al., 2008) (Figure 2-9).

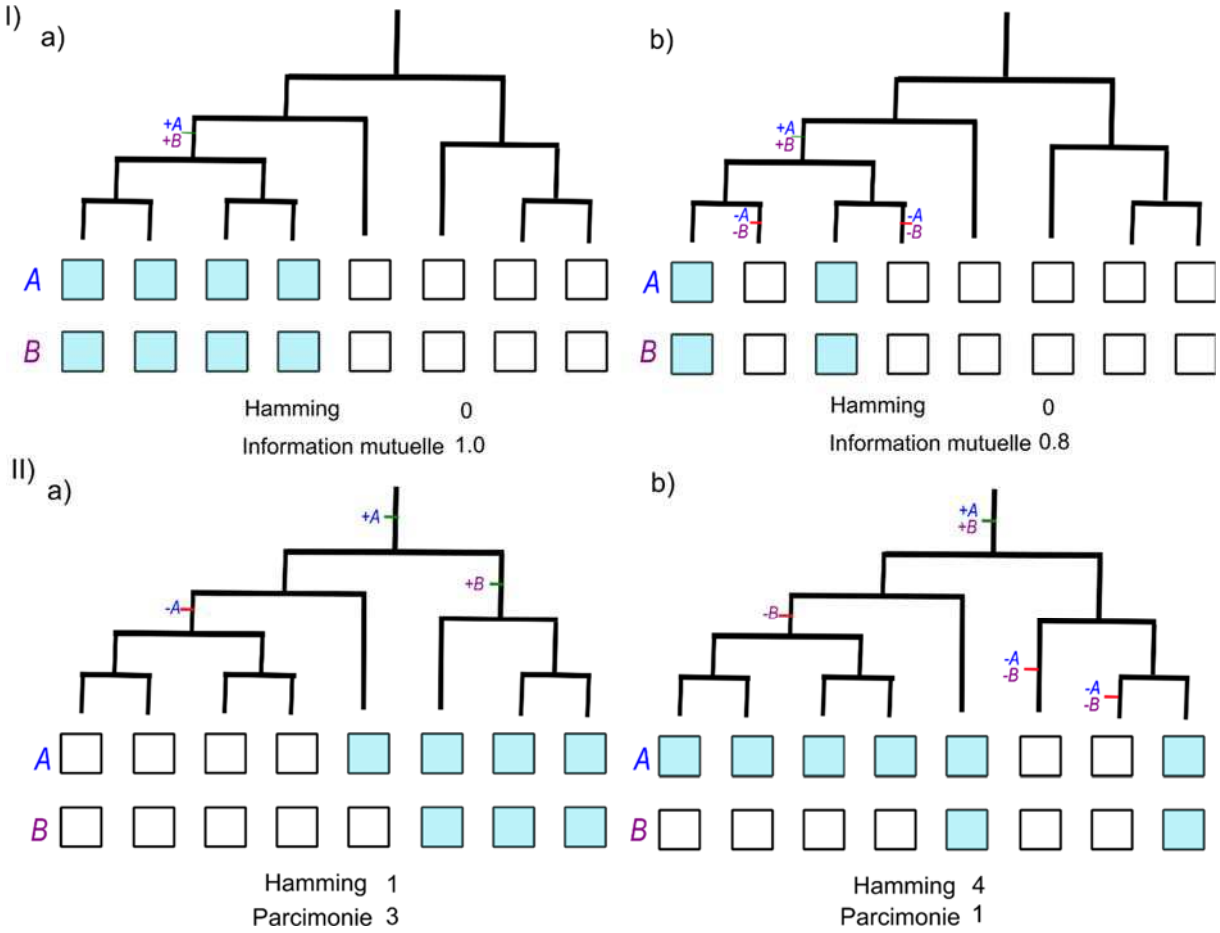


Figure 2-9 **Influence de la non-indépendance sur les méthodes naïves**. (I) Dans le cas de figure a, deux gènes sont présents dans la moitié des espèces considérées, suite à un événement de gain, et les méthodes de Hamming et d'Information Mutuelle indiquent une forte similarité. Dans le cas de figure b, deux événements de pertes supplémentaires renforcent les preuves de co-occurrence mais l'Information Mutuelle est moins forte. (II) Dans le cas de figure a, les profils similaires sont dus à des événements indépendants mais la distance de Hamming indique de la similarité. Dans le cas de figure b, le gène B a seulement subi un événement de perte supplémentaire par rapport à A, mais la distance de Hamming est élevée. L'utilisation de méthodes basées sur les arbres permet de détecter ces événements et renforcent les prédictions, comme l'indique la valeur de parcimonie. Exemples tirés de *(Kensche et al., 2008)*.

2) Méthodes avec modèle d'évolution *:*

Un profil de présence-absence est en principe le résultat d'un évènement d'acquisition de gène ancestral (plusieurs en cas de transferts horizontaux) et d'éventuels événements de perte. L'absence de prise en compte de ces informations peut facilement mener à des aberrations dans

les mesures de similarité (Figure 2-9), à plus forte raison quand certains clades sont surreprésentés. Certaines des méthodes de comparaison utilisent donc des modèles d'évolution basés sur des arbres phylogénétiques, elles comptent les modèles de parcimonie, de Maximum de Vraisemblance et la méthode de *tree kernel* (Kensche et al., 2008), dont je fais une brève description ici.

Le modèle de parcimonie reconstruit, à partir des profils observés, la succession d'évènements de gains et de pertes de gènes nécessitant le moins d'évènements évolutifs. Ces événements sont reportés dans un vecteur sur lequel on peut appliquer les méthodes de comparaison vues plus haut (Barker et al., 2007) et la distance ainsi observée dépendra du nombre de différences dans la succession d'événements ayant menée à chaque profil (Figure 2-10). Les méthodes de Maximum de Vraisemblance prennent en compte les différents scénarios évolutifs et scorent l'efficacité respective d'un modèle de gains et pertes dépendants (les gènes sont perdus et gagnés ensemble) et d'un modèle indépendant pour expliquer le profil (Barker et al., 2007; Barker et Pagel, 2005). Dans la méthode à 'noyau d'arbre', les profils sont décrits, dans un espace vectoriel à grande dimension, par leurs différentes histoires évolutives possibles dont la probabilité dépend du modèle d'évolution par arbre bayésien (Vert, 2002). L'utilisation d'arbres phylogénétiques et de modèles d'évolution donne un avantage à ces méthodes par rapport à des distances simples, cependant leur mise en application est généralement couteuse en termes de ressources informatiques, d'autant plus que le nombre d'espèces considérées est important.

3) Méthodes de prise en compte indirecte de la phylogénie.

Il est possible de prendre en compte des informations issues de la phylogénie de manière plus simple et moins couteuse en ressources informatiques. Ces méthodes, à mi-chemin entre les deux types d'approches décrites précédemment utilisent les arbres phylogénétiques pour construire des vecteurs estimant les événements de gains et de pertes à partir du vecteur de présence absence.

La mesure de co-occurrence implémentée dans la base de données STRING (von Mering et al., 2003) réduit d'abord, pour chaque comparaison deux à deux de profils, les cas où plusieurs variables sont homogènes dans plusieurs espèces descendant d'un même ancêtre commun par une seule variable représentant l'état ancestral pour réduire l'impact des clades de taille importante (Figure 2-10). La méthode introduite par (Dey et Meyer, 2015) ordonne quant à elle les espèces selon un arbre phylogénétique et construit un vecteur représentant les transitions observées entre présence et absence comme estimation des événements de gains et de pertes (Figure 2-10).

Dans les deux cas, des méthodes de mesures « naïves » sont ensuite appliquées pour comparer les vecteurs ainsi générés, spécifiquement STRING utilise l'Information Mutuelle et Dey une mesure de similarité basée sur le nombre de transitions partagées entre les vecteurs. En principe, on peut toutefois y appliquer toute mesure adaptée à des vecteurs binaires et le résultat obtenu est plus proche des mesures basées sur les modèles (Figure 2-10). Ces méthodes indirectes, bien

que moins précises que les méthodes basées sur des modèles sont toutefois rapides et donc applicables à des volumes de données importants.

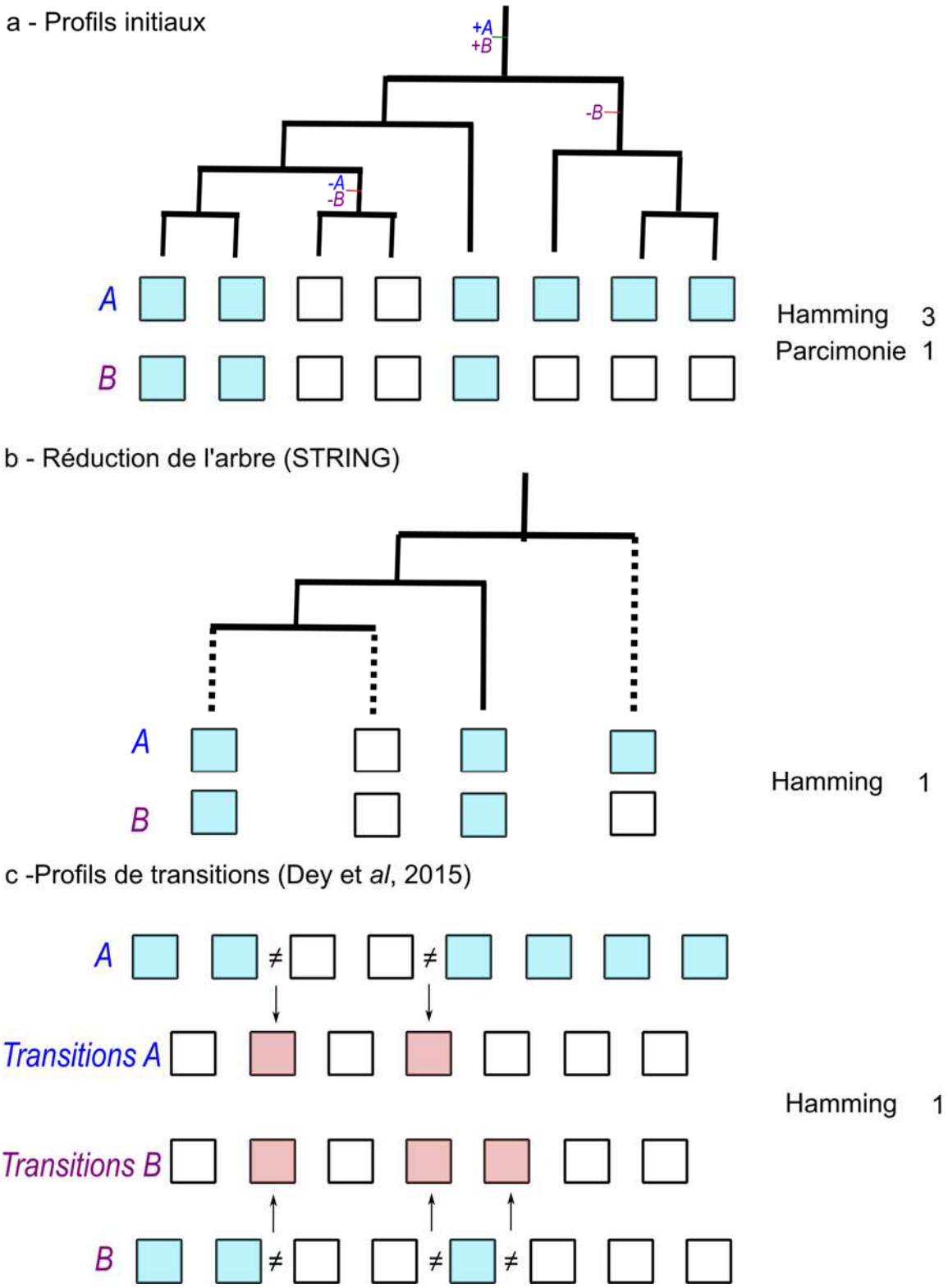


Figure 2-10 **Transformation de profils basés sur la phylogénie**.a - Les profils initiaux présence-absence des gènes A et B, et leur histoire évolutive respective. Pour ces profils, la mesure de Hamming indique des différences importantes alors que la parcimonie reconnait la similarité d'histoire évolutive. b –La réduction de l'arbre, utilisée dans String, consiste à donner une seule valeur de présence-absence à chaque branche de l'arbre phylogénétique où les profils sont homogènes (branches en pointillés). c –

Les profils de transitions *(Dey et Meyer, 2015)* indiquent les transitions entre présence et absence des profils initiaux préalablement ordonnés selon la taxonomie. Pour b et c, une mesure naïve (Hamming pour cet exemple) reflète mieux la similarité d'histoire évolutive que lorsqu'elle est calculée sur les profils initiaux.

Bien qu'en principe plus efficaces que les comparaisons naïves, toutes les méthodes basées sur les arbres, directes ou indirectes ont quelques faiblesses majeures. Elles supposent notamment la connaissance exacte de l'arbre des espèces, un pré-requis qui n'est pas toujours confirmé au vu de notre connaissance actuelle de la taxonomie ; et elles sont également plus sensibles aux erreurs dans les prédictions de présence-absence.

2.4.3.3 Classification des profils

Les mesures de distance vues ici mesurent la similarité entre deux profils donnés. Elles sont en général appliquées à l'ensemble des profils d'une matrice phylogénétique pour réaliser des classifications automatiques et identifier des groupes de gènes avec une histoire évolutive similaire et membre d'un même complexe macromoléculaire ou simplement impliqués dans une même fonction (modules fonctionnels). On peut les classifier selon plusieurs protocoles :

1) Classements des profils similaires

La méthode la plus simple, lorsque l'on cherche les gènes co-occurrents à un gène ou un phénotype dont le profil de présence-absence est connu, consiste à sélectionner les gènes dont le profil est proche, c'est-à-dire ceux ne dépassant pas un seuil de dissimilarité et de les classer par ordre de proximité. Les gènes ainsi identifiés ont une probabilité importante d'être associés au gène ou au phénotype recherché. Cette méthode ne nécessite des mesures de distances qu'entre le profil connu et chacun des profils de la matrice phylogénétique, par opposition aux autres méthodes de classifications qui reposent sur des mesures entre tous les profils de la matrice.

2) *Clusterings*

Les matrices phylogénétiques de taille importante peuvent être exploitées d'un point de vue plus global en réalisant une classification de l'ensemble des profils en groupes montrant une forte similarité entre eux, de façon à identifier des modules fonctionnels (Snel et Huynen, 2004). Les algorithmes de classification non supervisée (*clustering*) sont particulièrement adaptés à cette tâche et ne nécessitent qu'une matrice représentant les distances entre chacun des profils. On utilise souvent les algorithmes de *clustering* hiérarchique (Niu et al., 2017; Staub et al., 2006; Tabach et al., 2013b) pour ce type de classification car ils ne nécessitent pas de connaître *a priori* le nombre de classes de profils qui existent, à la différence notamment de l'algorithme des *k-means* (Glazko et Mushegian, 2004). Brièvement, le *clustering* hiérarchique regroupe itérativement les profils les plus similaires entre eux jusqu'à obtenir un arbre représentant leur relation de proximité. On peut ensuite, à partir de cet arbre, regrouper les profils selon différents niveaux de granularité : peu de classes dont les profils ont seulement en commun des

caractéristiques facilement distinguables (présence dans un clade seulement) ou de nombreuses classes avec des profils plus homogènes. La stratégie « agglomérative » du *clustering* hiérarchique, qui consiste à étendre les classes itérativement en rajoutant un profil proche à ceux déjà dans la classe, peut être adaptée pour prendre en compte la notion de seuil de dissimilarité vu plus haut, à la manière de Dey and Meyer, 2015, et ainsi se limiter à la création de classes pouvant refléter des modules fonctionnels.

3) Classifications de graphes

Alternativement, les modules de gènes peuvent être définis à partir d'une représentation sous forme de graphes où les profils sont les nœuds et leur similarité les arrêtes. On peut par exemple construire ces réseaux en créant des arrêtes uniquement pour les meilleurs scores de similarité de chaque gène (Shin et Lee, 2017) ou en considérant seulement les relations sous un seuil de dissimilarité donné (Beck et al., 2018). On peut bien sûr choisir d'intégrer au graphe les scores de similarité en attribuant un poids à chaque arrête. Dans ce genre d'association, les modules de gènes similaires sont identifiés par des algorithmes de *clustering* adaptés aux graphes (Beck et al., 2018). L'avantage des structures en réseaux pour étudier les relations entre profils est la possibilité d'exploiter de nombreux outils mis au point pour étudier les réseaux biologiques ou celle de les intégrer à des réseaux issus d'autres données *omiques*.

4) Identifications des modules fonctionnelles

Une fois les gènes classés en catégories, il reste à identifier à quels processus biologiques ces catégories représentent, le cas échéant. Pour ce faire, on part de ressources liant des gènes à une ou plusieurs fonctions ou processus biologiques tels que *Gene Ontology* (Ashburner et al., 2000) ou Reactome (Croft et al., 2014) et on effectue des tests d'enrichissement (Dey et Meyer, 2015; Pellegrini, 2012), de façon à déterminer si l'on retrouve dans chaque module une proportion plus importante de gènes associés à une fonction particulière que dans la population totale. On associe ainsi aux différents modules les fonctions moléculaires, processus biologiques ou composants cellulaires pour lesquels l'enrichissement est significatif. Si les enrichissements de ce type sont le plus souvent réalisés pour des processus biologiques, il est conceptuellement possible de les réaliser pour d'autres types d'annotation (par exemple pathologies ou phénotypes (Tabach et al., 2013b)), voire par rapport aux résultats d'autres expériences biologiques. Associer les modules de gènes à une fonction ou un phénotype donné aide ensuite à l'inférence de la fonction des autres gènes du module, par le principe de « culpabilité par association ».

2.4.3.4 Apprentissage automatique

Une méthode émergente pour identifier des gènes liés à des fonctions ou des phénotypes donnés à partir de leurs profils sont les méthodes d'apprentissage automatique (*Machine Learning*), au sens large.

D'une manière générale, ces techniques utilisent les profils d'une liste de gènes fournis par l'utilisateur, connus pour être associé à la fonction ou au phénotype recherché, les algorithmes d'apprentissages dérivent ensuite cette information pour identifier d'autres gènes dont les profils correspondent à l'ensemble d'apprentissage. Le programme CLIME (*CLustering by Inferred Model of Evolution*) (Li et al., 2014) illustre ce concept efficacement : il partitionne l'ensemble de gènes fourni en fonction de leur histoires évolutives probables (gains et pertes), en s'appuyant sur un arbre phylogénétique. Pour chacune des partitions de l'ensemble, l'algorithme score ensuite les autres gènes du génome selon la probabilité qu'ils aient suivies l'histoire évolutive correspondante. Ainsi, cet algorithme permet de retrouver d'autres gènes pouvant être liés à cette fonction, en prenant en compte l'hétérogénéité des histoires évolutives des gènes liés à un processus.

Une autre méthode d'apprentissage, Clus-HMC-Ens (Skunca et al., 2013; Škunca et Dessimoz, 2015; Weißenborn et Walther, 2017), utilise des annotations de référence (telles que *Gene ontology*) comme ensemble d'apprentissage, pour annoter ensuite les gènes en fonction de leur profil. Il s'agit d'un algorithme basé sur les arbres de décision capable de gérer des classifications hiérarchiques à étiquetage multiple (un gène peut avoir plusieurs annotations). Les annotations de gènes utilisées dans ces méthodes peuvent être des descripteurs de fonctions biologiques, d'appartenance à une voie métabolique ou de signalisation, ou leur association à des traits phénotypiques.

2.4.3.5 Analyses guidées par les connaissances (knowledge-driven approaches)

Qu'il s'agisse des méthodes de classification basées sur les distances ou d'apprentissage automatique, les méthodes que nous venons de voir sont automatisées sans *a priori* quant à ce qui est recherché. Lorsqu'elles sont appliquées à la recherche d'association à un phénotype d'intérêt (Takabayashi et al., 2009), on peut diriger l'analyse des profils phylogénétiques à la lumière des informations connus sur le phénotype, sa distribution dans les espèces considérées en utilisant des métriques adaptées à la question biologique.

Ainsi pour l'identification de gènes liés à la thermophilie, Makarova et collègues (Makarova et al., 2003) sélectionnèrent les gènes selon un ensemble de règles définies : (i) présence dans au moins trois espèces thermophiles, (ii) présence dans plus de thermophiles que dans d'autres espèces, et (iii) présence à la fois dans des bactéries et des archées thermophiles.

La mise en place de telles règles permet de prendre en compte plus facilement les biais éventuellement introduits par la phylogénie. Dans cet exemple précis, la troisième contrainte permet d'éviter un biais en gènes spécifiques des Archées qui ne seraient pas forcément lié à la thermophilie : les archées considérées dans cette étude étaient en majorité thermophiles et représentaient la plupart des espèces thermophiles considérées.

Le second avantage de ce type d'approches est qu'elles ne se limitent pas à retrouver des profils similaires au profil de présence-absence du caractère comme les méthodes basées sur les distances. Elles permettent à la fois d'être moins restrictif par rapport aux pertes dans les

espèces ayant le caractère recherché et plus strict par rapport à l'absence dans d'autres espèces, l'ensemble des paramètres de score étant contrôlé par l'expérimentateur.

Dans le principe, ces approches guidées par les connaissances sont des extensions de l'approche soustractive adaptées pour intégrer un nombre plus important d'espèces dans l'analyse. C'est donc sans surprise que ces analyses guidées ont également été utilisées pour la prédiction de protéines du cil eucaryote au cours des années (Avidor-Reiss et al., 2004; Hodges et al., 2011; Merchant et al., 2007). La dernière en date est notamment intéressante car elle profite de l'augmentation du nombre de génomes pour automatiser la recherche et attribuer un score à chaque protéine en fonction du nombre d'espèces ciliées et non-ciliées dans laquelle on la retrouve. Bien que simple, un tel système permet de prendre en compte la nature non déterministe de la biologie tout en donnant un score de confiance aux gènes retrouvés de cette manière.

2.5 Le cil eucaryote, un cas d'étude pour les relations génotype-phénotype

Comme nous l'avons vu, le cil eucaryote a reçu une attention particulière lorsqu'il s'agit d'études basées sur la présence-absence et notamment le profilage phylogénétique. Dans cette section, qui conclue ce chapitre, nous verrons en quoi l'histoire évolutive unique du cil en fait un objet d'étude de choix pour la génomique comparative. Nous aborderons ensuite ses différents rôles dans l'organisme et les phénotypes variés résultant de son dysfonctionnement, en faisant, là encore, un cas d'étude pour les relations génotype-phénotype.

2.5.1 Pertes et profits : le cil sous l'œil de la génomique comparative

Le cil eucaryote est présent dans l'ensemble des clades majeurs des eucaryotes et considéré, à ce titre, comme l'un des caractères du dernier ancêtre commun des eucaryotes (*Last Eukaryotic Common Ancestor*, LECA). Pour autant, il est sujet à une grande diversité chez les Eucaryotes, à tel point qu'il régulièrement pris en compte comme critère de classification des espèces (Adl et al., 2012). Plus remarquable encore, cette organelle a été perdue plusieurs fois au cours de l'évolution et de manière indépendante, ce qui aboutit à des représentants ciliés et non ciliés au sein de chaque clade majeur (Figure 2-11). Si l'on se réfère à nos connaissances actuelles de la taxonomie, cela correspond à un minimum de cinq à dix événements de perte dans l'histoire évolutive des eucaryotes. Parmi les cas emblématiques de perte du cil, on note deux pertes chez les champignons : l'une chez l'ancêtre des Microsporidies et une seconde chez l'ancêtre commun aux Zygomycètes et aux Dikarya (Liu et al., 2006). On peut également remarquer son absence quasi-totale chez les plantes terrestres, due à deux événements de pertes : l'un chez les Gymnospermes et un second à la base des Angiospermes (Hodges et al., 2012).

De plus, de grandes différences phénotypiques existent entre les cils des espèces l'ayant conservé, l'un des points majeurs étant la motilité. Si le cil est, en l'état de nos connaissances, observable chez tous les métazoaires où il est généralement essentiel au déplacement des gamètes, l'ensemble des nématodes, dont l'organisme modèle *Caenorhabditis elegans,* possèdent seulement des cils non motiles.

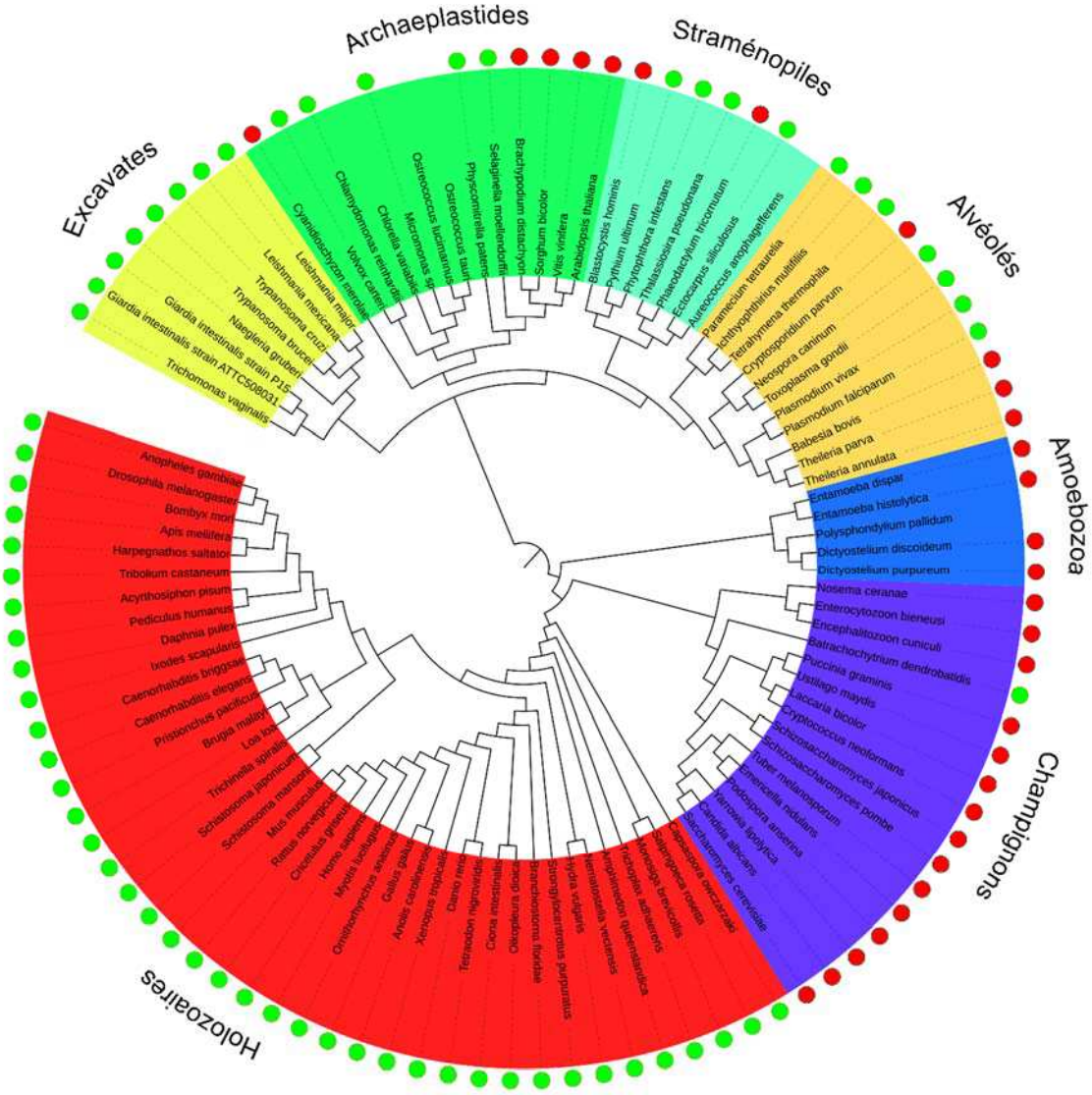


Figure 2-11 **Distribution du cil chez les eucaryotes.** Arbre taxonomique de 100 espèces d'eucaryotes, dont les colorations correspondent à des taxons majeurs. La présence du cil est représentée par un cercle vert, son absence par un cercle rouge. Si le phénotype n'est pas renseigné, aucun cercle n'est présent. Le cil est présent et absent dans des représentants de la plupart des clades.

Selon l'hypothèse fondatrice du profilage phylogénétique, les gènes liés à un phénotype sont conservés et perdus avec celui-ci. Le nombre de pertes importantes marquant l'histoire du cil conduit à un profil de présence-absence très particulier. Cela est également vrai pour les variations phénotypiques qui le caractérisent. Cette particularité du cil explique qu'il ait été la cible de tant d'attention pour les inférences par génomique comparative (mentionnées plus tôt dans ce chapitre). Ces études ont également été en grande partie motivées par l'importance de cette organelle dans le fonctionnement des cellules eucaryotes et par extension, chez les Métazoaires, des organismes multicellulaires.

2.5.2 Organisation du cil

Avant d'entrer en détail dans le fonctionnement du cil, il est important de rappeler sa structure et sa place dans la cellule. Physiquement, le cil est une extension de la membrane plasmique soutenue par un cytosquelette de microtubules, l'axonème (Figure 2-12). Bien que son milieu interne ne soit pas séparé du reste de la cellule par une membrane, il possède une composition moléculaire différente du reste de celle-ci grâce à sa structure compartimentalisée et ses mécanismes de transport moléculaire dont on fera ici une description rapide.

Le corps basal. Situé à la base du cil, le corps basal dérive du centriole mère, le centre organisateur des microtubules de la cellule, et est structuré autour de 9 triplets de microtubules en disposition circulaire. Il est associé à deux structures, appelées appendices distaux et sous-distaux, qui sont impliqués dans la fixation du corps basal à la membrane plasmique (Huang et al., 2017). La fixation du corps basal au niveau de la membrane est la première étape de la génération du cil, d'où l'axonème s'étend.

L'axonème. L'axonème est, chez la plupart des eucaryotes, composés de 9 doublets de microtubules ainsi que d'une paire de microtubules centraux. On retrouve sur l'axonème les bras de dynéine internes et externes, ainsi que les rayons radiaux, dont l'action est responsable du mouvement du cil. Certains types de cils, incapables de mouvements, sont dépourvus de la paire centrale et des bras de dynéine (cil primaire sur la Figure 2-12), ces différentes catégories de cil seront détaillées dans la section suivante.

La zone de transition : La zone de transition désigne l'ensemble des structures moléculaires situées à la base du cil qui permettent la compartimentalisation du cil et son intégrité moléculaire vis-vis de la cellule, en agissant comme une 'porte ciliaire' (Reiter et al., 2012). La zone de transition abrite des structures en forme de Y faisant la jonction entre la membrane plasmique et l'axonème dont la fonction supposée est de créer une barrière de diffusion avec le reste de la cellule. Les jonctions en Y organisent le collier ciliaire, des structures circulaires fixées à la membrane, autour de la zone transition. On ne connait pas en détail la composition moléculaire de ces structures mais les complexes MKS et NPHP essentiels au fonctionnement du cil constituent de bons candidats.

La machinerie de transport intra-flagellaire. Aucune protéine n'étant produite dans le cil et la diffusion avec le reste de la cellule rendue impossible par la zone de transition, l'apport en protéine sur toute la longueur de l'organelle est assuré par une machinerie de transport composée de deux grands complexes moléculaires : les complexes IFT (*Intra-Flagellar Transport*) et BBSome (*Bardet-Biedl Syndrome*). Le complexe IFT est composé de deux sous-complexes (Lechtreck, 2015). L'IFT-B est responsable du transport antérograde du cargo le long de l'axonème en s'associant à la kinésine-2, un moteur moléculaire ; réciproquement IFT-A s'associe à la dynéine et est responsable du transport rétrograde des protéines. Le BBSome se déplace de concert avec l'IFT le long du cil, et est considéré comme un 'adaptateur de cargo', permettant de lier des protéines supplémentaires aux complexes d'IFT. La machinerie de transport intra-flagellaire est nécessaire à la genèse et à la maintenance du cil car elle permet

l'acheminement des protéines constitutives de l'axonème, les tubulines à l'extrémité du cil où la réaction de polymérisation des microtubules a lieu. En plus de ce rôle essentiel, elle participe également au bon fonctionnement du cil en y acheminant d'autres protéines notamment les récepteurs membranaires.

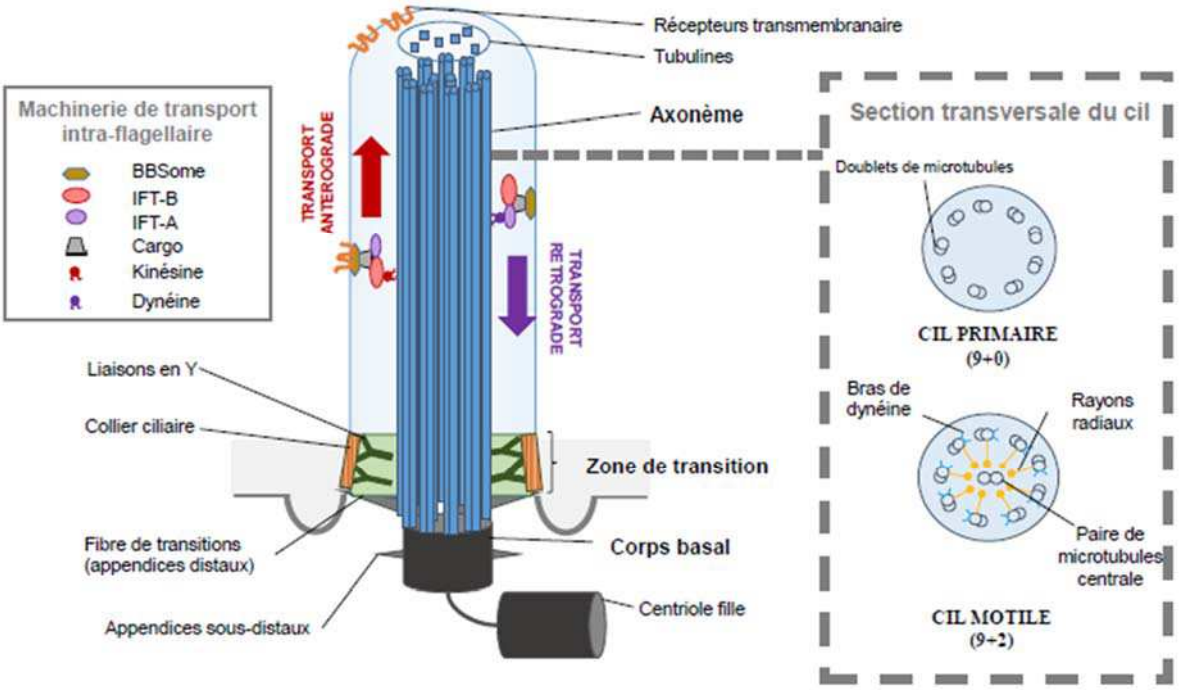


Figure 2-12 **Organisation du cil eucaryote**. La section transversale indique les différences entre les deux catégories de cils.

2.5.3 Les fonctions du cil

Grâce à sa position unique comme extension de la cellule, le cil possède deux grandes catégories de fonctions : le déplacement et la réception de signaux externes à la cellule. Si ces fonctions sont associées chez les eucaryotes unicellulaires dont les cils sont pourvus de mouvement, chez l'homme et les métazoaires de manière générale, elles peuvent être dissociées. De fait, on classe traditionnellement les cils en deux catégories sur la base de leurs caractéristiques structurales et fonctionnelles.

Les cils motiles ont pour fonction principale le mouvement de la cellule ou de fluides à la surface des cellules. Leur particularité structurale est un axonème possédant une paire centrale de microtubules et 9 doublets externes (configuration 9+2, Figure 2-12). Le mouvement est permis par des structures spécifiques du cil motile : les bras de dynéine et les rayons radiaux. Chez l'homme, on retrouve un cil motile unique sous la dénomination de flagelle au niveau des spermatozoïdes et plusieurs cils sur les cellules épithéliales ciliées du tractus respiratoire, des conduits génitaux féminins, et les cellules de l'épendyme (Satir et Christensen, 2007). Ces épithéliums multiciliés sont nécessaires aux déplacements de mucus et de fluides.

Les cils primaires, à l'inverse des cils motiles, reposent sur un axonème dépourvu de microtubules centraux (configuration 9+0) et sont généralement dépourvus de mouvement. On les retrouve au niveau de la plupart des cellules de vertébrés lorsqu'elles sont en phase G0. Leur rôle est essentiellement sensoriel, et ils sont capables selon les cellules de traiter des signaux mécaniques, chimiques ou encore photonique (Ke et Yang, 2014). D'une façon générale, le cil primaire est essentiel à la régulation du cycle cellulaire et dans la régulation de certaines voies de développement tels que *Sonic Hedgehog* et Wnt, mais il tient également un rôle spécifique dans certains organes : les tubules rénaux où il agit comme méchanosenseur de flux et régule la prolifération cellulaire et la rétine où un cil modifié permet de former les photorécepteurs (May-Simera et al., 2017).

La classification du cil en deux grandes catégories peut être remise en cause et il existe des exceptions (Takeda et Narita, 2012). Par exemple, le cil nodal embryonnaire, impliqué dans la mise en place de la symétrie droite-gauche, est doué de motilité malgré une structure axonémale en 9+0. Cependant, ces deux catégories fonctionnelles englobent la plupart des cas et permettent de mieux comprendre les différentes classes de ciliopathies, les pathologies associées au cil.

2.5.4 Les ciliopathies : des pathologies aux phénotypes complexes

La variété des fonctions cellulaires dans lesquelles les cils sont impliqués entraîne une quantité tout aussi pléthorique de syndromes en cas de dysfonctionnement du cil. Les ciliopathies couvrent une dizaine de maladies dont les résultats phénotypiques sont complexes. On distingue deux grandes classes de ciliopathies, liées à chacune des fonctions prédominantes du cil (Figure 2-13).

Les ciliopathies affectant le cil motile sont les dyskinésies ciliaires primaires (*primary ciliary dyskinesia* ou PCD). Elles sont principalement dues à des atteintes des bras de dynéine internes ou externes, des paires centrales de microtubules ou des complexes de régulation de la dynéine. Les symptômes de ces pathologies sont directement liés à la motilité et comptent des atteintes des voies respiratoires, l'infertilité masculine (atteinte du spermatozoïde) et des anomalies développementales : le *situs inversus* et l'hydrocéphalie. Les PCD (Praveen et al., 2015) forment une classe à part des autres ciliopathies, aussi bien au niveau du type de symptômes (Figure 2-13) qu'à leur diversité, moindre en comparaison des ciliopathies 'sensorielles'.

Les ciliopathies atteignant les mécanismes de biogenèses et les mécanismes sensoriels du cil, englobent plusieurs pathologies marquées par des phénotypes distincts bien qu'ils se recoupent entre différentes classes. Certaines sont associées aux dysfonctions d'un organe particulier, comme la néphronophtisie (NPHP) marquée par la formation de kystes au niveau du rein ou l'amaurose congénitale de Leber (LCA) et ses atteintes de la rétine. D'autres, dues à l'altération des fonctions essentielles du cil, sont pléiotropiques et se manifestent par des symptômes affectant plusieurs organes : des atteintes du rein et de la rétine, mais également plusieurs types d'atteintes développementales telles que la polydactylie, et des phénotypes globaux tels que l'obésité. On compte parmi les ciliopathies pléiotropiques, les syndromes de Bardet-Biedl (BBS), de Joubert, de Jeune et de Meckel (MKS). Ces grandes classes de ciliopathies sont

associées à des atteintes de gènes de la zone de transition et du transport intraflagellaire, mais les mécanismes exacts menant à l'émergence de ces phénotypes sont encore peu connus, et un même gène peut être impliqué dans différentes ciliopathies en fonction des mutations qui l'affectent et du contexte génétique (Reiter et Leroux, 2017).

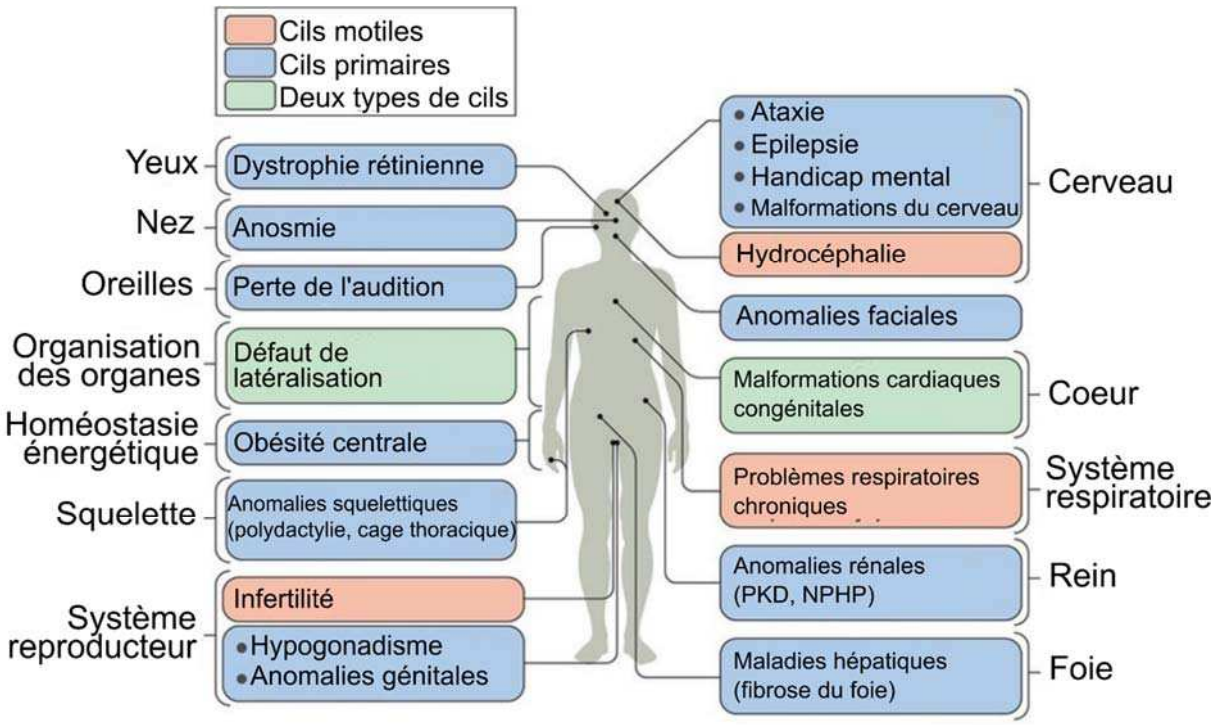


Figure 2-13 **Diversité des symptômes associés aux ciliopathies**. Les différents phénotypes impliquent de nombreux organes ou processus biologiques et sont différents selon que la ciliopathie touche le cil motile ou le cil primaire, sensoriel. Figure tirée de (Reiter et Leroux, 2017)

2.5.5 Problématiques omique et génomique comparative

La complexité de ces pathologies, probablement due à l'ubiquité du cil et à ses diverses fonctions, nécessitent une meilleure compréhension des relations génotype-phénotype qui sont en jeu ici. Cet objectif est au centre des efforts de la communauté scientifique. Il s'agit d'identifier les gènes impliqués dans le cil, de comprendre comment les protéines ciliaires s'organisent dans la cellule et d'identifier comment ces modules interagissent avec le reste des processus cellulaires pour provoquer les symptômes observés. Pour cela, des exploitations de technologies à haut débit de plus en plus ambitieuses sont réalisées comme le montre la publication récente d'un réseau d'interactions protéines-protéines dédié, basée sur la caractérisation par protéomique des intéractants de 217 protéines ciliaires (Boldt et al., 2016).

La tendance est aussi à l'exploitation simultanée des différentes données omiques déjà générées, à travers des ressources visant à agréger l'ensemble des expériences pertinentes pour l'analyse du cil : ciliomeDb (Inglis et al., 2006) et CilDb (Arnaiz et al., 2009, 2014). Parallèlement, des outils ont été développés pour faciliter l'analyse des données. Le consortium

Syscilia a mis en place une liste *Gold Standard* des gènes liés au cil (van Dam et al., 2013) annotés par des experts. Ce standard peut aider à calibrer les méthodes d'inférence *in silico* en mettant à disposition un ensemble de référence. Plus récemment, le même consortium associé au consortium *Gene Ontology* a organisé une refonte des annotations GO consacrées à l'organelle (Roncaglia et al., 2017), plus en phase avec notre compréhension actuelle du cil. Cette ontologie mise à jour facilitera d'autant plus l'intégration des données.

Le cil est ainsi représentatif des problématiques associées aux relations génotype-phénotype, d'une part à cause du rapport encore mal compris entre les atteintes des gènes ciliaires et les syndromes des ciliopathies, d'autre part car son étude bénéficie également d'une couverture par des données omiques diverses et bénéficierait de l'intégration de données diverses. En particulier, les données d'évolution apportent une perspective inédite, notamment grâce à l'histoire évolutive unique de l'organelle. Il s'agit donc d'un cas d'application particulièrement adapté pour le développement d'un marqueur évolutif.

3 Inférer et représenter l'orthologie : méthodes et défis

Le profilage phylogénétique, comme la génomique comparative au sens large repose sur une bonne définition des relations d'orthologie et de paralogie entre gènes. Comme la connaissance de l'enchaînement exact des événements de spéciation et de duplication ancestraux nous est inaccessible, ceux-ci sont inférés à partir de divers éléments à notre disposition, selon des méthodes variées

Dans ce chapitre, j'aborde les défis afférents à ces prédictions et détaille les méthodes existantes, en insistant sur leur diversité. Dans un second temps, je précise le besoin d'en standardiser les résultats pour évaluer objectivement les caractéristiques de chacune des méthodes et orienter les choix des méthodes adaptées, mais également faciliter les combinaisons de ces données entre elles ou avec des données d'autre nature. Je termine par une description des ressources en ligne mettant à disposition ces prédictions en soulignant l'importance de la représentation, de la contextualisation ainsi que des outils d'exploration des données qui permettent de faciliter la compréhension des données évolutives.

3.1 De la séquence à l'orthologie

Comme nous l'avons vu au chapitre précédent, le concept d'homologie est la clé de voute de toute la génomique comparative. Les prédictions des relations d'orthologie sont donc un préalable à la plupart de ces analyses mais elles ne sont pas triviales, pour la bonne raison qu'il n'y a pas de trace directe des événements de spéciation ou duplication passés. De fait, on se base généralement sur des traces indirectes pour retrouver l'homologie, en premier lieu la similarité de séquences. Principalement, considérant que deux gènes orthologues descendent d'un ancêtre commun, on considère que leurs séquences présentent des similarités, héritées de leur ancêtre. On fait donc l'hypothèse qu'un gène orthologue *Ga* d'une espèce A aura plus de similarité avec son orthologue *Gb* de l'espèce B qu'avec toutes les autres séquences de l'espèce B.

Cette hypothèse est le socle des méthodes de prédiction d'orthologie basées sur la similarité, dont on compte aujourd'hui presqu'une cinquantaine (voir Tableau 3-1 en fin de chapitre). On peut les séparer en grandes catégories : les méthodes basées sur les graphes, les méthodes basées sur les arbres, les méthodes hybrides et les méthodes intégratives, dont je détaille ici la diversité méthodologique.

3.1.1 Les méthodes basées sur les graphes

3.1.1.1 Principes généraux

Comme leur nom l'indique, les méthodes basées sur les graphes traitent les relations d'orthologie comme des graphes dont les nœuds sont les séquences et les arrêtes les relations proprement dites. Les relations sont avant tout construites à partir des comparaisons de séquences deux-à-deux à l'échelle du génome. Ces comparaisons sont implémentées par des outils de recherche heuristiques de similarité de séquences tels que BLAST.

L'approche la plus naïve des méthodes basées sur les graphes, dite du « meilleur hit » (*Best hit*, BH), consiste à considérer comme orthologue d'un gène d'intérêt *Ga,* le gène *Gb* d'une autre espèce B ayant la similarité la plus importante avec elle, à condition que cette similarité soit significative. C'est une méthode rapide, mais qui ne considère pas la nature symétrique d'une relation d'orthologie, ce qui peut entraîner des erreurs dans les cas de duplications et de pertes différenciées aux cours de l'évolution (Figure 3-1).

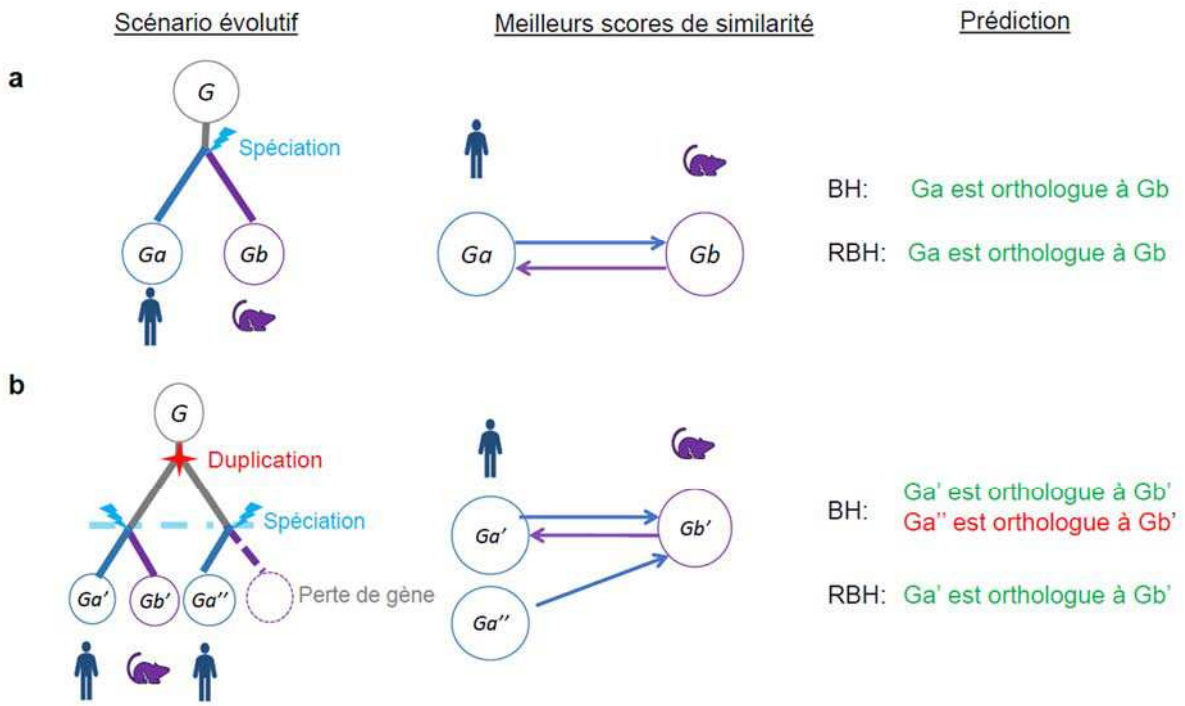


Figure 3-1 **Inférences de relations d'orthologie selon deux scénarii évolutifs**. a. Cas 'simple' avec un événement de spéciation seulement. Les deux gènes sont les plus similaires l'un à l'autre. Les méthodes du meilleur hit (BH) et du meilleur hit réciproque arrivent à une prédiction correcte. b. Duplication précédent la spéciation et la perte du gène Gb'' chez la souris. Ga'' a donc pour plus proche homologue non pas son orthologue mais son outparalogue Gb'. La méthode du meilleur hit produit un faux positif dans ce scénario alors que le meilleur hit réciproque n'est pas impacté.

La méthode du « meilleur hit réciproque » (*reciprocal best hit*, RBH) (Overbeek et al., 1999) permet de pallier ce défaut, en n'admettant une relation d'orthologie que si la plus forte similarité entre séquences est réciproque. En reprenant notre exemple (Figure 3-1), cela signifie

que *Gb* est le gène de l'espèce B avec la plus forte similarité à *Ga,* et que *Ga* est le gène de l'espèce A avec la plus forte similarité à *Gb*.

La méthode du RBH est la base conceptuelle sur laquelle reposent la plupart des algorithmes basés sur les graphes utilisés aujourd'hui bien que chacun apporte des modifications et des étapes supplémentaires pour mieux couvrir la réalité évolutive derrière les relations d'orthologie.

3.1.1.2 Intégration des inparalogues

Lorsque l'on prend en compte les événements de duplication, les relations d'orthologie peuvent être de trois types (Figure 3-2) :

1. Un-à-un : un gène de l'espèce A est orthologue à un gène de l'espèce B.
2. Un-à-plusieurs : un (ou plusieurs) évènement(s) de duplication a eu lieu dans le lignage d'une des espèces après l'événement de spéciation. Dans ce cas, un gène de l'espèce A est orthologue à plusieurs gènes inparalogues de l'espèce B.
3. Plusieurs-à-plusieurs : des événements de duplication ont eu lieu indépendamment dans le lignage de chaque espèce après l'événement de spéciation. Dans ce cas, plusieurs gènes inparalogues de l'espèce A sont co-orthologues à plusieurs gènes inparalogues de l'espèce B.

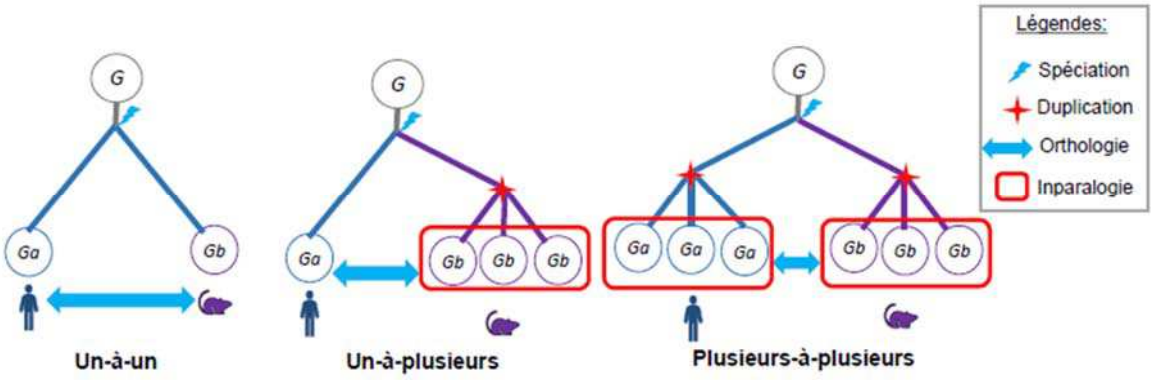


Figure 3-2 **Cardinalité des relations d'orthologie**. Différents types de relations d'orthologie existant entre descendants d'un gène G en fonction des événements de duplication postérieurs à la spéciation.

Le RBH ne permet de prédire clairement que le premier type de relations et peut donc mener, entre autres, à l'exclusion de vrais orthologues. De nombreuses méthodologies, incluant les algorithmes les plus populaires, étendent la stratégie du RBH pour permettre de représenter tous les types de relations et se prémunir de ses limites.

Les relations d'inparalogie peuvent être inférées à partir des comparaisons d'espèces deux-à-deux, en identifiant les séquences de l'espèce A qui sont plus similaires entre elles qu'elles ne sont avec une séquence de l'espèce B. Sur cette base, Inparanoïd (O'Brien et al., 2005) et OrthoInspector (Linard et al., 2011) effectuent la prédiction d'inparalogues à la phase de comparaison deux-à-deux des séquences, mais procèdent dans un ordre différent. Inparanoïd

identifie un premier lieu les RBH entre les protéines de deux génomes et considère comme inparalogues les séquences du même organisme ayant un meilleur score que le RBH lui-même. OrthoInspector commence par construire les groupes d'inparalogues en identifiant les protéines d'un même organisme plus similaires entre elles qu'avec toutes protéines d'un autre organisme et recherche ensuite les RBH entre groupes d'inparalogues. Ainsi, des relations d'orthologie entre groupes d'inparalogues peuvent être identifiées sans nécessairement qu'il existe un RBH entre les protéines composant ces groupes. Pour les deux méthodes, les trois types de relations d'orthologie (un-à-un, un-à-plusieurs, plusieurs-à-plusieurs) entre deux organismes sont directement déterminés.

Une autre possibilité pour intégrer les inparalogues est d'utiliser un intervalle de confiance lors de la prédiction des meilleurs hits réciproques pour admettre plusieurs 'meilleurs hits' si ceux-ci ont un score de similarité proche du véritable meilleur hit réciproque, c'est notamment la stratégie utilisée par OMA (Roth et al., 2008) dans l'étape de prédiction des relations deux-à-deux.

3.1.1.3 Améliorer les mesures de similarité de séquences

Comme on l'a vu, les méthodes basées sur les graphes, inspirées du RBH reposent sur l'évaluation de leur similarité réciproque. De façon générale, on ne considère cette similarité suffisante pour inférer une relation que si elle dépasse un certain seuil, en dessous duquel elle pourrait ne pas être due à une relation d'orthologie. En plus d'être utilisés pour les prédictions proprement dites, les scores de similarité peuvent être utilisés pour évaluer la pertinence de relations d'orthologie ou d'inparalogie (ex. *Inparanoïd*) ou pondérer ces relations.

Pour l'ensemble de ces tâches, les indicateurs issus du BLAST, bit score et *e-value*, sont souvent utilisés, mais il est important de noter qu'ils ne sont pas des mesures exactes de similarité et dépendent notamment de la longueur des séquences alignées. Une méthode récente, OrthologFinder (Horiike et al., 2016), utilise un score normalisé en fonction de la taille des séquences et de la distance phylogénétique pour réaliser ses prédictions.

Alternativement, un moyen de prendre en compte la similarité entre séquence est de la calculer avec des méthodes de maximum de vraisemblance utilisant des modèles d'évolution. Ces scores, calculés sur la base d'alignements globaux (rsd (Wall et al., 2003)) ou locaux (OMA (Roth et al., 2008)), ne sont pas soumis aux mêmes biais que les scores de BLAST et peuvent être utilisés sur le même principe que le RBH : on s'intéresse encore aux séquences les plus proches.

3.1.1.4 Les orthogroupes

Une fois les relations d'orthologie prédites entre deux génomes, on peut vouloir s'intéresser à ces relations dans le contexte de l'ensemble des génomes analysés. Une part considérable des méthodes basées sur les graphes étendent donc les relations d'orthologie deux-à-deux à la création de groupes d'orthologues, que l'on appellera par la suite « orthogroupes ».

Conceptuellement, les gènes homologues retrouvés dans un ensemble d'espèces sont les descendants d'un gène ancestral présent chez le dernier ancêtre commun de ces espèces (à l'exception des cas de transferts horizontaux). Les méthodes de création d'orthogroupes cherchent à regrouper l'ensemble de ces descendants. Plusieurs stratégies existent pour constituer ces orthogroupes à partir des relations deux-à-deux.

Une première stratégie consiste à construire les groupes par transitivité, en ajoutant successivement au groupe d'orthologues prédits entre deux espèces, l'ensemble des gènes des autres espèces ayant une relation d'orthologie avec ceux déjà présents. Dans le cas idéal, si l'ensemble des relations est bien défini, procéder de cette façon mène à la constitution du groupe complet des descendants du gène ancestral (DeLuca et al., 2012). Cette stratégie est cependant très sensible aux erreurs de prédictions, il suffit d'une relation erronée entre gènes appartenant à deux orthogroupes pour amener à l'agglomération de ceux-ci.

Pour éviter cela, l'algorithme d'OMA (Roth et al., 2008) définit un orthogroupe comme une entité où la majorité des membres sont interconnectés, les cliques. Une définition aussi stricte des orthogroupes conduit à une forte spécificité à défaut de sensibilité : le critère d'interconnexion renforce la cohérence des groupes ainsi produit, rendant peu probable une inclusion erronée, mais impliquant l'exclusion possible d'orthologues réels, dont la séquence divergente ne permettrait pas de valider une relation avec l'ensemble du groupe. Le critère d'interconnexion signifie également que les orthogroupes ainsi définis ne comprennent que des relations d'orthologie un-à-un et n'incluent pas d'inparalogues, deux paralogues ne pouvant être reliés par une relation d'orthologie. Ils ne comprennent donc pas tous les gènes orthologues entre eux, mais uniquement ceux dont l'événement de spéciation est antérieur à l'événement de duplication le plus récent.

Alternativement, Multiparanoïd (Alexeyenko et al., 2006) définit ces orthogroupes par une méthodologie moins stricte : il applique un algorithme de *trimming* de cluster au graphe établis par transitivité. Cette étape vise à exclure uniquement les relations aberrantes entre orthogroupes, en prenant notamment en compte les scores de confiance des relations considérées.

Une autre stratégie basée sur la transitivité, consiste à utiliser non pas les paires, mais des triangles de relations entre gènes de trois espèces. Ici, les triangles sont constitués dans le cas où les gènes des trois espèces sont chacun les meilleurs hits les uns des autres et forment un groupe minimal. Chaque groupe minimal est ensuite étendus en le fusionnant à d'autres triangles partageant une arrête commune. Cette méthode est ainsi moins sensible aux erreurs de prédiction car il est toujours nécessaire pour l'extension d'un groupe qu'il y ait toujours au moins deux orthologues communs à chaque membre du groupe. C'est sur une telle stratégie que repose le programme COG , (Tatusov et al., 1997), le premier à implémenter une méthode de création d'orthogroupes. Une stratégie similaire est reprise dans les programmes EggNog (Jensen et al., 2008) et OrthoDb (Kriventseva et al., 2008).

Plutôt que de procéder par transitivité, une dernière stratégie d'orthogroupe utilise des algorithmes dédiés au *clustering* dans les graphes. Cette stratégie a été utilisée pour la première fois par OrthoMCL (Chen et al., 2006) et a été reprise par les programmes plus récents OrthAguoge et ProteinOrtho (Ekseth et al., 2014; Lechner et al., 2011). Brièvement, les algorithmes de *clustering* identifient les groupes de gènes fortement connectés dans un graphe d'orthologie dont les arrêtes sont pondérées en fonction de la similarité de séquence. Par conséquent, les groupes constitués incluent souvent les orthologues et les paralogues issus de duplications récentes. Les paramètres des programmes peuvent être modifiés pour adapter la granularité des groupes, et notamment retrouver des paralogues issues de duplications plus anciennes.

Les groupes issus de ces différentes méthodes, à l'exception des plus strictes, incluent à la fois des séquences ayant des relations d'orthologie et de paralogie. Plusieurs espèces étant représentées dans les groupes, il peut devenir difficile de distinguer les inparalogues des outparalogues par rapport à une espèce donnée et donc, d'ordonner dans le temps les événements de duplication et de spéciation. Pour contourner le problème, EggNog (Jensen et al., 2008), Kegg Orthologs (Nakaya et al., 2013) et OrthoDB (Kriventseva et al., 2008) utilisent un arbre des espèces, qui apporte l'information sur les événements de spéciation pour diriger la création d'orthogroupes. Les groupes sont construits par rapport à plusieurs clades et on peut donc déterminer la hiérarchie des événements en comparant les groupes des taxons récents à ceux des plus ancestraux.

Faire entrer la taxonomie pour prédire les relations d'orthologie est le principe utilisé par une autre catégorie de méthodes, celles basées sur les arbres.

3.1.2 Méthodes basées sur les arbres

Les méthodes d'inférence basées sur les arbres reposent sur un principe simple, reconstituer l'histoire évolutive d'un gène en réconciliant cet arbre avec l'arbre des espèces. Ces méthodes se déroulent en plusieurs étapes : l'ensemble des homologues est d'abord identifié en fonction de leur similarité de séquences. A partir de ces homologues, on construit un alignement multiple et l'on génère un arbre phylogénétique des gènes. Finalement, les événements de duplication, de spéciation et de perte sont identifiés dans l'arbre phylogénétique, en le comparant avec un arbre des espèces. Il s'agit de l'étape de réconciliation. Les relations de paralogie et d'orthologie sont triviales à définir, une fois les spéciations et les duplications identifiées (Figure 3-3).

Les méthodes de réconciliation reposent généralement sur un critère de parcimonie et cherchent donc, à placer les événements de duplication de façon à minimiser leur nombre ainsi que le nombre de pertes nécessaires à expliquer la topologie de l'arbre (Zmasek et Eddy, 2002). Un des risques des méthodes basées sur les arbres est l'incertitude liée aux arbres de gènes et même parfois, à l'arbre des espèces, ce qui peut être à l'origine d'erreurs lors de la réconciliation. Pour prendre en compte ces incertitudes dans l'arbre des espèces, les approches RIO (Zmasek et Eddy, 2002) et OrthoStrapper (Storm et Sonnhammer, 2002) s'appuient sur des méthodes d'échantillonnage pour évaluer la confiance dans les relations d'orthologie. Le programme

RAP, utilisé dans les bases de données Hogenome, Homologene et Hoverlens (Penel et al., 2009), prend en compte les branches peu résolues dans les deux arbres lors de l'étape de réconciliation.

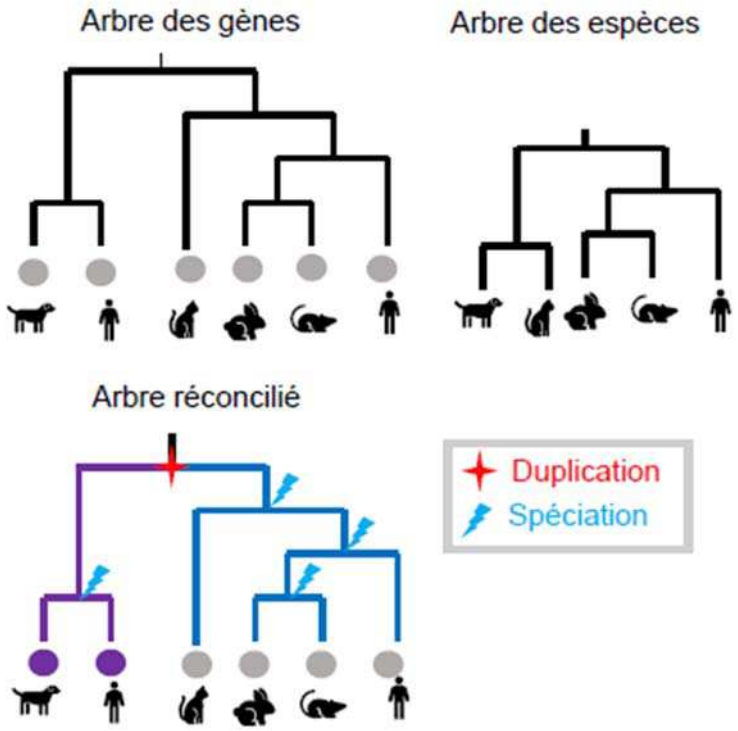


Figure 3-3 **Réconciliation de l'arbre des gènes et de l'arbre des espèces**. La réconciliation consiste à identifier les différents événements de duplication et de spéciation dans l'arbre des gènes. Dans ce cas simple, des gènes humains se retrouvent dans les branches filles du nœud ancestral mettant en évidence un évènement de duplication. L'ordre des spéciations correspond à l'arbre des espèces, les arbres sont donc réconciliés.

Sur le principe, les méthodes basées sur les arbres sont à même de résoudre des scénarios d'évolution plus complexes que celles basées sur les graphes. L'étape de réconciliation peut cependant être coûteuse en ressources informatiques et peu adaptée pour un grand nombre d'espèces. De plus, elles nécessitent une taxonomie suffisamment certaine pour pouvoir y baser ses prédictions, ce qui n'est pas toujours garanti.

Pour éviter ces problèmes tout en s'appuyant sur la taxonomie, une autre approche s'appuie sur la comparaison, à chaque nœud, des espèces présentes dans ses branches filles (van der Heijden et al., 2007). Si l'intersection entre les espèces représentées dans chaque branche est nulle, on attribue au nœud un événement de spéciation, si elle ne l'est pas on y attribue un événement de duplication. Cette méthode ne nécessite pas directement d'arbre des espèces, puisque les opérations à réaliser consistent simplement à une comparaison des espèces présentes. Une implémentation de cette méthode est notamment utilisée dans la base de données PhylomeDB (Huerta-Cepas et al., 2007). Une variante de la stratégie est utilisée dans les méthodes dites de *tree splitting*, implémentées pour la base de données MBGD (Uchiyama, 2003), et les programmes PhyTreePruner (Kocot et al., 2013) et Ortholog-Finder (Horiike et al., 2016). Pour ces programmes, le but est d'identifier les sous-arbres représentant des groupes

d'orthologie sans duplication, c'est-à-dire regroupant une seule séquence par espèce. Dans ce cas, les nœuds identifiés comme des duplications (reliant plusieurs espèces) sont 'coupés', de façon à obtenir uniquement les groupes d'orthologues.

Comme ces méthodes ne reposent sur aucun arbre des espèces, leur exécution est moins coûteuse et n'est pas affectée par les incertitudes dans la taxonomie, pour autant elles deviennent aussi moins efficaces pour résoudre les scénarios évolutifs complexes.

3.1.3 Méthodes hybrides

Les méthodes basées sur les arbres extraient plus d'informations des séquences, et en prenant en compte la taxonomie, permettent en principe de situer plus efficacement les événements de duplication et de perte, ce qui améliore les prédictions. En revanche, les ressources de calcul nécessaires pour ce genre d'approches sont plus importantes que les méthodes basées sur les graphes. Il y a donc un équilibre à trouver entre spécificité des prédictions et temps de calcul. Les méthodes hybrides utilisent des attributs des deux catégories de méthodes, de façon à garder les avantages des méthodes basées sur les arbres tout en nécessitant moins de ressources.

Dans le principe, les méthodes hybrides sont des extensions de la création de groupes successifs en remontant dans la taxonomie, décrites en fin de section consacrée aux méthodes basées sur les graphes. La différence ici est que les relations d'orthologie ne sont pas reconstruites *de novo* à chaque nœud de l'arbre taxonomique, mais utilisent les orthogroupes des nœuds enfants.
Une méthode de ce type est utilisée pour générer les *Hierarchical Orthologous Group*s (HOG) (Altenhoff et al., 2013) disponibles sur la base de données OMA. A chaque nœud de l'arbre taxonomique, les relations d'orthologie sont utilisées pour fusionner les groupes des niveaux inférieurs ayant des relations entre eux. La méthode HyPPO (Lafond et al., 2018) procède d'une façon similaire et donne également la possibilité de générer un arbre des espèces à partir de l'ensemble des groupes d'orthologues si l'arbre n'est pas connu à l'avance.

Une variante de cette méthode d'inférence successive consiste à recalculer les relations deux-à-deux à chaque nœud de l'arbre, en construisant une séquence consensus obtenue à partir de des séquences orthologues identifiées dans les branches précédentes. Cette séquence consensus est utilisée à chaque étape pour estimer les relations d'orthologie par RBH avec d'autres séquences consensus ou avec des séquences d'autres espèces, en fonction de la position dans l'arbre. Cette méthode itérative est implémentée dans Hieranoïd (Schreiber et Sonnhammer, 2013) et PHOG (Merkeev et al., 2006) et réduit considérablement le nombre de comparaisons deux-à-deux à réaliser.

Une autre sous-catégorie de méthodes hybrides utilise des méthodes proches de la réconciliation d'arbres, mais à partir d'orthogroupes basés sur les arbres et sans nécessiter la construction d'un arbre phylogénétique. On compte deux exemples de cette stratégie COCO-CL (Jothi et al., 2006) et MultiMSOAR (Shi et al., 2011). COCO-CL part des groupes d'orthologues produits par des méthodes basées sur les graphes et établit un *clustering* hiérarchique des séquences en utilisant les distances issues d'un alignement multiple. A chaque étape du *clustering*

hiérarchique, le recouvrement des espèces représentées dans les deux groupes de gènes regroupés à cette étape est analysé. Si le recouvrement dépasse un certain seuil, on sépare les orthogroupes à cette étape.

Dans le cas de MultiMSOAR, les duplications et pertes sont inférées à partir des orthogroupes en se basant sur un arbre des espèces, selon un critère de parcimonie. Dans les deux cas, les stratégies de reconstruction utilisées sont des équivalents de celles utilisées dans les méthodes basées sur les arbres, mais procèdent d'abord à l'identification d'orthogroupes.

3.1.4 Les méthodes intégratives

Ce tour d'horizon des méthodes de prédiction d'orthologie se voulait représentatif de la diversité des méthodes existantes pour la prédiction des relations d'orthologie. Bien qu'il serait tentant d'affirmer qu'une méthode ou une stratégie est supérieure aux autres, leur efficacité relative dépend des questions posées. Une dernière classe de méthodes d'inférence part de ce constat pour proposer des méthodes agglomératives, ou intégratives, prenant en compte les prédictions obtenues par différentes méthodes.

Dans leurs formes les plus simples, les prédictions par intégration de méthodes sont disponibles dans des bases de données telle que YOGY (Penkett et al., 2006), HCOP (Eyre et al., 2007), DIOPT (Hu et al., 2011) ou OGO (Miñarro-Gimenez et al., 2009), dans lesquelles les prédictions obtenues par les différentes approches sont indiquées, ainsi que le nombre de méthodes qui concurrent à une prédiction. Dans DIOPT, un score est attribué aux relations en fonction du nombre de méthodes indépendantes arrivant à ce résultat et les relations prédites par peu de méthodes peuvent être filtrées à l'avance.

Combiner les méthodes de cette façon permet à la fois d'identifier plus facilement les faux positifs des méthodes individuelles, qui seraient retrouvées par aucune autre, tout en retrouvant les faux négatifs (prédits par d'autres méthodes). Le programme MARIO (Pereira et al., 2014) fonctionne sur ce principe et intègre un filtre supplémentaire : les relations d'orthologie prédites par plusieurs méthodes sont utilisées pour former un groupe d'orthologues, qui est ensuite utilisé pour construire un profil HMM des séquences. Celui-ci sert ensuite à évaluer les prédictions réalisées spécifiquement par chaque méthode individuelle et les intégrer s'ils correspondent au profil.

Dans un autre registre, la base de données MetaPhOrs (Pryszcz et al., 2011) n'intègre pas directement les prédictions, mais les arbres phylogénétiques réalisés par plusieurs méthodes de prédiction d'orthologie. Elle utilise ces arbres pour prédire des réactions d'orthologie et attribue un score en fonction du nombre de fois où une relation d'orthologie est retrouvée. De cette manière, elle agrège plusieurs prédictions tout en évitant de prendre en considération les résultats peu fiables dus, par exemple, à une mauvaise résolution dans l'arbre phylogénétique.

Lorsque l'on parle d'intégration de données, il est difficile de ne pas mentionner les méthodes d'apprentissage automatique, qui permettent de combiner les prédictions de plusieurs

méthodes. Le programme Whormole (Sutphin et al., 2016) fonctionne selon ce principe, en exploitant un « classifieur » basé sur des machines à vecteur de support (SVM). Le programme nécessite un jeu d'entrainement positif de relations d'orthologie validées et un jeu négatif de paires de gènes non orthologues. L'algorithme est entraîné sur ces jeux de référence pour attribuer un poids à chacune des méthodes de prédiction en fonction de leur performance. Ce poids est ensuite utilisé pour réaliser les prédictions sur un jeu de données complet et en extraire les relations d'orthologie fiables. Son intérêt est d'identifier les algorithmes les plus efficaces pour certaines situations, par exemple en fonction de la proximité des espèces considérées, et d'y accorder plus ou moins de crédit, tout en gardant l'avantage de combiner plusieurs prédictions en cas de mauvaises résolutions de certains cas particuliers. Il est à noter que Wormhole a été appliqué uniquement à l'identification des 'orthologues de moindre divergence', c'est-à-dire un sous-ensemble de relations d'orthologie ne considérant que les paires avec la plus forte similarité de séquences dans les relations un-à-plusieurs ou plusieurs-à-plusieurs. Cette approche est conceptuellement applicable à l'inférence d'orthologie au sens large.

Un enseignement qui ressort des méthodes intégratives est la nécessité d'un format commun pour faciliter le croisement des résultats des méthodes d'inférences d'orthologie. Un autre réside dans la nécessité d'évaluer l'efficacité, la sensibilité et la spécificité de chaque méthode en fonction des contextes biologiques dans lesquels elles sont appliquées, ce que réalise en quelque sorte WormHole.

Tableau 3-1 **Liste non exhaustive des méthodes d'inférence d'orthologie**. Les méthodes sont classées par catégorie et par date de première publication. Les références associées sont indiquées ainsi qu'un indicateur montrant l'existence de ressources construites en utilisant ces méthodes.

Méthode	Catégorie	Description	Ressource
RBH	Basée sur les graphes	(Overbeek et al., 1999)	✗
COG	Basée sur les graphes	(Tatusov et al., 1997)	✗
Inparanoid	Basée sur les graphes	(Remm et al., 2001)	✓
EGO	Basée sur les graphes	(Lee et al., 2002)	✗
OrthoMCL	Basée sur les graphes	(Li et al., 2003)	✓
RoundUp	Basée sur les graphes	(Wall et al., 2003)	✗
MultiParanoïd	Basée sur les graphes	(Alexeyenko et al., 2006)	✗
OMA	Basée sur les graphes	(Roth et al., 2008)	✓
EggNOG	Basée sur les graphes	(Jensen et al., 2008)	✓
OrthoDb	Basée sur les graphes	(Kriventseva et al., 2008)	✓
GOOD	Basée sur les graphes	(Ho et al., 2010)	✗
OrthoInspector	Basée sur les graphes	(Linard et al., 2011)	✓
Proteinortho	Basée sur les graphes	(Lechner et al., 2011)	✗
ReMark	Basée sur les graphes	(Kim et al., 2011)	✗
panOct	Basée sur les graphes	(Fouts et al., 2012)	✗
SPOCS	Basée sur les graphes	(Curtis et al., 2013)	✗
Morfeus	Basée sur les graphes	(Wagner et al., 2014)	✗
OrthAgogue	Basée sur les graphes	(Ekseth et al., 2014)	✗
OrthoFinder	Basée sur les graphes	(Emms et Kelly, 2015)	✗

porthoDom	Basée sur les graphes	(Bitard-Feildel et al., 2015)	✗
Orthograph	Basée sur les graphes	(Petersen et al., 2017)	✗
JustOrthologs	Basée sur les graphes	(Miller et al., 2018)	✗
SonicParanoid	Basée sur les graphes	(Cosentino et Iwasaki, 2018)	✗
RIO	Basée sur les arbres	(Zmasek et Eddy, 2002)	✗
Orthostrapper	Basée sur les arbres	(Storm et Sonnhammer, 2002)	✗
PHOG	Basée sur les arbres	(Merkeev et al., 2006)	✗
MBGD	Basée sur les arbres	(Uchiyama, 2003)	✓
RAP	Basée sur les arbres	(Dufayard et al., 2005)	✓
Coco-CL-COG	Basée sur les arbres	(Jothi et al., 2006)	✗
OrthologId	Basée sur les arbres	(Chiu et al., 2006)	✗
PhiGS	Basée sur les arbres	(Dehal et Boore, 2006)	✗
COG-LOFT	Basée sur les arbres	(van der Heijden et al., 2007)	✗
Phylome Db	Basée sur les arbres	(Huerta-Cepas et al., 2007)	✓
GreenPhylDb	Basée sur les arbres	(Rouard et al., 2011)	✓
Ensembl Compara	Basée sur les arbres	(Vilella et al., 2009)	✓
Panther	Basée sur les arbres	(Mi et al., 2010)	✓
Phylotreepruner	Basée sur les arbres	(Kocot et al., 2013)	✗
OrthoLugeDb	Hybride	(Fulton et al., 2006)	✓
TreeFam	Hybride	(Li et al., 2006)	✓
MultiMSOAR	Hybride	(Shi et al., 2011)	✗
HOG	Hybride	(Altenhoff et al., 2013)	✓
Hieranoid	Hybride	(Schreiber et Sonnhammer, 2013)	✓
OrthologFinder	Hybride	(Horiike et al., 2016)	✗
Orthonome	Hybride	(Rane et al., 2017)	✓
Hyppo	Hybride	(Lafond et al., 2018)	✗
YOGY	Intégrative	(Penkett et al., 2006)	✓
HCOP	Intégrative	(Eyre et al., 2007)	✓
OGO	Intégrative	(Miñarro-Gimenez et al., 2009)	✗
DIOPT	Intégrative	(Hu et al., 2011)	✓
MetaPhOrs	Intégrative	(Pryszcz et al., 2011)	✓
MARIO	Intégrative	(Pereira et al., 2014)	✓
WormHole	Intégrative	(Sutphin et al., 2016)	✓

3.2 Standardiser et évaluer les méthodes de prédiction

La standardisation des formats de prédiction d'orthologie, et la comparaison objective des résultats des méthodes dans différents contextes sont des points clé tant pour l'amélioration des méthodes de prédiction que pour choisir le programme adapté à une problématique donnée. Ces thématiques sont au cœur d'un effort mené par la communauté, matérialisé par le consortium *Quest For Orthologs* (Gabaldón et al., 2009).

3.2.1 *Quest For Orthologs*

L'initiative *Quest For Orthologs* (QFO) regroupe des membres de la communauté scientifique intéressés par les relations d'orthologie, que ce soit comme utilisateur, comme développeur ou hôte d'une base de données. L'objectif est de travailler conjointement à résoudre les problèmes associés à la prédiction d'orthologues. Depuis la première réunion en 2009, QFO s'est réuni à quatre reprises (Dessimoz et al., 2012; Forslund et al., 2018; Sonnhammer et al., 2014). Ces réunions ont touché à la fois aux questions méthodologiques de prédiction, discutées plus haut, ainsi qu'aux défis techniques de la standardisation, de l'évaluation des méthodes et de la gestion des flux de données. Elles ont donné lieu à différentes initiatives qui structurent cette section.

3.2.2 Standardiser les méthodologies

Le besoin de standardisation des méthodes de prédiction d'orthologie, nécessaire à leur comparaison et leur intégration, se retrouve à deux niveaux : les espèces utilisées pour réaliser les prédictions et le format dans lequel ces prédictions sont accessibles.

La définition d'un ensemble de protéomes de référence QFO correspond directement à ce premier niveau. Ici ; le terme protéome désigne l'ensemble des séquences protéiques traduites à partir des gènes d'un génome, et non le résultat d'expériences de protéomique. Ces protéomes comprennent des séquences d'organismes modèles, des espèces d'intérêt pour la recherche biomédicale ou agronomique, ou encore des espèces intéressantes d'un point de vue phylogénétique (Sonnhammer et al., 2014). Les séquences de protéines inclues dans ce jeu de données sont spécifiquement choisies pour être complètes et non redondantes (une seule séquence protéique par gène, les variants ne sont pas pris en compte) et de façon à représenter le plus largement possible la diversité du vivant. Ils incluent notamment des représentants des 3 Domaines du vivant (Bactéries, Archées et Eucaryotes). Ces protéomes sont mis à disposition des hôtes de ressources d'orthologie de façon à constituer un jeu commun de données pour lesquelles plusieurs prédictions sont disponibles et à en faciliter l'intégration. Les protéomes et la liste d'espèces disponibles sont mis à jour régulièrement et regroupent aujourd'hui 78 espèces.

La standardisation du format de données des programmes d'inférence d'orthologie est rendue difficile par leurs natures différentes (méthodes basées sur les graphes ou sur les arbres, avec ou sans scores). Le format OrthoXML (Schmitt et al., 2011) est prévu pour représenter dans un même format, les résultats de ces différents types de méthodes et permet ainsi la comparaison de leurs résultats. Le traitement des données dans ce format est rendu possible par plusieurs bibliothèques logicielles et le format est supporté par quelques-unes des ressources majeures d'orthologie, assurant leur interopérabilité. Une ontologie (Fernández-Breis et al., 2016) est également en cours de développement pour permettre une représentation standardisée des relations d'orthologie. Cette ontologie permet la représentation des données selon un standard du Web sémantique, RDF (*Ressource Descriptions Framework*), rendant plus efficace l'interopérabilité de ces bases de données entre elles, mais aussi avec d'autres types de bases de données implémentant ces technologies, dont UniProt/SwissProt (The UniProt Consortium,

2017), neXtProt (Gaudet et al., 2015) ou l'EBI. Ce type de standardisation est particulièrement important pour faciliter l'intégration de données variées.

La standardisation à la fois au niveau d'un jeu de données d'espèces commun et au niveau du format de représentation rend possible la comparaison des méthodes à grande échelle.

3.2.3 Evaluation standardisée des méthodes

Grâce à la mise en place des standards que nous venons de voir, QFO a récemment publié les résultats d'une étude à grand échelle (Altenhoff et al., 2016) comparant 15 méthodes d'orthologie selon une vingtaine de *benchmarks* (Figure 3-4). L'ensemble des *benchmarks* est conçu de façon à être utilisable par toute nouvelle méthode de prédiction, et les résultats tenus à jour avec l'ensemble des méthodes évaluées, sont disponibles sur http://orthology.benchmarkservice.org/.

La nécessité d'utiliser plusieurs *benchmarks* vient du fait que les relations d'orthologie réelles ne peuvent pas être connues avec certitude, étant donné qu'elles sont le résultat d'évènements anciens. Pour autant, on peut utiliser des approches indirectes. Les *benchmarks* choisis évaluent les méthodes selon trois catégories de critères :

1. *Concordance avec un arbre des espèces.* Les arbres de gènes prédits comme étant orthologues entre eux sont sélectionnés pour réaliser un arbre phylogénétique des espèces et on évalue sa correspondance avec un arbre taxonomique. La comparaison peut se faire pour l'ensemble des espèces où seulement des clades donnés.
2. *Concordance avec des arbres des espèces connues.* On compare les arbres reconstitués à partir de relations d'orthologie de certains gènes à un ensemble d'arbres phylogénétiques de références et on en évalue la correspondance.
3. *Equivalence de fonction.* La similarité fonctionnelle est mesurée pour les paires de gènes prédits comme étant orthologues (Figure 3-4). Les catégories fonctionnelles utilisées pour ces comparaisons sont les annotations GO avec preuves expérimentales et les numéro EC de la base de données ENZYME (Bairoch, 2000).

Les conclusions générales des comparaisons réalisées à partir de ces différents *benchmarks* sont que, s'il n'existe aucune méthode clairement supérieure aux autres, certaines ont de bonnes performances en fonction du contexte étudié. Par exemple, MetaPhors, la méthode intégrative basée sur les arbres est particulièrement efficace sur les benchmarks de concordance avec un arbre de gènes connus. Pour chacun des *benchmarks*, chaque méthode se situe sur un axe entre spécificité et sensibilité (Figure 3-4). Dans, l'ensemble les prédictions d'orthogroupes d'OMA se distinguent par une forte spécificité alors qu'à l'inverse la méthode basée sur les arbres utilisée dans Panther présente une forte sensibilité. Finalement, Inparanoid, Hieranoid et OrthoInspector démontrent un bon équilibre en spécificité et sensibilité sur l'ensemble des *benchmarks*.

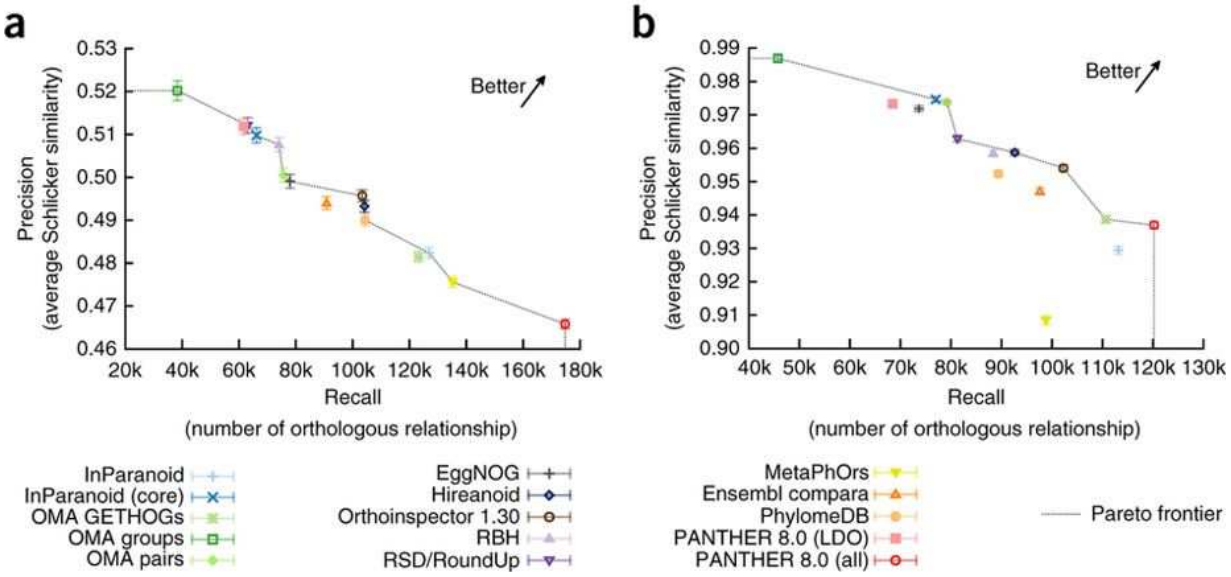


Figure 3-4 ***Benchmarking* des méthodes de prédiction d'orthologie**. Résultats obtenus par 15 méthodes sur un *benchmark* de similarité fonctionnelle entre orthologues. Deux types d'annotations sont utilisés : a- Annotations GO avec preuves expérimentales. b- Numéro EC. Les deux axes correspondent aux nombres d'erreurs dans les arbres (mesure de la spécificité) et la sensibilité (*recall*). Les différentes méthodes se répartissent à différents niveaux du rapport sensibilité/spécificité. Tirée de (Altenhoff et al., 2016).

3.2.4 Temps d'exécution à l'ère des *big data*

En plus de leur efficacité relative à réaliser des inférences d'orthologie correcte, le choix de la méthode d'orthologie à utiliser dépend aussi du temps de calcul nécessaire pour traiter un ensemble de données. Cette problématique est plus présente à l'ère des *big data* puisque le nombre de relations d'orthologie croît de manière quadratique avec le nombre de génomes considérés, à plus forte raison si ces génomes regroupent de nombreux gènes (Sonnhammer et al., 2014). Le temps de calcul nécessaire pour établir ces relations peut devenir limitant, plus encore pour les méthodes basées sur les arbres nécessitant une reconstruction de la phylogénie. Des stratégies ont vu le jour pour tenter de le réduire.

Un goulot d'étranglement commun à la plupart des méthodes d'orthologie basées sur les graphes concerne les comparaisons de gènes tous-contre-tous qui est souvent l'étape préalable à l'inférence d'orthologie. Des stratégies existent pour réduire cette étape, en réduisant à l'avance le nombre de comparaison à effectuer. Dans Hieranoïd, le nombre de comparaisons à réaliser est réduit considérablement par l'approche de recherche itérative basée sur un arbre des espèces (voir chapitre 3.1.3 : Méthodes hybrides) et croît de façon linéaire avec le nombre d'espèces considérées. Une autre approche consiste à pré-filtrer les séquences à comparer deux-à-deux, sur la base de critères de plus haut niveau que la similarité de séquences. Dans porthoDom (Bitard-Feildel et al., 2015), les séquences protéiques sont d'abord regroupées en fonction de leur similarité en terme d'architecture de domaines alors que JustOrthologs s'intéresse au nombre d'exons constituant la séquence codante d'un gène (Miller et al., 2018). Ce dernier programme intègre également une comparaison des séquences indirecte basée sur la

composition en dinucléotides (AT, CG, CC…) des séquences plutôt que sur leur similarité, ce qui permet un gain de temps considérable.

D'une façon générale, il est également possible d'obtenir des gains de temps sur les méthodes existantes en utilisant des algorithmes de comparaisons deux-à-deux plus rapide que la recherche par BLASTP, tel que DIAMOND (Buchfink et al., 2015) ou MMSeq2 (Steinegger et Söding, 2017), utilisables respectivement par OrthologFinder and SonicParanoïd (Cosentino et Iwasaki, 2018) ou en apportant des optimisations aux implémentations actuelles de l'algorithme. Alternativement, le temps de calcul peut être réduit drastiquement en parallélisant les différentes étapes des algorithmes (Ekseth et al., 2014), ce qui permet de tirer avantage des infrastructures de calculs à haute performance.

Finalement, bien que la quantité de génomes disponibles augmente régulièrement et que les annotations de ces génomes connaissent des mises à jour régulières, les relations d'orthologie déjà connues restent stables au cours du temps. On peut les utiliser comme base pour d'autres prédictions comme le font HamStr (Ebersberger et al., 2009), OrthoSelect (Schreiber et al., 2009) et OrthoGraph (Petersen et al., 2017). Dans le principe, de nouveaux gènes sont comparés directement aux groupes d'orthologues existant et ajoutés à ceux-ci en cas de similarité réciproque, ce qui évite la charge de calcul associée aux prédictions *de novo.* Un outil similaire est intégré à la base de données TreeFam (Schreiber et al., 2014).

Reposer sur des prédictions déjà réalisées est d'autant plus aisé qu'il existe aujourd'hui plus d'une trentaine de bases de données de relations d'orthologie dont les données précalculées sont directement disponibles sur des portails web. Ces ressources sont un portail d'accès à des prédictions réalisées sur un jeu de données de taille souvent considérable. Leur valeur ajoutée réside surtout dans la façon dont elles mettent à disposition les relations d'orthologie pour en faciliter l'analyse dans un contexte biologique.

3.3 Représenter l'orthologie : les ressources disponibles

Les bases de données d'orthologie ont une place centrale pour l'utilisation en routine du concept d'orthologie car il s'agit d'un moyen très accessible pour les non-experts d'y récupérer une information sans passer par des opérations de calculs longues ou difficiles à mettre en place.

Leur utilité ne s'arrête toutefois par-là, car elles peuvent exploiter ces données afin de les rendre plus visible pour l'utilisateur, les replacer dans un contexte biologique en y ajoutant des données pertinentes et fournir des outils d'analyses sophistiqués pour leur exploitation. Dans cette section, je fais un tour d'horizon de ces ressources et de leurs apports. Les descriptions à venir n'ont pas vocation à être exhaustives mais à montrer les différentes caractéristiques qui les distinguent. Dans cette optique, je séparerai les ressources en plusieurs grandes classes, il est important de noter que ces catégories ne sont pas mutuellement exclusives et que certaines ressources pourraient être inclues dans plusieurs d'entre elles. Pour faciliter le suivi de l'ensemble des ressources et de leurs caractéristiques, tous les points que j'aborderai dans cette section sont résumés dans un tableau synthétique (Tableau 3-2).

Tableau 3-2 **Les ressources d'orthologie et leurs caractéristiques**. Les ressources d'orthologie se divisent en trois grandes catégories, représentées par différentes couleurs, les ressources spécialisées sont une sous-catégorie des ressources dédiées qui se concentrent uniquement sur une division taxonomique. Les chiffres de la section couverture correspondent au nombre d'espèces représentées dans la ressource, si l'information est indisponible, la case est vide et grisée. Pour les autres sections, les symboles indiquent si l'option correspondante existe (✔) ou non (✗).

		Couverture					**Exploration**										**Représentation**						
Type	**Ressource**	Espèces	Bactéries	Eucaryotes	Archées	Virus	Par id gènes	Par id groupes	Par séquence	Par description	Par fonction	Par domaines	Par distribution	Comp. génomes	Webservice	SPARQL	Orthologues	Fonction	Domaines	MSA	Arbre phylo.	Synténie	Distribution
Ressources d'orthologie dédiées	**Inparanoid**	273				0	✔	✗	✔	✔	✗	✗	✗	✗	✗	✗	✔	✗	✗	✗	✗	✗	✗
	OMA	2 103	1 635	383	149	0	✔	✔	✔	✔	✗	✗	✗	✗	✔	✗	✔	✔	✔	✔	✗	✔	✔
	EggNOG	2 031	1 678	115	238	352	✔	✔	✔	✗	✗	✗	✗	✗	✔	✗	✔	✔	✔	✔	✔	✗	✔
	OrthoDb	**4 667**	3 663	354	659	3 139	✔	✗	✔	✔	✔	✗	✔	✗	✗	✔	✔	✔	✔	✗	✗	✗	✗
	OrthoMCL	150	36	98	16	0	✔	✔	✔	✔	✔	✔	✔	✗	✔	✗	✔	✔	✔	✔	✗	✗	✔
	Hieranoid	66	20	40	6	0	✔	✗	✔	✔	✗	✗	✔	✗	✗	✗	✗	✗	✗	✔	✔	✗	✗
	OrthoInspector	1 947	1 568	259	120	0	✔	✗	✔	✗	✗	✗	✗	✗	✗	✗	✔	✗	✗	✗	✗	✗	✗
	MBGD	**4 742**	4 350	166	226	0	✔	✗	✔	✔	✔	✗	✔	✔	✗	✔	✔	✔	✗	✔	✔	✔	✔
	OtholugeDb	2069		0		0	✔	✗	✗	✗	✗	✗	✔	✔	✗	✗	✔	✗	✗	✗	✗	✔	✗
	HOGENOME	1 470	1 233	140	97	0	✔	✔	✔	✗	✗	✗	✗	✗	✗	✗	✔	✗	✗	✔	✔	✗	✗
	PhylomeDb	1 862				0	✔	✗	✔	✗	✗	✗	✗	✗	✗	✗	✔	✗	✗	✔	✔	✗	✗
	Orthonome	20	0	20	0	0	✔	✗	✔	✔	✗	✗	✗	✗	✗	✗	✔	✔	✔	✔	✗	✗	✗
Ressources spécialisées	**TreeFam**	109	0	109	0	0	✔	✗	✔	✗	✗	✗	✗	✗	✗	✗	✔	✔	✔	✔	✔	✗	✔
	FungiPath	165	0	165	0	0	✔	✔	✔	✔	✔	✗	✔	✗	✗	✗	✔	✔	✗	✔	✔	✗	✗
	Greenphyl	37	0	37	0	0	✔	✔	✔	✔	✔	✔	✔	✗	✗	✗	✔	✔	✔	✔	✔	✗	✔
	PLAZA	119	0	119	0	0	✔	✗	✔	✗	✔	✗	✔	✔	✗	✗	✔	✔	✔	✔	✔	✔	✔
Ressources intégratives	**P-POD**	12	1	11	0	0	✔	✗	✗	✔	✗	✗	✗	✗	✗	✗	✔	✔	✗	✗	✔	✗	✔
	MetaPhOrs	2 714	1 720	877	116	1	✔	✗	✔	✗	✗	✗	✗	✗	✗	✗	✔	✔	✗	✗	✗	✗	✗
	WORMHOLE	6	0	6	0	0	✔	✗	✗	✗	✗	✗	✗	✗	✗	✗	✔	✗	✗	✗	✗	✗	✗
	DIOPT	10	0	10	0	0	✔	✗	✗	✗	✗	✗	✗	✗	✗	✗	✔	✗	✗	✗	✗	✗	✗
	GOOD	4	0	4	0	0	✔	✗	✗	✗	✗	✗	✗	✗	✗	✗	✔	✔	✗	✗	✗	✗	✗
	YOGY	11	1	10	0	0	✔	✗	✗	✗	✗	✗	✗	✗	✗	✗	✔	✗	✗	✗	✗	✗	✔
	HCOP	19	0	19	0	0	✔	✗	✗	✗	✗	✗	✗	✗	✗	✗	✔	✗	✗	✗	✗	✗	✗
Ressources généralistes	**Panther**	112				0	✔	✗	✗	✗	✗	✗	✗	✗	✔	✗	✔	✗	✔	✗	✔	✗	✗
	Ensembl	>1 000				0	✔	✗	✗	✗	✗	✗	✗	✔	✔	✗	✔	✔	✔	✔	✔	✔	✔
	Homologene	0	21	0	0	0	✔	✗	✗	✗	✗	✗	✗	✗	✗	✗	✔	✔	✗	✗	✗	✗	✔

3.3.1 Les ressources dédiées à l'orthologie

La plupart des programmes de prédictions d'orthologie, en plus d'être disponibles sous forme de programme exécutable localement, ont été utilisés pour générer une base de données précalculée regroupant les relations prédites pour un certain nombre d'organismes. Parmi ces bases de données, on compte Inparanoid (Sonnhammer et Östlund, 2015), OrthoInspector (Linard et al., 2015), OMA (Altenhoff et al., 2018), EggNog (Jensen et al., 2008), OrthoDb (Huerta-Cepas et al., 2016), OrthoMCL (Chen et al., 2006), MBGD (Uchiyama et al., 2015), Orthonome (Rane et al., 2017), PhylomeDB (Huerta-Cepas et al., 2014), TreeFam (Schreiber et al., 2014), HOGENOME (Penel et al., 2009), FungiPath (Grossetête et al., 2010), GreenPhylDB (Rouard et al., 2011) et PLAZA (Van Bel et al., 2018).

Ces bases de données diffèrent en termes d'espèces représentées et de services mis à disposition.

3.3.1.1 Nombre et diversité des espèces disponibles

La taille des ressources dédiées à l'orthologie (Section Couverture dans le Tableau 3-2) a cru de concert avec le nombre de génomes disponibles et une majorité d'entre elles couvrent aujourd'hui plus d'un millier d'espèces, avec 4667 espèces pour OrthoDB et 4742 pour MBGD. Cette couverture s'étend sur les 3 Domaines du vivant, avec une majorité de bactéries (jusqu'à 4350 espèces), des centaines d'eucaryotes (jusqu'à 659 espèces) et d'archées (jusqu'à 354 espèces). Pour couvrir l'intégralité des entités biologiques connues, EggNog et OrthoDB intègrent également des gènomes de virus.

De manière à s'adapter à ce grand nombre de relations, les bases de données les plus massives mettent en place des stratégies évitant d'avoir à calculer l'ensemble des relations entre toutes les espèces. Par exemple, OrthoDb met à disposition les relations entre membres de clades majeurs uniquement, MBGD le fait également et fait le lien entre les trois Domaines en utilisant des organismes de référence.

Certaines ressources couvrent seulement des clades spécifiques (Vert foncé dans le Tableau 3-2) : TreeFam couvre uniquement des espèces métazoaires, Orthonome plusieurs espèces de drosophiles, FungiPath les champignons et finalement, PhylDB et PLAZA regroupent les relations d'orthologie entre plantes. Cette dernière ressource est elle-même séparée en plusieurs bases de données pour les différents clades de plantes. D'un point de vue technique, s'intéresser à un embranchement donné permet d'inclure une grande diversité d'espèces sans risquer une augmentation trop importante du nombre de relations à calculer, ainsi qu'à proposer des fonctionnalités adaptées à l'étude du taxon d'intérêt.

Dans les cas où malgré tout, une espèce d'intérêt ne serait pas présente dans les espèces disponibles, EggNOG et MBGD proposent d'utiliser leur ensemble de relations d'orthologie déjà connues pour les étendre à un génome d'intérêt et aider à son annotation. EggNOG propose son service eggNog-Mapper (Huerta-Cepas et al., 2016) qui associe directement les séquences

soumises par un utilisateur aux groupes d'orthologues précalculés pour un clade donné (NOG), permettant ensuite le transfert d'annotations fonctionnelles. Le service MyMBGD permet de calculer *de novo* les relations d'orthologie entre le génome soumis et une partie de ceux présents dans MBGD.

La diversité et le nombre des espèces présentes dans une base déterminent la granularité avec laquelle l'on pourra exploiter les relations d'orthologie en génomique comparative. Les analyses qu'il est possible de réaliser dépendent également de la façon dont on peut accéder aux données.

3.3.1.2 Interroger les données

Les données d'orthologie peuvent être utilisées pour répondre à plusieurs questions biologiques, il existe donc plusieurs façons de les interroger (la plupart du temps par le biais d'un portail web) qui répondent à différents types de besoin : connaitre les relations d'orthologie d'un gène ou d'une famille de gènes en particulier, effectuer une comparaison à l'échelle des génomes complets ou intégrer les résultats à des analyses automatisées.

3.3.1.2.1 Rechercher un gène et ses orthologues

Une des utilisations principales des ressources d'orthologie est la recherche des orthologues d'un gène, soit dans une espèce donnée, soit à l'échelle de la famille de ce gène. L'ensemble des ressources d'orthologie permet d'accéder à cette information par leurs portails web. Les modalités d'accès sont différentes selon les bases et vont de la recherche simple d'un gène à des recherches impliquant la fonction ou l'histoire évolutive (Section Exploration dans le Tableau 3-2).

La méthode la plus simple pour retrouver les orthologues d'un gène est d'utiliser son identifiant ou celui des protéines correspondantes : symbole du gène, identifiants dans les bases de données de référence comme Uniprot (The UniProt Consortium, 2017) ou RefSeq (Sayers et al., 2009). Lorsque l'on s'intéresse à une famille de gènes en particulier, l'identifiant de la famille de gènes ou de l'orthogroupe est une alternative, à condition que celui-ci soit connu. Dans le cas où la séquence dont on veut rechercher les orthologues n'est pas identifiée ou indisponible, il est généralement possible de rechercher ses homologues les plus similaires pour identifier la famille d'orthologues à laquelle elle appartient. Ces fonctions de recherches reposent généralement sur le programme BLAST et permettent, en outre, de retrouver les plus proches homologues du gène recherché si celui-ci n'est pas présent dans la base de données.

Les gènes, protéines ou orthogroupes peuvent également être recherchés sur la base de leur fonction en s'appuyant sur des mots clé dans la description de la protéine ou du gène (Par description, dans le Tableau 3-2) ou des annotations standardisées. Certaines ressources proposent donc de lancer la recherche par terme *Gene Ontology* (OrthoDB, FungiPath, GreenPhyl, PLAZA), identifiants Enzyme Classification (OrthoMCL, FungiPath, GreenPhyl), les annotations de voies de signalisations KEGG (GreenPhyl, MBGD) ou d'autres indicateurs

de fonctions. Similairement, OrthoMCL et GreenPhylDB proposent de chercher les groupes présentant un domaine protéique donné en utilisant les processus PFAM (Finn et al., 2014) ou Interpro (Finn et al., 2017). La recherche par des processus fonctionnels est une valeur ajoutée lorsque l'on s'intéresse aux correspondances de fonction entre gènes ou à l'histoire évolutive de cette fonction.

Comme nous l'avons vu dans le chapitre précédant avec le profilage phylogénétique, la distribution taxonomique d'une séquence et ses orthologues peut apporter des informations sur sa fonction biologique. Elle constitue donc également un paramètre pour l'interrogation de données. Des variantes de la recherche de gènes selon leur distribution évolutive sont implémentées dans OrthoDB, OrthoMCL, MBGD, OrtholugeDB et GreenPhylDB. Leurs implémentations sont diverses selon les bases de données, j'en donne les grandes lignes ici.
Dans OrthoDB, l'option '*Phyloprofile'* permet de rechercher les orthogroupes dont les séquences sont présentes dans tout ou une majeure partie (>80%) des espèces dans une sélection de clades. On ne peut cependant pas spécifier les éléments d'absence, ce qui empêche de spécifier la distribution exacte du gène.

La recherche par profil phylogénétique ('Par distribution' dans le Tableau 3-2) permettant de préciser à la fois des contraintes de présence et d'absence est possible dans les ressources OrtholugeDB et OrthoMCL. Le premier pour 20 espèces seulement, alors qu'OrthoMCL la propose pour 150 espèces avec une interface complète, autorisant des contraintes pour les espèces individuellement et pour les grandes divisions taxonomiques qui les englobent. La recherche par profil phylogénétique présente dans MGBD intègre directement les méthodes adaptées au profilage phylogénétique et permet de rechercher non seulement les orthogroupes correspondant exactement aux contraintes définies, mais également celles qui sont similaires selon plusieurs mesures de distances (Hamming, Correlation, Information Mutuelle). Pour finir, GreenPhylDb pousse le principe de recherche par distribution en permettant de spécifier des contraintes d'histoires évolutives : l'interface de recherche par 'Motif d'arbre' permet de définir manuellement les événements de spéciation et duplication, ainsi que les contraintes de présence et d'absence par le biais d'un arbre phylogénétique. Cette recherche permet ainsi de retrouver toutes les familles dont l'arbre phylogénétique présente ces motifs spécifiques d'évolution.

Les recherches par profils permettent de s'intéresser à l'ensemble des gènes qui ont émergé dans un clade donné ou partagent une histoire évolutive. Elles ont, également, un intérêt certain car elles permettent de tirer profit de certains des avantages du profil phylogénétique.

Que ce soit la recherche par la fonction biologique ou l'histoire évolutive, ce genre de recherches permet d'étudier l'orthologie selon un autre point de vue, mais il s'agit toujours en définitive d'accéder à l'information d'orthologie au niveau du gène. Certaines ressources proposent également d'y accéder au niveau de l'ensemble du génome.

3.3.1.2.2 Comparaison de génomes

La connaissance de l'ensemble des relations d'orthologie offre la possibilité directe de comparer le répertoire de gènes de deux espèces, permettant de retrouver toutes les correspondances entre les gènes des deux espèces et de comparer, par exemple, l'architecture des génomes. Inparanoïd, OrtholugeDB, MDGB et PLAZA permettent d'interroger leurs données dans une perspective orientée sur les génomes, à différents degrés. Les options présentes dans Inparanoïd, OrtholugeDB et OMA permettent d'explorer l'ensemble des relations existantes entre toutes les protéines de deux espèces et d'en exporter le résultat. OrtholugeDB ajoute à cela des comparaisons de l'ordre des gènes entre les génomes, permettant d'en comparer directement le contexte.

OMA et PLAZA vont plus loin dans cette logique en allant jusqu'à proposer la comparaison de l'architecture génomique au sens large en se basant sur les relations d'orthologie. On peut construire sur le site web de ces deux ressources, des diagrammes de points (Figure 3-5) pour identifier les correspondances entre les régions génomiques de deux espèces. Comme on pourrait l'attendre d'une ressource spécialisée, PLAZA intègre également d'autres outils spécifiquement adaptés à l'étude des génomes de plantes, ils permettent notamment d'étudier les événements de duplications de génomes (polyploïdie).

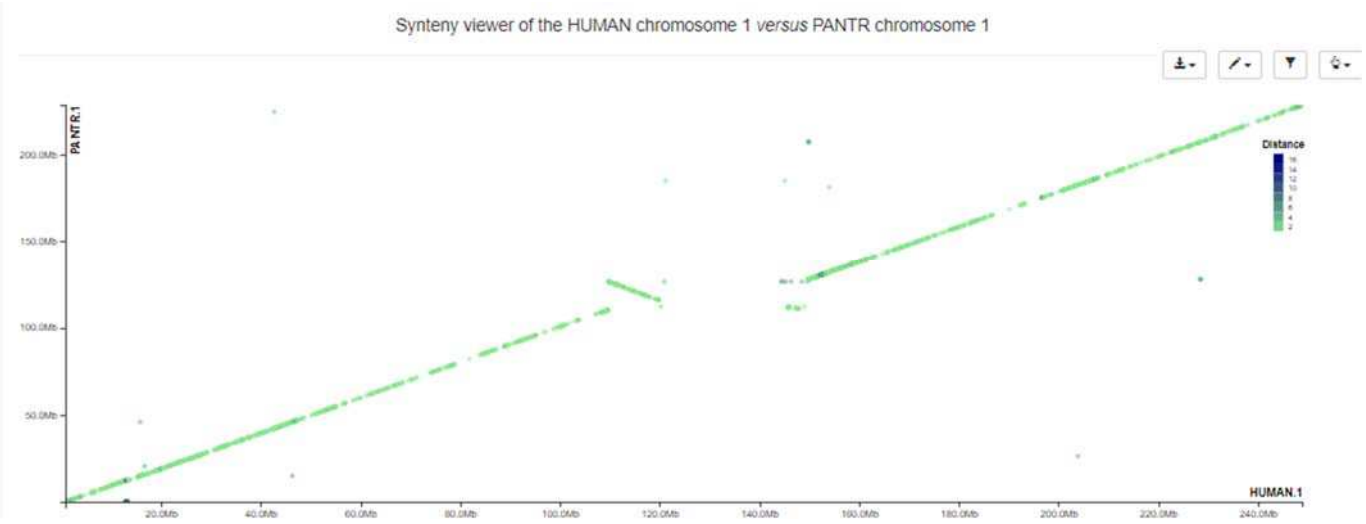


Figure 3-5 **Comparaison de l'architecture génomique dans OMA**. *Dot plot* du chromosome 1 de l'homme avec celui du chimpanzé, en fonction de leurs orthologues, selon l'outil disponible sur le site d'OMA (https://omabrowser.org)

Les comparaisons inter-génomiques dans MBGD se font à l'échelle des clades et sont exploitées de plusieurs manières. Premièrement, le nombre de relations d'orthologie entre gènes est utilisé pour calculer la similarité entre répertoires de gènes de chaque espèce. Une seconde option permet de classer les orthogroupes en fonction du nombre d'espèces du clade où il est représenté, cette classification est précisée et les orthogroupes sont catégorisés en fonction de leur distribution dans les sous-divisions taxonomiques de ce clade.

Ces approches, utilisant les relations d'orthologie pour comparer les génomes, sont une façon d'intégrer directement à ces ressources des exploitations de génomique comparative tirant parti des grands volumes de données d'orthologie déjà calculées.

3.3.1.2.3 Les accès programmatiques

Nous l'avons vu, les ressources d'orthologie donnent accès, *via* leur interface web, aux relations d'orthologie. On peut ainsi explorer un à un les orthogroupes et exporter les résultats pour d'autres analyses. Pour autant, ces moyens d'accès ne sont pas adaptés pour effectuer un grand nombre de requêtes automatiquement, extraire de grands volumes de données, et/ou les intégrer dans des protocoles d'analyse automatisée. Ce rôle est en général rempli par des *webservices* qui permettent directement l'interrogation des ressources et l'extraction de données dans des formats adaptés aux traitements automatiques.

Les ressources EggNOG, OMA et OrthoMCL mettent à disposition de tels services sous la forme d'interfaces de programmation REST (*Representational State Transfer*). Brièvement, ces services permettent d'interroger directement les ressources par requête HTTP, en intégrant l'ensemble des paramètres de recherche dans l'URL de la requête. Les requêtes qu'il est possible de réaliser sont décrites dans une documentation disponible sur les sites de chaque ressource. Elles recoupent généralement les options de requête permises par le portail web :

- Accéder aux informations sur les protéines et les orthogroupes dont on connait l'identifiant (EggNog, OMA, OrthoMCL),
- Effectuer des requêtes basées sur d'autres critères (Recherche par séquences, informations fonctionnelles, domaines) (OMA, OrthoMCL),
- Lister l'ensemble des données (génomes, protéines, groupes) présentes dans la base (OMA),
- Retrouver les relations d'orthologie entre deux espèces (OMA).

Comme mentionné plus haut, les réponses rendues par ce genre de requêtes sont conçues pour être facilement interprétables par l'ordinateur, elles sont donc le plus souvent fournies au format JSON (*Javascript object notation*) ou XML. Ces formats obéissent à une structuration stricte et sont spécialement conçus pour passer des informations d'un programme à l'autre.

L'extraction automatique des données permet de les intégrer, entre autres, dans des protocoles de génomique comparative à grande échelle (e.g analyse phylogénétique de plusieurs familles de protéines) ou pour croiser des données issues d'espèces différentes. Croiser les données de plusieurs sources de données de façon automatisée repose sur l'interopérabilité des bases de données permise par le développement récent d'un autre type d'interface programmatique, les interfaces SPARQL (*SPARQL Protocol and RDF Query Language*). Des interfaces de ce type sont notamment intégrées à de grandes ressources bio-informatiques généralistes dont Uniprot, NextProt ou les ressources développées à l'EBI (Jupp et al., 2014) . Ces interfaces sont basées sur le standard de représentation de données RDF que l'on a mentionné dans la section

précédente. OMA et MGDB ont développé des interfaces de ce type, permettant l'interrogation de leurs données selon une nomenclature commune (*OrthologyOntology*).

3.3.1.3 Représenter les relations

Dans cette section, le terme représentation désigne les approches utilisées pour synthétiser les relations d'orthologie afin d'en faciliter l'exploitation et de les remettre dans leur contexte biologique en y ajoutant des informations externes.

Le mode de représentation des relations d'orthologie le plus simple, adopté par la plupart des ressources, est de fournir une liste de l'ensemble des protéines ou gènes, soit ayant une relation avec une protéine d'intérêt, soit constituant un orthogroupe. Il s'agit là de l'information essentielle, qui peut être suffisante selon les cas, mais qui peut aussi être complétée par des informations de deux types : les annotations fonctionnelles et l'architecture en domaines.

Les annotations fonctionnelles visent à renseigner, dans un vocabulaire standardisé, la fonction biologique ou moléculaire d'une unité biologique. Les types d'annotations que l'on peut retrouver, selon les ressources sont les termes *Gene Ontology* (Gene Ontology Consortium, 2015), les numéros *Enzyme Classification* (Bairoch, 2000) spécifiques de la fonction moléculaires et dans certains cas, des références croisées avec la base bibliographique Pubmed (Sayers et al., 2009). Elles permettent de mettre en avant la signature fonctionnelle d'un orthogroupe, donnant ainsi des pistes pour les protéines du groupe qui ne sont pas encore annotées. Si ces informations sont le plus souvent disponibles sous forme textuelle, indiquant par exemple la proportion de membres d'un orthogroupe annotés par le terme, une représentation visuelle facilite l'interprétation. A titre d'exemple, on peut citer la représentation disponible sur l'interface *ontology* des orthogroupes OMA, qui synthétise pour les orthologues de chaque espèce l'annotation GO et le type de données sur lequel s'appuie l'annotation (Figure 3-6).

Figure 3-6 **Représentation graphique des annotations fonctionnelles dans OMA**. Les colonnes représentent les protéines de l'orthogroupe et les lignes les annotations GO. Un carré indique que la protéine est annotée par le terme correspondant. La couleur représente la source de l'annotation : le bleu correspond à une inférence par similarité de séquences validée par un expert, le gris à une inférence par preuve expérimentale d'interaction protéine-protéine, l'orange à une annotation automatique. Capture d'écran effectuée sur le page de l'orthogroupe OMA 789321 correspondant à la protéine P53 de *Rattus norvegicus*.

Dans le cadre des relations d'orthologie, les informations concernant l'architecture en domaines des protéines permettent une comparaison synthétique des séquences, sans nécessiter un alignement multiple. Cette information renseigne notamment sur la conservation de fonctions entre orthologues. Là encore, la façon de présenter cette information compte. Si certaines ressources font le choix d'indiquer les domaines de manière textuelle, les plus informatives en fournissent une représentation schématique (OMA, OrthoMCL) parfois intégrée à un arbre phylogénétique, pour les remettre dans le contexte de l'histoire évolutive (EggNOG) (Figure 3-7).

Les arbres phylogénétiques et les alignements qui permettent de les construire, sont par ailleurs, un outil d'analyse essentiel pour comprendre en détail les relations entre orthologues et paralogues. A ce titre, ils sont souvent mis à disposition par les ressources, à plus forte raison pour les méthodes basées sur les arbres qui reposent méthodologiquement sur ce type de données. Les alignements multiples sont, soit pré-calculés et disponibles directement sur l'interface, soit réalisables 'à la volée' sur une sélection de séquences. La mise à disposition de ces données sur les sites facilite l'analyse des familles de gènes, en évitant d'avoir à exporter les séquences pour réaliser l'alignement par la suite.

Les arbres phylogénétiques aident à compléter cette analyse, en permettant en principe d'identifier les différents évènements de duplication et de spéciation. Dans cette logique, et pour en faciliter la visualisation, plusieurs ressources indiquent d'ailleurs directement ces événements aux niveaux des nœuds de l'arbre (EggNOG, PhylomeDb, treeFam, Hieranoid). Comme on l'a vu plus tôt, la visualisation des arbres phylogénétiques peut aussi être complétée par des informations de domaines ou une version schématique de l'alignement multiple (Figure 3-7).

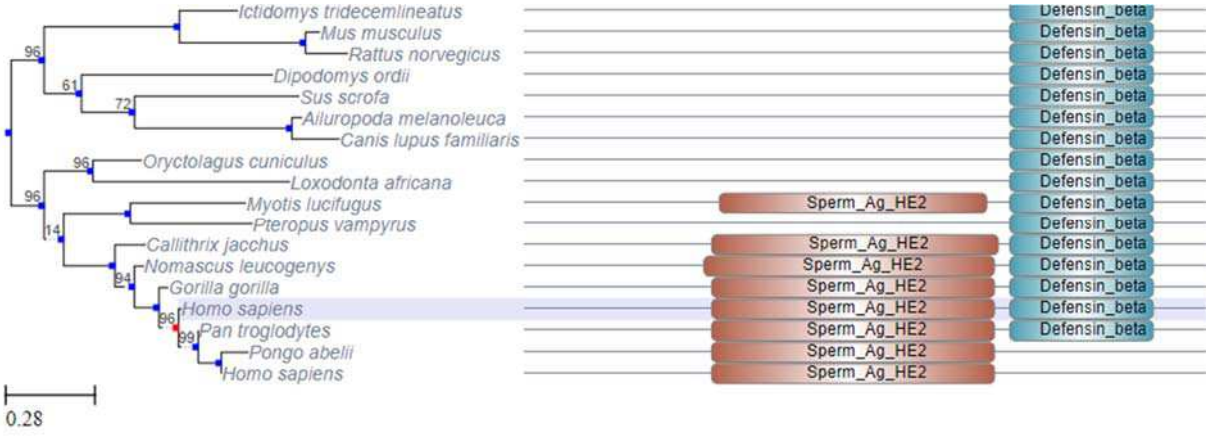


Figure 3-7 **Représentation sous forme d'arbre phylogénétique des relations d'orthologie**. L'arbre phylogénétique représenté indique les évènements de spéciation (en bleu) et de duplication (en rouge). En face de l'arbre, les domaines des séquences sont également indiqués. Capture d'écran modifiée, tirée du site d'EggNOG (http://eggnogdb.embl.de) pour l'orthogroupe ENOG4111B8J.

Comme nous avons pu le voir dans le chapitre précédent, on peut également exploiter les relations d'orthologie pour étudier l'organisation des gènes sur le génome, la synténie. Brièvement, en plus d'aider à valider les relations d'orthologie (deux gènes prédits comme

orthologues, dans un contexte génomique similaire, ont plus de chance de l'être), la synténie peut conduire à l'identification de gènes ayant une fonction proche, notamment chez les procaryotes. On retrouve, sans surprise, des représentations des orthologues orientées vers la synténie dans les bases de données spécialisées pour les Procaryotes comme OrtholugeDb et MGDB. Un outil de ce type est également disponible dans OMA et dans PLAZA. Dans ce dernier, le contexte génomique des paralogues y est également intégré ce qui facilite l'étude des duplications de génomes. Là encore, les options de visualisations sont sensiblement différentes entre les ressources, mais dans l'ensemble, les comparaisons de synténie se font selon une représentation schématique avec un code couleur indiquant les gènes orthologues. On peut ainsi évaluer la conservation de la synténie en un coup d'œil (Figure 3-8). OMA et PLAZA exploitent également l'intéractivité permise par les technologies web pour en faciliter la visualisation : il suffit de passer le curseur sur un gène pour surligner l'ensemble de ses orthologues.

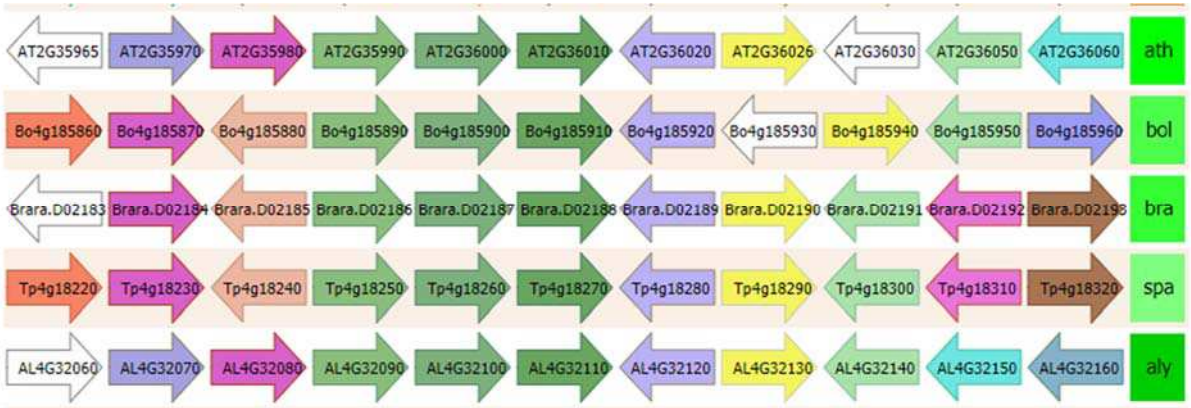


Figure 3-8 **Représentation de la synténie**. Schéma représentant l'ordre et l'orientation des gènes entourant le gène AT2G36010 d'*Arabidopsis thaliana (ath)* et ses orthologues dans 4 espèces : *Brassica oleraceae* (bol), *Brassica rapa* (bra), *Schrenkiella parvula* (spa) et *Arabidopsis lyrata* (aly). Les gènes homologues sont représentés dans la même couleur. Capture d'écran extraite de PLAZA Dicotylédone pour le groupe HOM04D001363.

Le dernier type de représentations que j'aborderai ici suit aussi ce principe, mais concerne cette fois-ci la distribution des orthologues dans les différentes espèces et par extension, dans les grandes divisions taxonomiques. J'ai déjà insisté dans le détail sur les applications des informations de distribution, qui sont centrales à ces travaux de thèse, j'en souligne ici les modalités de représentation et ce qu'elles impliquent (Figure 3-9).

Pour les bases de données comprenant peu d'espèces, il est possible d'indiquer la présence ou l'absence dans chacune d'entre elles individuellement. C'est la stratégie choisie par GreenphylDb et PLAZA, qui affichent la distribution par espèce sous formes de diagrammes.
Dans les cas où les espèces sont plus nombreuses, il est essentiel d'ajouter des informations taxonomiques pour aider à l'interprétation de la distribution. Pour cette raison, les représentations en profil de présence-absence par espèces d'OrthoMCL (Figure 3-9a) et MGBD comprennent une information d'appartenance à de grandes divisions taxonomiques.

Le choix de la représentation de présence ou absence par espèce apporte des avantages, notamment la possibilité d'ajouter à cette information le nombre de gènes inparalogues détectés dans chaque espèce cependant, cette information devient vite difficile à interpréter par l'homme quand plusieurs centaines d'espèces sont répertoriées.

Pour les ressources proposant des milliers d'espèces, la représentation des distributions est une affaire de synthèse des informations. Les ressources résolvent cette problématique en l'affichant par division taxonomique. A titre d'exemple, dans le module de visualisation par profil taxonomique, eggNOG (Figure 3-9b) affiche les espèces représentées dans l'orthogroupe par un diagramme taxonomique dynamique, mimant l'arbre du vivant. Similairement, TreeFam (Figure 3-9c) affiche la proportion des espèces ayant un orthologue dans les clades majeurs de la base de données dans une représentation schématique de la taxonomie. Dans les deux cas, utiliser la distribution par clade permet de garder le sens général des distributions sans avoir à en représenter l'intégralité.

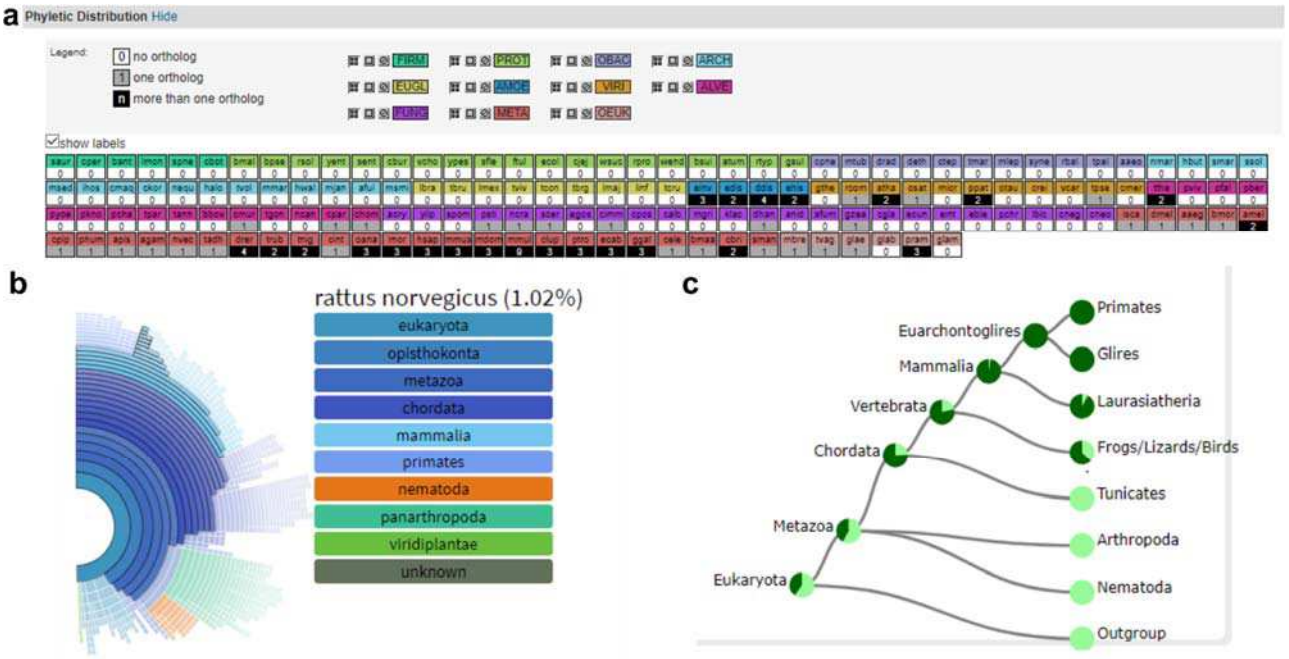


Figure 3-9 **Représentations synthétiques de la distribution des orthologues**. a) Profil phylogénétique par espèce dans OrthoMCL. Les couleurs représentent les divisions taxonomiques. b) Arbre du Vivant dans EggNOG, la taxonomie de chaque espèce possédant un orthologue est affiché sur l'arbre. Les codes couleur indiquent ici aussi les grandes divisions taxonomiques. c) Arbre synthétique des grands clades présents dans TreeFam. La proportion d'espèces du clade ayant un orthologue est indiqué en vert foncé dans les camemberts.

Les différentes façons de représenter et contextualiser l'orthologie référencées ici donnent un aperçu des analyses qui sont mises à disposition dans les ressources d'orthologie. A ce niveau, elles ne constituent pas seulement un dépôt de données mais des plateformes d'analyses complètes où les choix de visualisations jouent un rôle crucial. Cette idée est plus aboutie encore dans les bases de données 'généralistes' intégrant les données d'orthologie que nous verrons par la suite. Avant cela, j'aborde une autre classe de ressources complémentaires aux ressources décrites ici.

3.3.2 Les ressources intégratives

Les ressources d'inférences d'orthologie intégratives (En violet dans le Tableau 3-2) sont une catégorie à part, dans le sens où leur rôle n'est pas de mettre à disposition des prédictions réalisées par une méthode, mais par plusieurs. Leur principe est avant tout de rendre disponible, au même endroit, les prédictions réalisées par différentes méthodes ainsi que d'évaluer leur pertinence. Le concept des prédictions par intégration de plusieurs méthodes ayant été vu dans la section 3.1.4, ce chapitre s'intéressera avant tout aux spécificités de leurs interfaces et à ce qu'elles apportent.

Le premier point spécifique des méthodes intégratives est la variété des méthodes disponibles pour la prédiction d'une relation. Le nombre des méthodes intégrées pour ces prédictions est donc un critère important car elles spécifient le nombre de références croisées que l'on peut y effectuer. Les requêtes peuvent se faire en combinant toutes les sources ou seulement les sources de son choix. Les méta-prédicteurs présentent deux avantages spécifiques. Premièrement, ils servent de point d'entrée pour les ressources d'orthologie, où l'on peut réaliser des analyses plus poussées. Dans cette logique, on retrouve dans HCOP (Eyre et al., 2007) un lien vers l'entrée correspondante des ressources dédiées à l'orthologie.

Le second avantage des méthodes intégratives est la possibilité d'utiliser les prédictions réalisées pour attribuer un score aux relations, que ce soit en prenant en compte le nombre de méthodes ayant participé à une production ou en calculant des scores plus sophistiqués (voir WormHole). Ce score a un intérêt considérable pour choisir la sensibilité et la spécificité voulues dans la détection des relations d'orthologie. La possibilité de filtrer les relations par leurs scores est proposée notamment par DIOPT (Hu et al., 2011), WormHole (Sutphin et al., 2016) et MetaPhors (Pryszcz et al., 2011).

On note que la plupart des ressources intégratives sont restreintes à un nombre peu élevé d'espèces, notamment les organismes modèles, pour lesquels elles regroupent les prédictions pour chaque méthode. MetaPhors à l'inverse, exploite l'avantage d'agréger les prédictions précalculées dans plusieurs bases de données pour en accroitre la couverture.

Pour résumer, les ressources intégratives sont des points d'entrée efficaces pour identifier et évaluer les relations d'orthologie en prenant avantage des différences entre les méthodes de prédictions. Elles ne disposent pas d'options de représentations sophistiquées comme certaines des ressources dédiées mais permettent déjà, à leur niveau, de servir de lien pour intégrer ces différentes ressources.

3.3.3 Les ressources généralistes : l'orthologie intégrée au contexte biologique

La dernière classe de ressources dédiées à la prédiction d'orthologie que nous verrons dans cette section se réfère aux ressources intégrées dans des bases de connaissances biologiques plus importantes (En jaune dans le Tableau 3-2): la section orthologie de Panther (Mi et al., 2010),

Ensembl Compara (Herrero et al., 2016) et la base de relations du NCBI, Homologene (Sayers et al., 2009). La stratégie adoptée pour mettre à disposition les données peut, sur certains points, être comparée aux bases de données 'dédiées' mais leur position particulière, dans un environnement intégrant une grande diversité de données biologiques justifie de les considérer séparément. Dans cette section, je soulignerai comment les relations d'orthologie s'intègrent dans ces systèmes de données biologiques en apportant un regard différent sur ce système et inversement.

3.3.3.1 L'orthologie intégrée à une base de connaissance

Homologene est la ressource d'orthologie du *National Center for Biotechnology Information* (NCBI) (Sayers et al., 2009), l'un des hébergeurs les plus importants de bases de connaissances biologiques. Le NCBI regroupe une quarantaine de bases de données, ayant trait notamment aux données génomiques (projets de séquençage, séquences génomiques...), géniques (séquences des gènes, informations sur les *loci,* données d'expression...) ou protéiques (séquences, structures, domaines), aux descriptions d'espèces (taxonomie) ou encore, à la littérature scientifique (Pubmed, Pubmed Central). Ces bases sont intégrées au sein d'un portail unique, le système Entrez. De fait, les entrées de chaque ressource du NCBI proposent des liens vers le contenu correspondant des autres ressources, pour permettre la description la plus exhaustive des objets biologiques. Homologene n'échappe pas à cette logique et on y retrouve, pour chaque page d'orthogroupe, un accès direct aux autres bases de données : les gènes présents dans le groupe, les protéines pour lesquelles ils codent, l'architecture en domaines ou la littérature ayant trait à ce groupe de gènes. Similairement, on peut accéder à cette base de données par un lien direct depuis les autres ressources. Dans ces cas-là, et bien que la base couvre seulement 21 espèces, les données d'orthologie sont une voie supplémentaire d'exploration des objets biologiques.

L'association des relations d'orthologie à un *hub* d'autres données auxquelles les comparer est aussi disponible sur Uniprot (The UniProt Consortium, 2017). Brièvement, Uniprot est une base de données dédiée aux protéines agrégeant entre autres des données fonctionnelles, d'expression, d'interactions, de pathologies ainsi que des liens avec de nombreuses bases de données biologiques. Elle se divise en deux bases de connaissance : Swissprot dont les annotations font l'objet de curation manuelle par des experts et TrEMBL dont les protéines sont annotées de façon automatique. Uniprot met également à dispositions une sélection de protéomes de référence non redondants pour plus de 10 000 organismes cellulaires choisis pour représenter l'arbre du vivant. Uniprot ne génère pas de prédictions d'orthologie propres mais propose des liens vers les ressources d'orthologie de référence.

3.3.3.2 L'orthologie pour la classification évolutive et fonctionnelle

Protein ANalysis THrough Evolutionary Relationships (PantherDB) (Mi et al., 2016; Thomas et al., 2003) est une ressource visant à classifier les gènes en fonction de leur histoire évolutive, pour faciliter leur analyse en termes fonctionnels. A partir d'un arbre phylogénétique, les gènes partageant une relation d'homologie sont regroupés au sein d'une famille tandis que les groupes d'orthologues et de paralogues récents forment des sous-familles. La particularité de

PantherDB par rapport aux autres ressources est l'accent porté sur le lien entre l'histoire évolutive et la fonction. Chaque famille et sous-famille est associée à des annotations fonctionnelles, principalement GO, qui sont propagées d'un gène à une famille selon la structure de l'arbre phylogénétique. Il est ainsi possible de naviguer dans la ressource non seulement à partir des gènes, mais également à partir de ces annotations. On peut ainsi retrouver l'ensemble des familles ou sous-familles de gènes liées à une annotation voulue.

Cette structure en familles permet également de tirer pleinement profit de la conjecture d'orthologie en permettant de propager des annotations aux gènes non caractérisés d'une famille. Ce potentiel est exploité par les outils d'analyse fonctionnelle disponibles dans Panther, proposant la caractérisation de listes de gènes pour 112 espèces des trois Domaines.

3.3.3.3 Ensembl Compara : l'orthologie pour l'analyse du génome

Ensembl (Zerbino et al., 2018) est une ressource d'exploration et de visualisation de génomes de vertébrés (il existe des portails dédiés à d'autres divisions taxonomiques : métazoaires, plantes, champignons, protistes et procaryotes), intégrant des données d'annotation de gènes et de régions intergéniques, de régions de régulation épigénomique, de variants ainsi que de génomique comparative. C'est de cette dernière catégorie de données, regroupées dans la base de données Ensembl Compara (Herrero et al., 2016), dont il sera question ici.

Dans Ensembl, les relations d'orthologie et de paralogie sont disponibles en tant qu'information de chaque gène. En plus de la liste d'orthologues, comme dans d'autres ressources, on peut visualiser l'arbre phylogénétique ayant permis de les définir, ainsi qu'un arbre taxonomique montrant les événements probables de gains, pertes et duplications de gènes et la distribution du gène dans les espèces considérées.

Une fonctionnalité unique à cette ressource, néanmoins, est la possibilité de visualiser le contexte génomique d'un gène et de ses orthologues dans l'explorateur de génomes. Cela permet non seulement d'étudier la synténie, mais également d'analyser ces orthologues à la lumière d'autres éléments génomiques, tels que les éléments régulateurs. Cela est d'autant plus facilité qu'Ensembl Compara intègre non seulement les données orthologie entre gènes, mais aussi des données d'homologie au niveau du génome, sous forme d'alignements deux-à-deux des génomes, visualisables de concert avec les données d'orthologie.

Les données de génomique comparative d'Ensembl incluent d'autres données à l'échelle du génome, notamment des alignements multiples de génomes pour plusieurs groupes taxonomiques plus ou moins récents (des amniotes aux primates pour l'homme) et les scores des conservations qui en sont issus. A plus large échelle encore, Ensembl offre également une visualisation des blocs de synténie entre les chromosomes de deux espèces permettant d'avoir une vue d'ensemble des événements de réarrangements chromosomiques ayant eu lieu depuis leur séparation. Ainsi, les relations d'orthologie présentes dans Ensembl constituent un niveau dans un vaste panel d'outils de génomique comparative. Ici, celles-ci fonctionnent comme un indicateur parmi d'autres, permettant d'analyser le génome sous différents aspects.

Au-delà d'un état de l'art des méthodes et ressources existantes pour les relations d'orthologie, ce dernier chapitre introductif illustre l'importance de la représentation des données d'orthologie. Les choix de représentation ont un rôle central pour faciliter, démocratiser et permettre une pleine exploitation de l'orthologie et de l'évolution dans différents contextes biologiques et en particulier, dans l'étude des relations génotype-phénotype qui sont au cœur de mes travaux.

Contributions

4 Du vivant aux données et des données au vivant

Les analyses de génomique comparative, et spécifiquement les méthodes basées sur le profilage phylogénétique gagnent en puissance avec le nombre et la diversité des espèces considérées. A l'ère des données massives, la problématique consiste surtout à savoir comment exploiter les données et les rendre interprétables par l'homme. Ce chapitre décrit la façon dont j'ai abordé cette problématique, au niveau conceptuel comme au niveau technique, pour concevoir la ressource d'orthologie OrthoInspector 3.0. Cette ressource constitue le socle technique sur lequel s'appuie le principe des marqueurs évolutifs et les outils d'analyse que j'ai développés pour les exploiter.

Dans ce chapitre, je rappelle l'historique des précédentes versions de la ressource OrthoInspector et mes objectifs dans la mise en place de cette nouvelle version. Je détaille ensuite les choix réalisés lors de sa conception, répondant à plusieurs problématiques propres aux données massives : la qualité des données, le traitement et le stockage et finalement, la représentation synthétique de l'information.

4.1 OrthoInspector : historique et objectifs

4.1.1 Le programme OrthoInspector

OrthoInspector (Linard et al., 2011) est un programme d'inférence de relations d'orthologie basé sur les graphes et conçu pour être rapide et simple d'utilisation. Techniquement, OrthoInspector prédit les relations d'orthologie entre paires de gènes protéiques de plusieurs espèces, à partir de leurs protéomes (au sens de l'ensemble des séquences des protéines codées par un génome). Dans ce but, il exploite les résultats de comparaison de séquences fournis par BLAST et infère séparément le type de relations d'orthologie (un-à-un, un-à-plusieurs et plusieurs-à-plusieurs) entre chaque paire de protéomes. Il prédit donc les relations d'orthologie entre deux espèces mais, contrairement à d'autres méthodes basées sur les graphes, il ne les utilise pas pour créer des orthogroupes étendus.

L'algorithme, codé en JAVA, construit une base de données relationnelle SQL en plusieurs étapes. Pour des raisons de performance, cette base de données est utilisée pour héberger non seulement les prédictions finales, mais également les résultats des étapes intermédiaires dont je décris brièvement le déroulement ici :

1. *Organism_step*. Les espèces et toutes les séquences de leurs protéomes sont enregistrées dans la base.
2. *Blast_parsing*. Dans cette étape, les résultats des BLAST tous-contre-tous sont analysés. Pour chaque gène d'un organisme A, on identifie le meilleur hit pour chaque espèce,

ainsi que les autres gènes de l'organisme A dont le score est meilleur que ce meilleur *hit*. Les gènes dans ce cas sont enregistrés en tant qu'inparalogues potentiels par rapport à cette espèce (Figure 4-1)

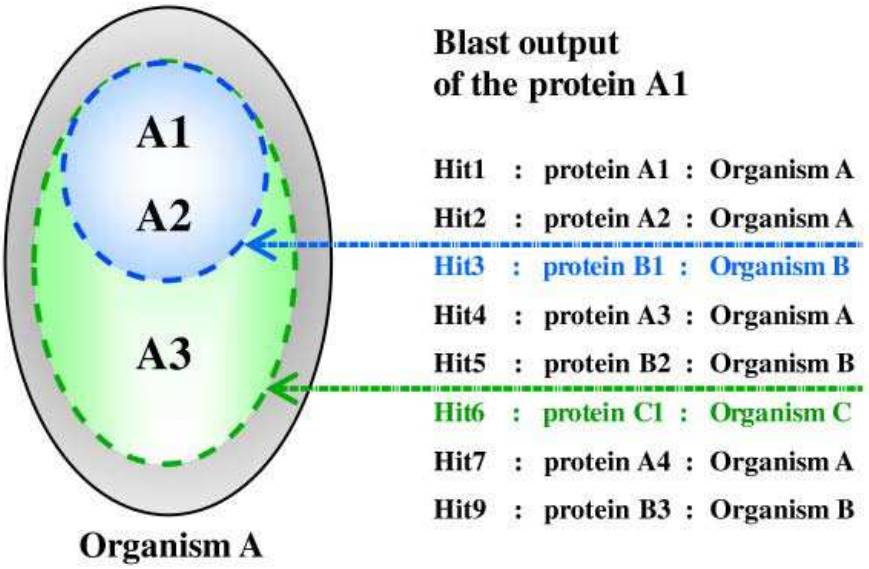


Figure 4-1 **Traitement du BLAST dans OrthoInspector**. Ici, B1 est le meilleur *hit* de l'espèce B pour la protéine A. La protéine A2, ayant un meilleur *hit* que B1, est considérée comme inparalogue potentiel. Tiré de (Linard et al., 2011)

3. *Inparalog validation.* Les groupes potentiels d'inparalogues sont comparés les uns aux autres pour identifier le plus grand groupe d'inparalogues consensus. Par exemple, si les résultats de BLAST du gène A1 donnent {A1, A2, A3} comme inparalogues par rapport à l'espèce C. On vérifie qu'un groupe identique est retrouvé dans les résultats de BLAST de A2 et A3. Si ce n'est pas le cas et que, par exemple, A3 est à l'origine d'un groupe {A1, A3}, la présence de A2 ne fait plus consensus et on conserve le groupe {A1, A3}, constitué uniquement des cas consensuels. Les inparalogues validés sont, à cette étape, enregistrés comme tel dans la base de données.
4. *Calcul des orthologies.* Dans cette dernière étape, les meilleurs *hits* des gènes entre eux ou avec les groupes d'inparalogues déterminés à l'étape précédente sont comparés, de façon à confirmer ou infirmer l'hypothèse du meilleur hit réciproque. Les éventuels conflits de relations, par exemple l'orthologie des membres d'un même groupe d'inparalogues avec des gènes non inparalogues entre eux, sont résolus. Les relations confirmées sont enregistrées dans la base de données, dans l'une des trois tables prévues à cet effet, une pour chaque type de relations : *one-to-one*, *one-to-many*, *many-to-many*.

Les différentes étapes détaillées ici peuvent être exécutées de façon parallèle, rendant possible le calcul de relations d'orthologie à partir de plusieurs centaines de protéomes dans un temps raisonnable. En termes de résultats, OrthoInspector obtient un bon équilibre entre sensibilité et spécificité sur l'ensemble des *benchmarks* QFO (Altenhoff et al., 2016). En cela, il s'avère

précieux pour de nombreuses analyses en particulier pour les méthodes de profilage phylogénétique, qui souffrent de façon équivalente des faux positifs et des faux négatifs.

4.1.2 Les ressources OrthoInspector

Comme les autres programmes majeurs d'inférence d'orthologie, OrthoInspector propose, depuis sa première publication, une ressource d'orthologie accessible par le web. Dans sa première version, cette ressource incluait les protéomes de 59 espèces, toutes Eucaryotes (Linard et al., 2011). Suivant l'apparition rapide de nouveaux protéomes, la deuxième version d'OrthoInspector (Linard et al., 2015) regroupe des prédictions d'orthologie entre 1947 espèces réparties en deux bases de données distinctes : une base dédiée aux Eucaryotes avec 259 espèces au total et une base dédiée aux Procaryotes incluant 1568 Bactéries et 120 Archées (Figure 4-2). A ces deux bases s'ajoute une troisième, centrée sur les protéomes de référence QFO et comprenant 147 espèces des trois domaines.

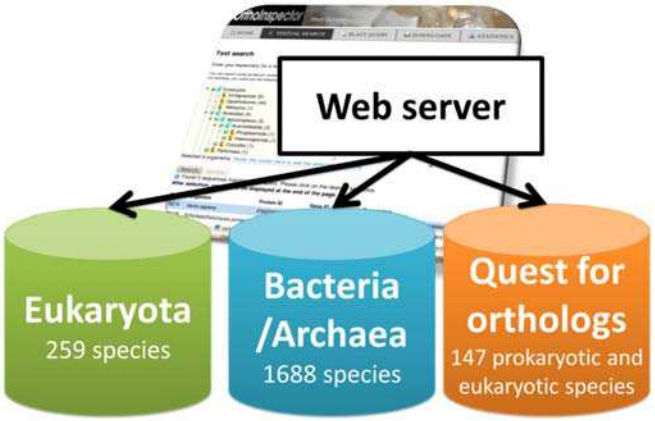


Figure 4-2 **Bases de données d'OrthoInspector v2**. Trois bases de données sont disponibles, selon les divisions taxonomiques d'intérêt. Chacune est accessible par un portail web spécifique.

Comme nous l'avons vu dans le chapitre 3, il est possible d'accéder aux relations d'orthologie de chacune de ces bases de données, par le biais d'un portail spécifique, en utilisant l'identifiant ou la séquence d'une protéine d'intérêt. On y retrouve, cependant, peu d'outils pour visualiser et analyser ces relations sous un autre angle.

Malgré une couverture en termes d'espèces relativement importante, les bases techniques des ressources OrthoInspector 2.0 en termes d'accès aux données ou de représentation, ne sont pas adaptées au flux de données actuel. Cet état de fait nous a donc mené au développement d'une nouvelle version de cette ressource, OrthoInspector 3.0 à même de relever ces défis et de servir de base solide à des analyses de génomique comparative.

4.1.3 OrthoInspector 3.0 : un socle pour développer de nouveaux outils

Comme je l'ai mentionné, OrthoInspector 3.0 est pensé pour constituer un socle sur lequel appuyer des analyses de génomique comparative. Elle doit donc correspondre à trois prérequis essentiels aux ressources d'orthologie.

Le premier est de fournir une bonne couverture du Vivant, ce qui implique à la fois d'inclure un grand nombre d'espèces, mais surtout une diversité taxonomique la plus complète possible avec des représentants de la plupart des phyla connus. On pourrait penser qu'augmenter le nombre de génomes augmente *de facto* la diversité, mais certaines catégories taxonomiques sont proportionnellement plus représentées que d'autres dans les bases de données génomiques. Il s'agit donc de sélectionner les protéomes à utiliser en apportant un soin particulier à leur diversité taxonomique. De plus, ces protéomes doivent être le reflet fidèle des répertoires protéiques des espèces qu'ils représentent, augmenter le nombre d'espèces couvertes ne doit donc pas se faire au détriment de la qualité des données.

Le second prérequis pour une ressource hébergeant des analyses de génomique comparative est l'accessibilité des données. L'objectif ici étant de permettre d'accéder, en un nombre d'étapes le plus réduit possible à l'ensemble des relations présentes dans la ressource. Cela revient donc, en contraste avec OrthoInspector 2.0, à mettre en place un portail d'interrogation unique pour l'ensemble des données. Finalement, l'objectif étant de permettre des analyses de génomique comparative, il est également important de rendre les données accessibles de manière automatique, par le biais d'une interface programmatique.

Pour finir, le dernier prérequis est la notion de visualisation et de représentation des données. Cet aspect est essentiel pour les bases couvrant de larges volumes de données, pour permettre d'en extraire rapidement les informations pertinentes. Nous le verrons, la notion de représentation implique non seulement la visualisation, mais également la mise en place d'une structure de données efficace, à même de s'adapter aux analyses dans différents contextes.

Ces trois prérequis sont les trois grands axes qui ont organisé mes choix dans la conception d'OrthoInspector 3.0 et qui structurent ce chapitre.

4.2 Sélection des protéomes

Le premier objectif dans la conception d'OrthoInspector 3.0 était donc d'obtenir une couverture du Vivant la plus complète possible dans l'état des données disponibles. Il s'agissait d'augmenter le nombre de protéomes couverts en prenant en compte deux critères principaux : la diversité et la qualité, essentielles pour représenter le Vivant efficacement.

4.2.1 La couverture du vivant : protéomes de référence UniProt

Le nombre et la diversité des protéomes complets, comme ceux de l'ensemble des données génomiques, est en augmentation constante. On compte aujourd'hui dans UniProt, sans prendre en compte les virus, plus de 90 000 organismes dont les protéomes sont disponibles. Pour autant, la façon dont ces données couvrent le Vivant est hétérogène en fonction des clades considérés. Premièrement, il existe fréquemment plusieurs protéomes pour des souches proches d'une même espèce, notamment chez les Bactéries. Les répertoires de gènes de ces souches étant relativement similaires, on parle de protéomes redondants. UniProt propose de filtrer ces données, ce qui permet donc d'en réduire le nombre, fin 2018, à 24 136 protéomes distincts

couvrant une majorité de 21 782 Bactéries, quelques 1 449 protéomes d'Eucaryotes et 905 protéomes d'Archées.

Même en excluant les protéomes redondants, les protéomes ne couvrent pas de façon équilibrée l'ensemble de la taxonomie et certaines divisions taxonomiques sont surreprésentées, par exemple près d'un quart (4678) des protéomes bactériens correspond à l'embranchement des Firmicutes. Un trop grand nombre de protéomes d'espèces proches, dont les répertoires de gènes sont souvent similaires, apporte peu à la génomique comparative et dans le pire de cas, peut être source de biais. J'ai donc choisi, dans le cadre d'OrthoInspector 3.0, de ne pas intégrer tous les protéomes connus, mais plutôt un échantillonnage suffisamment large pour être représentatif du plus grand nombre de branches de l'arbre du Vivant.

J'ai donc choisi de me restreindre aux protéomes de référence proposés par UniProt, ce qui correspond à 5443 protéomes répartis dans les 3 Domaines du Vivant (440 bactéries, 830 eucaryotes et 213 archées).Il s'agit d'un sous-ensemble des protéomes non redondants, qui sont sélectionnés manuellement et automatiquement pour garantir une représentation équilibrée du Vivant tout en incluant l'ensemble des organismes modèles ou d'intérêt pour la communauté scientifique au sens large. Les protéomes de références de virus sont également disponibles sur UniProt et nous les ajouterons ultérieurement à OrthoInspector, une fois que nous aurons défini des critères de qualités pertinents pour ce type de protéomes.

4.2.2 La qualité des protéomes : séparer le bon grain de l'ivraie

La qualité des données est une problématique récurrente dans le cadre des données massives et, comme nous l'avons vu, les génomes et les annotations de gènes protéiques sur ces génomes n'échappent pas à cette règle (voir section 1.31 pour les sources d'erreurs). En conséquence, le protéome d'une espèce peut être partiel et les protéines prédites fragmentaires ou erronées. Dans le cadre de la génomique comparative, l'utilisation de tels protéomes est potentiellement une source d'erreurs. C'est d'autant plus vrai pour le profilage phylogénétique, où une protéine non prédite sera considérée comme totalement absente d'une espèce et fausser les analyses ultérieures.

Pour anticiper ces erreurs, nous avons choisi de filtrer, en amont des prédictions d'orthologie, les protéomes de moindre qualité. Les méthodes existantes pour l'évaluation de la qualité des protéomes reposent principalement sur des comparaisons d'orthologie avec des génomes dont les annotations sont connues (Dunne et Kelly, 2017; Simão et al., 2015). Reposer sur les informations d'orthologie n'étant pas une option dans notre cas, nous avons cherché à identifier d'autres indicateurs à même de nous informer, d'une façon rapide et automatique, sur la qualité des protéomes.

4.2.2.1 Filtration des protéomes avec des protéines tronquées

Comme je l'ai mentionné, l'un des cas de figure que nous souhaitions éviter est un nombre important de protéines fragmentaires, de taille réduite par rapport aux séquences protéiques

réelles. Afin de déterminer si la taille de l'ensemble des protéines d'un protéome pouvait être un premier indicateur de sa qualité, nous avons examiné la distribution de la longueur des séquences protéiques pour chacun des protéomes de référence (Figure 4-3).

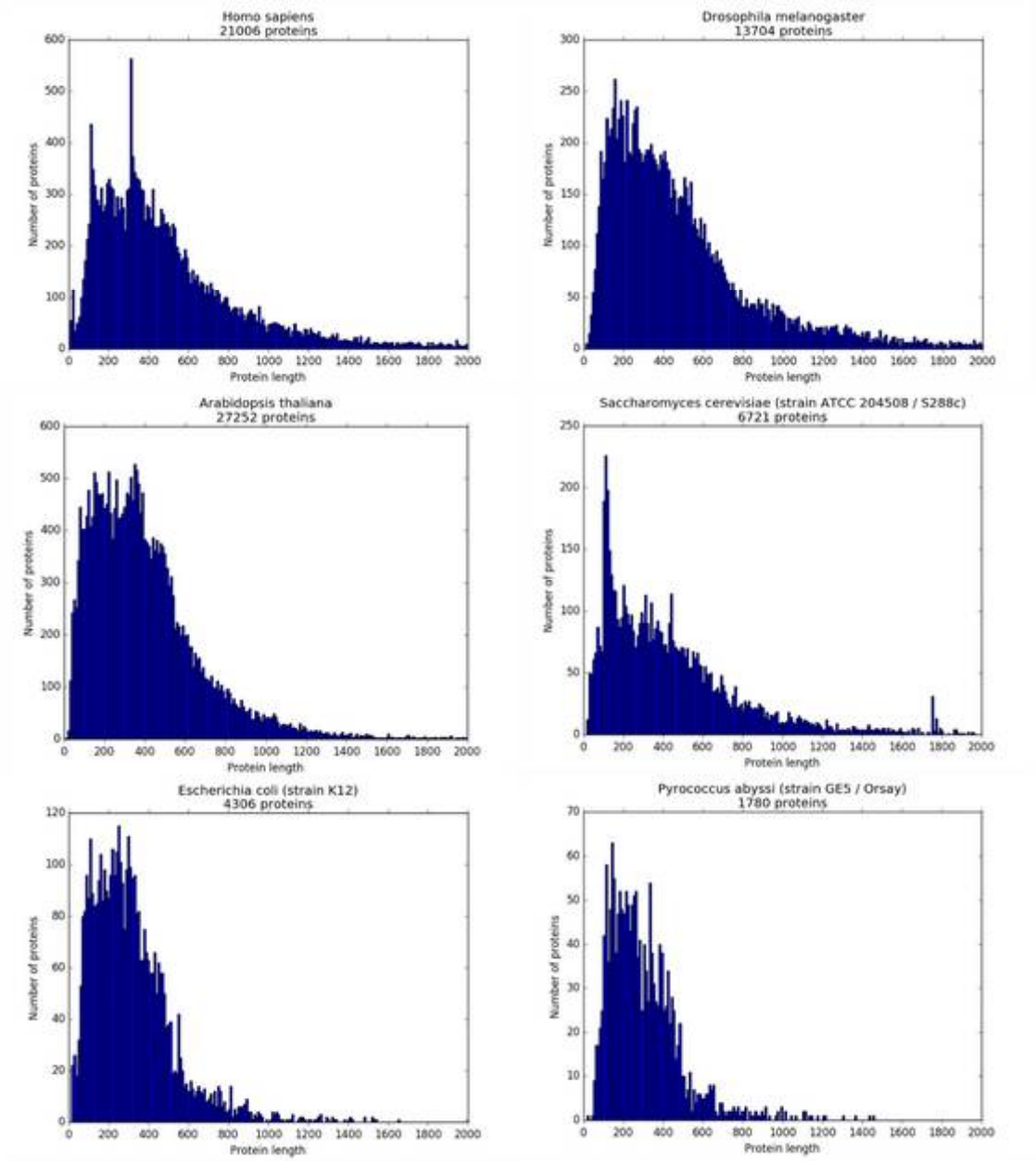


Figure 4-3 **Distribution de la longueur des séquences protéiques chez plusieurs organismes modèles**. L'axe des abscisses correspond à la longueur des protéines (en acides aminés) et s'arrête à 2000 pour des raisons de visualisation. Les protéomes choisis ici ont fait l'objet d'annotations manuelles et représentent plusieurs groupes taxonomiques : métazoaires (*Homo sapiens, Drosophila melanogaster*), plantes (*Arabidopsis thaliana*), champignons (*Saccharomyces cerevisiae*), bactérie (*Escherichia coli*) et archée (*Pyrococcus abyssi*). Les protéomes représentés ont une distribution similaire malgré les différences taxonomiques.

Les distributions obtenues pour des protéomes d'espèces très étudiées et dont l'annotation fait régulièrement l'objet de curations manuelles sont globalement similaires, asymétriques et centrées vers la gauche, avec un pic entre 200 et 400 acides aminés. La différence majeure que

l'on peut noter est un biais plus prononcé, chez les Eucaryotes, vers les protéines de grande taille. En outre, on observe pour certains protéomes une surreprésentation de protéines d'une certaine taille comme pour les protéines de 380 acides aminés pour *Homo sapiens* et de 180 acides aminés pour *Saccharomyces cerevisiae.* Ces différences mineures s'expliquent par l'expansion importante de certaines familles de gènes par duplications dans leur lignée, entrainant une surreprésentation de ces protéines par rapport à l'ensemble du répertoire de gènes. D'une façon générale, on retrouve une distribution de ce genre pour la plupart des protéomes considérés et elle correspond par ailleurs, à la distribution globale de toutes les protéines de UniProtKB/SwissProt (Figure 4-4). Il s'agit donc, selon toute vraisemblance, de la composition attendue pour un protéome issu d'annotations de qualité.

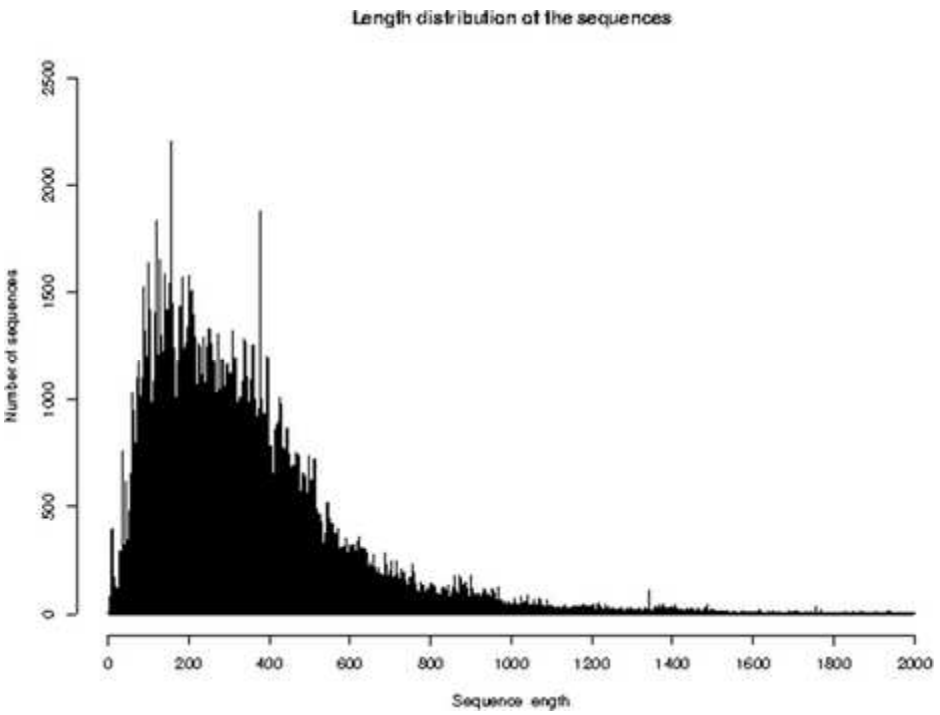


Figure 4-4 **Distribution de la longueur des protéines dans la banque SwissProt.** Figure établie à partir des statistiques officielles de la mise à jour d'août 2018 accessibles sur le site https://web.expasy.org/docs/relnotes/relstat.html

Si cette distribution est majoritaire, une inspection visuelle nous permet d'identifier des protéomes qui n'y obéissent pas. Ces protéomes se démarquent par des distributions clairement biaisées vers les protéines de petite taille (1 à 50 acides aminés), et qui décroit rapidement vers les tailles supérieures (Figure 4-5a). Ces distributions sont observées dans les trois Domaines du Vivant et ne semblent pas associées à une division taxonomique précise. On retrouve ainsi des protéomes présentant des distributions biaisées au sein de clades avec des protéomes présentant une composition « normale ». A titre d'exemple, la comparaison de *Caenorhabditis japonica* à *Caenorhabditis elegans*, l'espèce modèle du même genre, illustre directement ce contraste (Figure 4-5b). Il paraît peu probable qu'une telle différence au niveau du répertoire de gènes existe réellement entre espèces proches. Ces distributions biaisées indiquent donc probablement une forte proportion de séquences protéiques tronquées.

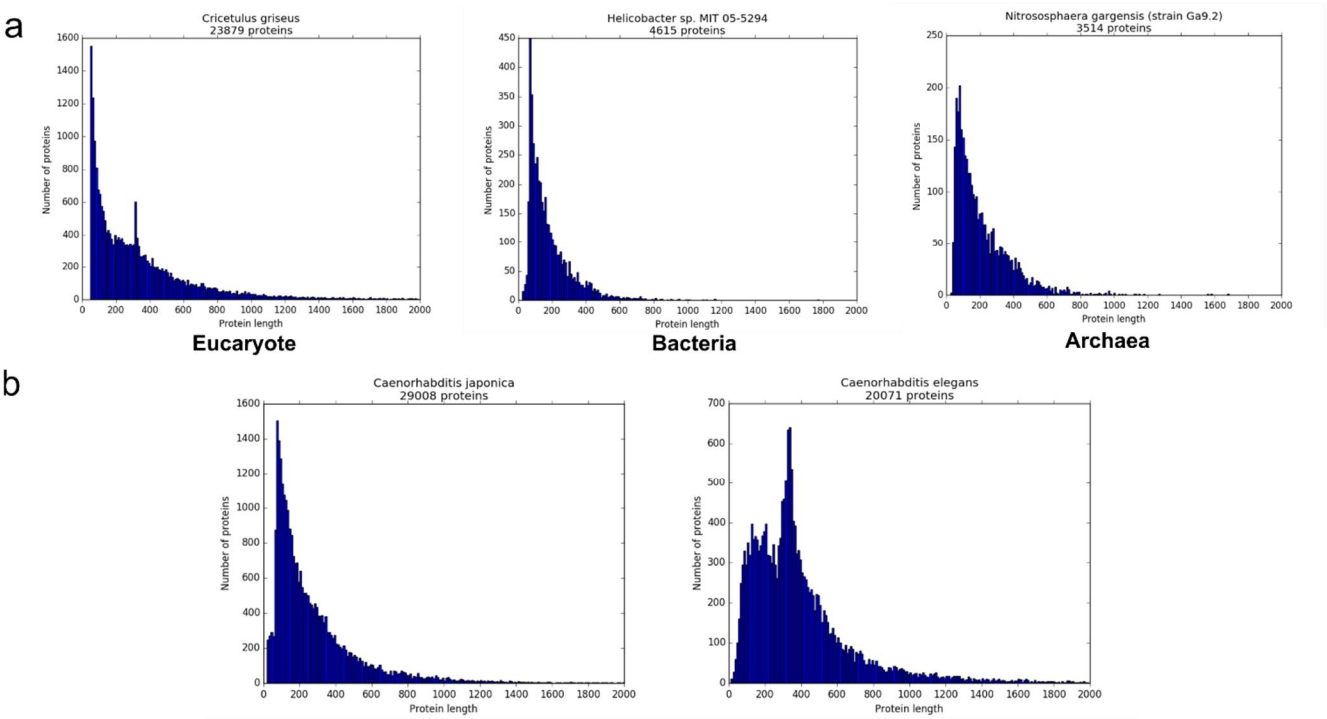


Figure 4-5 **Distributions biaisées de la taille des protéines dans certains protéomes**. a) Une distribution 'anormale' avec une représentation plus importante de petites protéines est observée dans les 3 domaines du vivant. b) Comparaison de la distribution de la taille des protéines du nématode *Caenorhabditis japonica* et de l'organisme modèle proche *Caenorhabditis elegans*, présentant une distribution classique.

Cette première analyse visuelle révèle l'existence d'une proportion non négligeable de protéines fragmentaires dans certains protéomes. De façon à filtrer ces protéomes, nous avons défini plusieurs critères :

- *La proportion de protéines annotées comme 'Fragments'* : dans UniProt, les protéines fragmentaires sont parfois annotées comme telles. Il s'agit d'une information directe qui permet de filtrer les protéomes avec une quantité anormale de ces protéines.
- *La proportion de protéines de petite taille* (inférieures à 100 acides aminés) : les protéines de moins de 100 acides aminés existent naturellement et constituent de 3 à 11% des protéomes d'organismes modèles, dont l'annotation est fiable. Similairement, elles représentent 10% des protéines de SwissProt, qui font l'objet de curation. L'analyse visuelle nous permet cependant de déterminer qu'une proportion anormalement élevée de petites protéines est indicatrice de protéomes de mauvaise qualité.
- *Proportion de protéines sans codon initiateur* : les protéines tronquées peuvent provenir d'un génome fragmenté en de nombreux contigs ne couvrant pas toujours l'extrémité 5' des régions codantes. Dans ces cas, le premier acide aminé ne correspondra pas à la méthionine du codon initiateur. Il existe chez les Procaryotes des codons initiateurs alternatifs ne correspondant pas au codon ATG, mais ceux-ci sont automatiquement traduits en méthionine. On peut donc considérer qu'une forte proportion de protéines ne commençant pas par une méthionine est une indication de protéines fragmentaires.

Les seuils ont été choisis afin d'éliminer les protéomes présentant des proportions trop élevées de séquences fragmentaires pour des prédictions d'orthologie, à savoir 10% de protéines annotées comme fragments, 20% de protéines de petite taille et 10% de protéines ne commençant pas par une méthionine pour les Bactéries et Archées. Pour les Eucaryotes, nous avons modifié ces seuils en prenant en compte les difficultés supplémentaires afférentes à l'assemblage des génomes, souvent plus longs et riches en éléments répétés et aux problèmes posés par l'épissage. Le seuil pour la proportion de protéines ne commençant pas par une méthionine a été porté à 55% du protéome.

4.2.2.2 Filtration des protéomes incomplets

Après avoir mis en place ces outils de filtration de protéomes riches en protéines tronquées, nous avons pris en compte un second paramètre, la complétion de ces protéomes. On ne peut *a priori* pas évaluer directement le nombre de gènes protéiques manquants dans un génome sans passer par une comparaison avec d'autres génomes, et donc par des relations d'orthologie. Il est en revanche possible d'avoir une idée approximative des tailles minimales des répertoires de gènes de chaque domaine du Vivant. De façon à éliminer les génomes incomplets, nous avons ainsi examiné les protéomes les plus réduits chez les Eucaryotes et Procaryotes et identifié la taille la plus basse pouvant être attribuée à des spécificités biologiques spécifiques de l'espèce : la symbiose ou le parasitisme.

Pour les Bactéries, cette limite est basse et se situe à 137 protéines, pour le génome de *Nasuia deltocephalinicola*. Cette espèce est impliquée dans une relation de symbiose avec l'insecte *Macrosteles quadrilineatus* et possède un génome de 112kb, l'un des plus petits jamais séquencés (Bennett et Moran, 2013). Sur la base de ce seuil, seulement trois protéomes bactériens avec un protéome plus réduit ont été exclus.

L'archée au protéome le plus réduit (535 protéines) dont le cycle de vie est caractérisé est *Nanoarchaeum equitans*. Là encore, cette espèce qui possède un génome de taille réduite (490kb) est impliquée dans une relation parasitique avec d'autres archées du genre *Ignicoccus* (Waters et al., 2003). Seulement deux protéomes d'archées ont un protéome de taille inférieure à 535 protéines dans notre jeu de données. Une vérification de la littérature a permis de confirmer qu'ils proviennent bien de séquences génomiques incomplètes, issues d'expériences de métagénomique (Castelle et al., 2015).

Chez les Eucaryotes, même les protéomes de taille réduite regroupent sensiblement plus de protéines que pour les autres Domaines. Parmi les protéomes de référence eucaryotes, le protéome le plus réduit que nous avons pu confirmer est celui de *Ordospora colligata* (Pombert et al., 2015), un champignon de l'embranchement des Microsporidies, dont les membres sont des parasites intracellulaires. Les Microsporidies sont justement connus pour la réduction extrême de leur génome et de leur répertoire de gènes, le protéome de référence d'*Ordospora colligata* regroupe 1810 séquences protéiques, à peine moins que ceux d'*Encephalitozoon cuniculi* et de *Nosema ceranae* deux autres représentants du groupe ayant respectivement 2008 et 2060 gènes codant pour des protéines. Dix autres protéomes eucaryotes ne dépassant pas le

seuil de 1810 protéines furent donc filtrés, la littérature ne faisant pas mention d'une réduction du génome.

4.2.2.3 Protéomes sélectionnés

En définitive, ces deux jeux de filtres préliminaires appliqués aux 5443 protéomes de référence UniProt, nous ont permis d'éliminer 690 protéomes (12,7%), répartis proportionnellement entre les trois Domaines du Vivant. Le Tableau 4-1 résume le résultat de cette étape. On remarque un recouvrement entre les différents filtres pour les Archées et Bactéries et notamment, qu'une forte proportion de séquences ne commençant pas par une méthionine corrèle bien avec les autres indicateurs de forte proportion de séquences fragmentaires, tous les protéomes filtrés par ce critère l'étant également par d'autres.

Tableau 4-1 **Résumé de l'étape de contrôle qualité**. Les colonnes 3 à 6 indiquent les nombres de protéomes exclus uniquement par un filtre ou, entre parenthèses, en combinaison avec un ou plusieurs autres filtres. La colonne 'combinés' indique le nombre de protéomes éliminés par plusieurs filtres. Les astérisques représentent les catégories où est compté le protéome de *Lokiarchaeum*, initialement filtré, mais que nous avons choisi d'inclure dans notre jeu final de protéome.

	Initial	Annotations « fragment »	Petites protéines	Codon initiateur	Tailles protéomes	Combinés	Eliminés	Final
Archées	**213**	15* (24)	10 (16)	0 (4)	1 (2)	9	35*	**179**
Bactéries	**4 400**	219 (323)	212 (263)	0 (84)	0	106	537	**3 863**
Eucaryotes	**830**	/	66 (67)	42 (43)	10 (10)	1	119	**711**
Total	**5 443**	234* (347)	288 (346)	42 (131)	11 (12)	116	691*	**4 753**

Les problématiques de contrôle qualité peuvent avoir un effet sur la diversité des protéomes couverts en risquant d'entraîner l'absence de certains clades. Ainsi, seulement trois représentants du clade des oiseaux (*Aves*) sur 45 ont été conservés de même que le seul protéome de référence des Gymnospermes (*Picea glauca*). Dans ces deux cas, les indicateurs de qualité ne permettent pas de faire exception pour les représentants de ces clades. A l'inverse, chez les Archées, nous avons fait le choix de conserver le protéome de *Lokiarchaeum sp. GC14_75*, l'un des seuls représentants du clade ASGARD dont l'intérêt est considérable pour la génomique comparative entre domaines. Le protéome de ce dernier dépassant uniquement le seuil d'annotation de fragments (14% de protéines au lieu de 10%), nous avons considéré que son inclusion présentait, *in fine*, plus d'avantages que d'inconvénients.

Ce dernier exemple conclut cette étape de sélection du protéome et en souligne bien le principe : il s'agissait d'assurer une large couverture tout en maintenant un objectif de qualité. Dans les deux cas, le but est d'apporter le moins de biais possible dans les données, qui pourraient fausser les analyses réalisées par la suite. Avec une sélection de 4753 espèces réparties dans l'ensemble des clades, la mise à jour d'OrthoInspector 3.0 en fait l'une des ressources d'orthologie les plus exhaustives en termes d'espèces représentées. Une telle quantité de protéomes à traiter et à représenter entraîne son lot de problèmes qui font l'objet des sections suivantes.

4.3 Une architecture à dimension variable

Un des problèmes posés par des relations d'orthologie établies pour des milliers de protéomes réside dans la façon de permettre l'accès à l'ensemble des données tout en mettant en avant les informations pertinentes. Les informations considérées comme telles dépendent cependant de la question biologique qui motive les recherches de relations d'orthologie. En amont de la mise en place des bases de données, l'architecture de données d'OrthoInspector 3.0 a été conçue pour s'adapter à plusieurs contextes.

4.3.1 Différents niveaux de granularité : exhaustivité contre pertinence

Pour appuyer notre réflexion sur l'architecture de la ressource OrthoInspector, nous avons considéré plusieurs cas d'utilisation d'une ressource d'orthologie :

- La recherche d'orthologie entre deux espèces pour des transferts d'informations expérimentales par exemple,
- L'analyse de génomique comparative à grande échelle (échelle des domaines ou de grands clades),
- L'analyse de génomique comparative à l'échelle d'un clade d'intérêt.

L'exhaustivité des données est essentielle pour le dernier cas, où avoir de nombreux représentants du clade étudié permet une analyse détaillée et précise. A l'inverse, dans le premier cas, on s'intéresse à une partie de l'information et il est important de pouvoir y accéder rapidement malgré la masse de donnée. Finalement, le second cas est à mi-chemin des deux, l'analyse à grande échelle s'appuyant souvent sur un sous-ensemble d'espèces représentatif de la diversité. Notre choix ici est donc de fournir à la fois une information synthétique de haute qualité pour les études globales et l'accès aux données complètes pour les cas de figure plus spécifiques.

Pour ce faire, nous avons défini un sous-ensemble d'espèces représentées par un protéome de référence, que nous appellerons ici organismes 'modèles' regroupant les organismes les plus importants dans la communauté scientifique (dont les modèles expérimentaux) tout en assurant une représentation synthétique de la diversité du Vivant. Pour la sélection de ces protéomes, nous nous sommes appuyés sur les différentes divisions taxonomiques officielles (embranchement, classe, ordre) en sélectionnant au minimum une espèce par division choisie. Lorsque plusieurs espèces étaient représentées, les espèces servant de modèles expérimentaux pour ce clade ont été sélectionnées. En l'absence de tels organismes, nous avons privilégié la qualité des protéomes en choisissant ceux avec le plus de protéines issues de SwissProt et avec de bons scores pour les indicateurs de qualités vus dans la section précédente. Selon les domaines considérés, nous avons adapté cette stratégie, notamment dans le choix des niveaux taxonomiques pour prendre en compte leurs spécificités.

4.3.1.1 Sélection des archées 'modèles'

Pour les Archées (*Tableau* 4-2), nous avons sélectionné une espèce par ordre. Au vu de la répartition taxonomique des archées, cela revient souvent à une espèce par classe, voire par embranchement. Dans deux cas, nommément l'ordre des Halobactériales et celui des

Thermococcales, nous avons fait exception en y sélectionnant deux espèces en fonction de leur importance pour la communauté scientifique. Nous avons aussi choisi d'inclure dans notre panel, les deux seules représentantes du groupe DPANN, malgré leur appartenance à un même ordre, en raison du peu d'espèces disponibles pour ce groupe. Notre sélection finale comprend ainsi 31 Archées modèles sur les 179 disponibles dans les protéomes de référence.

Tableau 4-2 **Choix des organismes modèles pour les Archées**. Nombres d'organismes choisis selon l'embranchement, la classe et l'ordre. Les espèces dont la classification est incertaine sont grisées.

	Embranchement	Classe	Ordre	Modèles	Total
DPANN	Nanoarchaeota	Nanoarchaeota	Nanoarchaeales	2	2
	Euryarchaeota	Archaeoglobi	Archaeoglobales	1	6
		Hadesarchaea		1	4
		Halobacteria	Halobacteriales	2	11
			Haloferacales	1	6
			Natrialbales	1	9
		Methanobacteria	Methanobacteriale	1	15
		Methanococci	Methanococcales	1	5
		Methanomicrobia	Methanocellales	1	2
			Methanomicrobiales	1	12
			Methanosarcinales	1	13
			Unclassified	0	1
		Methanopyri	Methanopyrales	1	1
		Thermococci	Thermococcales	2	10
		Thermoplasmata	Methanomassiliicoc.	1	6
			Thermoplasmatales	1	15
			Unclassified	0	1
		Unclassified		1	4
TACK group	Korarchaeota			1	1
	Crenarchaeota	Thermoprotei	Acidilobales	1	2
			Desulfurococcales	1	12
			Fervidicoccales	1	1
			Sulfolobales	1	6
			Thermoproteales	1	12
	Thaumarchaeota	Cenarchaeles	Cenarchaeaceae	1	1
		Nitrosopumilales	Nitrosopumilaceae	1	2
		Unclassified		1	3
	Bathyarchaeota			1	8
ASGARD group	Lokiarchaeota			1	1
	Thorarchaeota			1	3
	Unclassified			0	1
	Environnementales			0	3
TOTAL				**31**	**179**

4.3.1.2 Sélection des bactéries 'modèles'

En ce qui concerne les Bactéries, vu le nombre de protéomes et leur diversité taxonomique, une représentation équilibrée des branches taxonomiques était difficile à atteindre. Cela est

compliqué par l'importante asymétrie dans le nombre d'espèces entre les embranchements. Nous avons sélectionné les espèces de façon à avoir au moins un représentant pour chaque embranchement et pour chaque classe et ordre quand cela était possible. Le positionnement taxonomique de certaines Bactéries étant indéfini, nous avons choisi de ne pas les considérer pour la sélection d'organismes modèles, à l'exception notable de *Kinetoplastibacterium galatii TCC219,* une Betaprotéobactérie endosymbionte dont le génome réduit peut avoir un intérêt pour la génomique comparative. Certaines divisions ont dû être écartées en raison d'un faible nombre de représentants, tous de qualité insuffisante au vu des indicateurs mentionnés plus haut. Pour autant, l'ensemble des espèces 'modèles' choisies représentent 124 des 146 ordres couvrant nos 3863 espèces bactériennes initiales. Dans la majorité des cas, notre sélection se bornait à choisir un représentant par ordre en s'appuyant sur des indicateurs de qualité, notamment le nombre de protéines ayant fait l'objet de revue dans SwissProt. Nous avons fait des exceptions en conservant plusieurs espèces pour certains ordres comptant un grand nombre de représentants (nommément, les Lactobacillales (205 espèces), Bacillales (210), Micrococcales (149), Corynebacteriales (121), Enterobacterales (81), Burkholderiales (151) et Rhizobiales (148)) et surtout les ordres comptant plusieurs espèces d'intérêt biologique et/ou médical (par exemple, pour les Enterobacteriales, la sélection comprend *Escherichia coli* (souche K12), *Salmonella enterica* (salmonellose), *Yersinia pestis* (peste) et l'endosymbiote *Buchnera aphidicola*). En définitive, et selon les différents critères entrant en jeu, notre sélection de protéomes modèles de Bactéries comprend 142 espèces (Tableau 4-3).

Tableau 4-3 **Choix des organismes modèles pour les Bactéries**. Les embranchements et les classes sont représentés, ainsi que la proportion d'ordres représentés par au moins une espèce modèle (en vert si tous les ordres sont représentés). Les divisions pour lesquelles plusieurs espèces par ordre ont été choisies sont en gras.

Groupe	Embranchement	Classe	Modèles	Total	Ordres
	unclassified		0	308	
	Thermotogae	Thermotogae	3	12	3/3
	Thermodesulfobacteria	Thermodesulfobacteria	1	3	1/1
Terrabacteria group	unclassified		0	1	
	Tenericutes	Mollicutes	3	70	3/3
	Firmicutes	Unclassified	0	45	
		Tissirellia	1	21	1/1
		Negativicutes	3	37	3/3
		Limnochordia	1	1	1/1
		Erysipelotrichia	1	21	1/1
		Clostridia	4	366	4/4
		Bacilli	**6**	**416**	2/2
	Deinococcus-Thermus	Deinococci	2	17	2/2
	Cyanobacteria		6	80	6/7
	Chloroflexi	Unclasssified	0	4	
		Thermomicrobia	1	2	1/2
		Ktedonobacteria	0	1	0/1
		Dehalococcoidia	1	5	1/2
		Chloroflexia	1	4	1/2
		Caldilineae	1	1	1/1

		Ardenticatenia	0	1	0/1
		Anaerolineae	1	9	1/1
	Armatimonadetes	Unclasssified	0	1	
		Fimbriimonadia	0	1	0/1
		Chthonomonadetes	1	1	1/1
	Actinobacteria	Thermoleophilia	1	2	1/1
		Rubrobacteria	1	2	1/1
		Coriobacteriia	2	31	2/2
		Actinobacteria	**17**	**534**	14/17
		Acidimicrobiia	1	7	1/1
	Synergistetes	Synergistia	1	11	1/1
	Spirochaetes	**Spirochaetia**	**4**	**33**	3/3
PVC group	Verrucomicrobia	Unclassified	0	4	
		Verrucomicrobiae	1	4	1/1
		Spartobacteria	1	3	
		Opitutae	1	8	1/2
	Planctomycetes	Planctomycetia	1	15	1/1
		Phycisphaerae	0	1	0/1
	Lentisphaerae	Lentisphaeria	0	1	0/1
	Chlamydiae	Chlamydiia	1	9	1/1
Proteobacteria		Unclassifed	0	1	
		Zetaproteobacteria	1	1	1/1
		Gammaproteobacteria	**19**	**495**	16/17
		Epsilonproteobacteria	1	41	1/2
		Deltaproteobacteria	8	99	8/8
		Betaproteobacteria	**11**	**226**	9/9
		Alphaproteobacteria	**10**	**474**	8/10
		Acidithiobacillia	0	4	0/1
	Nitrospirae	Nitrospira	1	10	1/1
	Nitrospinae	Nitrospina	0	3	0/1
	Fusobacteria	Fusobacteriia	1	15	1/1
FCB group	Gemmatimonadetes	Gemmatimonadetes	1	4	1/1
	Fibrobacteres	Fibrobacteria	1	1	1/1
		Chitinivibrionia	0	1	0/1
		Chitinispirillia	0	1	0/1
	Other/Unclassified		0	9	
	Ignavibacteria	Ignavibacteriae	1	2	1/1
	Chlorobi	Chlorobia	1	13	1/1
	Bacteroidetes	Unclassified	0	10	
		Sphingobacteriia	1	24	1/1
		Flavobacteriia	**2**	105	1/1
		Cytophagia	1	44	1/1
		Chitinophagia	1	11	1/1
		Bacteroidia	2	141	2/2
	Elusimicrobia	Endomicrobia	0	2	0/1
		Elusimicrobia	1	1	1/1
	Dictyoglomi	Dictyoglomia	1	1	1/1
	Deferribacteres	Deferribacteres	1	6	1/1

Chrysiogenetes	Chrysiogenetes	1	1	1/1
Caldiserica	Caldisericia	1	1	1/1
Aquificae	Aquificae	2	11	2/2
Acidobacteria	Solibacteres	1	1	1/1
	Blastocatellia	1	2	
	Acidobacteriia	1	6	1/1
	Unclassified	0	3	
Total		**142**	**3863**	**124/146**

4.3.1.3 Sélection des eucaryotes modèles

Pour les Eucaryotes, les divisions taxonomiques sont hétérogènes et beaucoup d'entre elles ne correspondent pas à des embranchements, classes ou ordres (Amoebozoa, Apusozoa…). En absence de définition taxonomique précise, le Tableau 4-4 montre les différents clades que nous avons considérés (au niveau de l'embranchement ou plus) ainsi que le nombre d'espèces modèles pour chacun de ces clades. Lorsque les classes étaient renseignées, nous avons cependant fait en sorte de sélectionner au minimum une espèce par classe ; à titre d'exemple, chez les Champignons les espèces modèles choisies représentent chacune des 29 classes présentes dans les protéomes de référence. Les sélections ont été faites dans un souci de diversité taxonomique, mais comme pour les Bactéries, nous avons pris en considération l'importance scientifique des espèces dans la recherche expérimentale, ce qui entraîne chez les Eucaryote une surreprésentation de certains clades dans notre sélection d'espèces modèles en fonction de leur proximité avec l'Homme. Le clade des Mammifères euthériens regroupe ainsi 6 espèces modèles relativement proches taxonomiquement (*Homo sapiens*, *Pan troglodytes*, *Mus musculus*, *Rattus norvegicus*, *Canis lupus familiaris* et *Bos taurus*).

Pour résumer, la sélection d'organismes modèles pour chaque domaine a été réalisée dans un souci de représenter équitablement les différentes divisions taxonomiques et d'inclure des espèces d'intérêt pour la communauté scientifique. Cette sélection de 317 organismes modèles comprend un nombre équivalent de bactéries et d'eucaryotes, respectivement 142 et 144, et 31 espèces d'Archées. Cette liste relativement réduite comparée à la totalité des protéomes de référence (environ 6%) reste représentative des grandes divisions taxonomiques, elle est donc adaptée aux études de génomique comparative à l'échelle du Vivant tout en restant accessible en termes de données à traiter. Pour cette raison, elle occupe une place centrale dans l'architecture d'OrthoInspector.

Tableau 4-4 **Choix des organismes modèles chez les Eucaryotes**. Les espèces sont classées selon les divisions taxonomiques auxquelles elles appartiennent (embranchement ou plus large).

	Embranchement	Modèles	Initial
Alveolata	Apicomplexa	6	35
	Chromerida	1	1
	Others	2	7
Amoebozoa		3	8
Apusozoa		1	1
Cryptophyta		1	1
Euglenozoa		4	18
Fornicata		1	2
Haptophyceae		2	2
Heterolobosea		1	1
Metazoa	Placozoa	1	1
	Porifera	1	1
	Cnidaria	2	2
	Platyhelminthes	2	14
	Nematoda	4	32
	Arthropoda	12	42
	Annelida	1	1
	Mollusca	2	2
	Echinodermata	1	1
	Chordata	21	62
Fungi	Blastocladiomycota	1	1
	Chytridiomycota	3	4
	Cryptomycota	1	1
	Entomophthoromycota	1	1
	Glomeromycota	1	1
	Mucoromycotina	1	8
	Microsporidia	3	16
	Ascomycota	18	269
	Basidiomycota	10	82
	Unclassified	0	1
Autres Opisthokontes		1 1 2	1 1 2
Parabasalia		1	1
Rhizaria		1	2
Rhodophyta		2	2
Stramenopiles	Bacillariophyta	2	3
	Eustigmatophyceae	1	2
	Phaeophyceae	1	1
	Others	6	15
Viridiplantae	Chlorophyta	3	11
	Streptophyta	15	52
TOTAL		**144**	**711**

4.3.2 Organisation de la base de données : des modules complets autour d'un axe central

Du fait de leur spécificité par rapport à l'ensemble des données, nous avons fait le choix d'organiser les données d'OrthoInspector 3.0 autour des protéomes modèles, le but étant qu'ils soient les points d'accès principaux aux prédictions d'orthologie. Du fait de notre étape de sélection, ces protéomes sont plus équilibrés en termes taxonomiques que l'ensemble de protéomes au niveau des trois Domaines du Vivant, et évitent ainsi, au moins en partie, des biais de surreprésentation. Ils sont, par conséquent, particulièrement adaptés pour servir de base à des analyses de génomiques comparatives entre les domaines. Pour cette raison, nous avons utilisé exclusivement ces protéomes modèles pour construire une base de données d'orthologie dite inter-domaines.

Par opposition, l'exhaustivité de nos protéomes de référence convient parfaitement à des analyses plus fines, à l'échelle des classes et des ordres, pour lesquels on dispose de nombreux représentants. Nous avons donc également construit trois bases de données spécifiques à chaque domaine en utilisant l'ensemble des protéomes de référence associés.

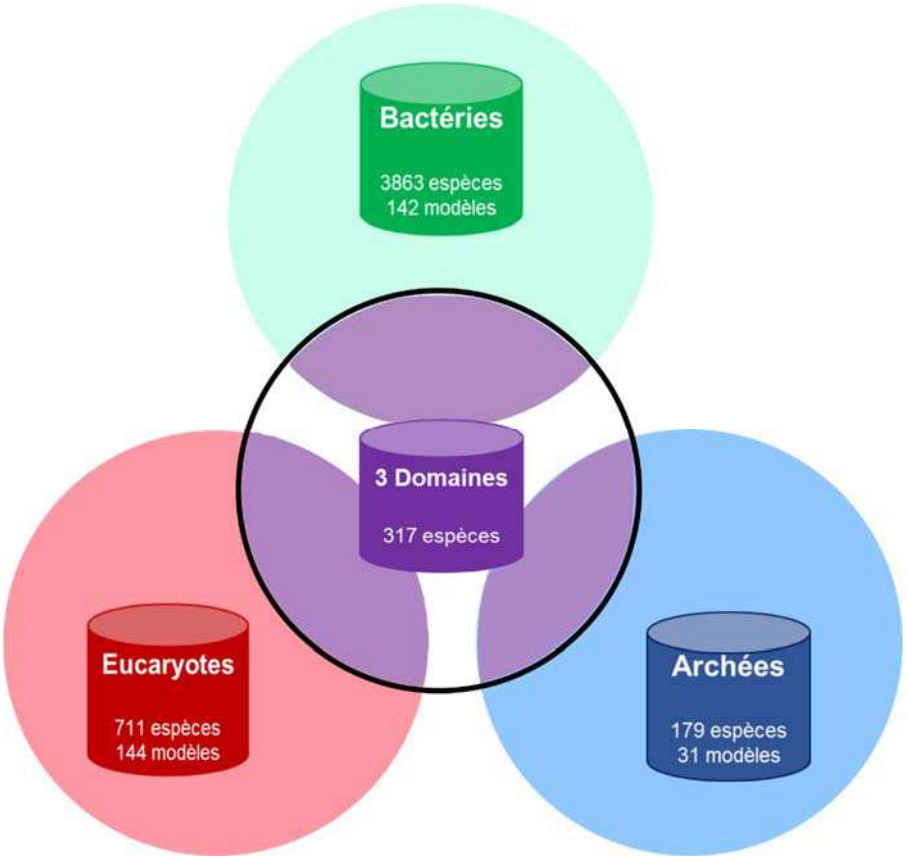


Figure 4-6 **Architecture de données d'OrthoInspector 3.0**. La ressource repose sur quatre bases complémentaires, trois bases de données 'exhaustives' et spécifiques à chaque domaine, et une base de données inter-domaines avec seulement des espèces modèles de chacun des trois domaines.

Ces quatre bases de données couvrent l'ensemble des cas d'utilisations principaux d'une ressource d'orthologie. Les organismes modèles permettent un accès synthétique aux relations d'orthologie entre les espèces les plus étudiées et pour lesquelles on réalise le plus souvent des transferts d'annotations, la base de données inter-domaines permet des analyses de génomique comparative à travers les Domaines du Vivant et les bases de données spécifiques de s'intéresser aux relations d'orthologie à un niveau plus fin de granularité

En regroupant un sous-ensemble d'espèces de chacune des bases intra-domaine, la base de données inter-domaines permet également de faire le lien entre les trois autres bases de données (Figure 4-6). En effet bien que les relations d'orthologie entre deux organismes non-modèles A et B provenant de domaines différents ne soient pas directement évaluées, il est possible de les retrouver par transitivité. Il s'agit d'identifier les orthologues de A dans l'organisme modèle le plus proche A' et d'identifier les orthologues de cette dernière dans un organisme B', modèle proche de B. Finalement, en connaissant les relations d'orthologie de B avec B', on permet ainsi le transfert de relations (Figure 4-7).

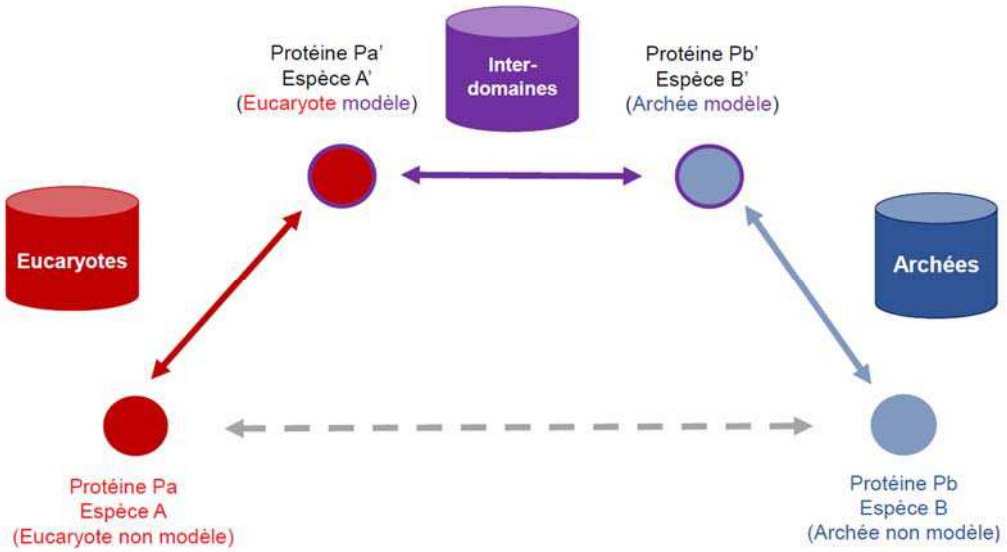


Figure 4-7 **Prédiction par transitivité entre espèces non-modèles.** La relation d'orthologie entre Pa et Pb (en pointillé), deux protéines d'espèces non modèles Eucaryotes et Archées, n'est pas directement disponible dans OrthoInspector 3.0. On peut les retrouver par transitivité en passant par des orthologues dans les espèces modèles proches.

D'un point de vue technique, une telle organisation a un avantage considérable dans le contexte des données massives. On peut, indirectement retrouver les relations d'orthologie sans calculer l'intégralité des relations entre l'ensemble des espèces. Le nombre de relations s'accroissant de manière quadratique avec le nombre de protéines, cela a une importance tant au point de vue de la génération des données que de leur stockage. Faire reposer une même ressource sur des bases de données distinctes implique cependant plusieurs contraintes techniques à prendre en compte lors de sa conception.

- *Identifiants uniques*. Pour garantir l'interopérabilité, les données de chaque base doivent être reconnues par un identifiant unique partagé entre elles. Les bases de données intra comme inter-Domaines reposant sur des protéomes de référence UniProt. Nous avons

choisi de faire reposer cette correspondance et donc, l'ensemble de l'interrogation des données sur les numéros d'accès UniProt, uniques pour chaque protéine.

- *Accès distincts.* La ressource d'orthologie reposant non pas sur une, mais sur de multiples bases de données relationnelles, chacune contenant une partie de l'information, cela doit être pris en compte dans l'interface d'accès aux données. Il s'agit notamment d'éviter d'interroger simultanément les bases de données et créer de potentiels conflits. Pour cela, l'interface doit être conçue de façon à ce que la base de données contenant l'information recherchée soit directement interrogée à chaque étape.

Nous reviendrons sur ce dernier point dans la section consacrée à l'interface du portail d'orthologie.

L'architecture choisie permet jusqu'à un certain degré de faire une économie de ressources informatiques, en termes de puissance de calcul comme de volumes des relations. Ces données restent massives et impliquent la mise en place d'une stratégie de gestion du flux adaptée.

4.4 Construction des bases : gérer les flux de données

Dans leur totalité, les bases de données intra-domaines représentent 12 408 471 protéines pour la base bactérienne, 10 290 183 pour la base eucaryote et 394 539 pour la base archée. La base inter-domaine représente quant à elle 2 874 537 protéines. Les calculs de relations d'orthologie à l'échelle d'autant de protéines nécessitent des adaptations pour les différentes étapes de la prédiction : les comparaisons de similarité tous-contre-tous, les prédictions d'orthologie proprement dites ainsi que les modalités de stockage des données.

4.4.1 Les comparaisons tous-contre-tous

Comme mentionné dans l'introduction, les comparaisons de relations tous-contre-tous reposent sur des programmes tels que BLAST. Leur mise en place nécessite tout d'abord la création d'une base de données indexant l'ensemble des séquences à comparer. Pour chaque séquence, on effectue ensuite une comparaison à la base de données de façon à identifier les meilleurs *hits*. Individuellement, ces opérations sont courtes (de l'ordre de quelques secondes à plusieurs minutes, en fonction du volume de la base de données), mais les réaliser à l'échelle de plusieurs millions de séquences peut prendre un temps considérable. Ces opérations de comparaison n'étant pas dépendantes les unes des autres, il est possible de les paralléliser, c'est-à-dire faire réaliser ces opérations par des unités de calcul différentes, ce qui permet d'effectuer plusieurs opérations simultanément et donc, de diminuer le temps nécessaire d'un facteur égal au nombre d'unités utilisées.

Dans la logique des données massives, nous avons parallélisé les exécutions du programme BLAST à une échelle irréalisable pour une infrastructure informatique classique. Nous avons pour cela utilisé *l'European Grid Infrastructure* (EGI). Ce système de grille de calcul a pour principe de mettre en commun les ressources de plusieurs ordinateurs géographiquement séparés pour effectuer des calculs massivement parallélisés. Pour l'exploitation de cette

infrastructure, le goulot d'étranglement n'est plus réellement le nombre de processeurs disponibles, mais le transfert des données entre les ordinateurs ainsi que l'espace disque disponible sur chaque machine qui ne dépasse pas quelques GO de mémoires dans certains cas. Pour exploiter son potentiel, il nous a donc fallu réduire au minimum la quantité de données à transférer pour chaque étape. Il était par exemple impossible d'utiliser des bases de données BLAST de plusieurs millions de séquences (jusque 6,9 Go pour les données bactériennes).

En prenant en compte ces contraintes, nous avons mis en place un protocole d'exécution des programmes BLAST exploitant le potentiel de la grille, qui se divise en trois étapes (Figure 4-8)

1. *Fragmentation des protéomes*. Pour profiter au maximum de la parallélisation, il s'agit de partitionner les comparaisons tous-contre-tous en tâches indépendantes. Il est bien sûr possible d'effectuer les comparaisons séquence par séquence, mais les opérations étant courtes, les transferts entre machines prennent alors proportionnellement plus de temps que la tâche elle-même. Pour maximiser le ratio calcul/transfert, nous avons donc décidé de fragmenter les protéomes en groupes de 500 séquences.

2. *Fragmentation des bases*. Les bases de données de plusieurs Go peuvent être trop volumineuses pour certains des ordinateurs de la grille. Nous avons fragmenté ces bases de données en partitions de 1Go. Les valeurs *d'Expect* de BLAST dépendant directement de la taille des bases de données interrogées, les commandes BLAST intègrent un paramètre pour indiquer la taille de la base de données totale et rendre ainsi les résultats obtenus sur les différentes portions de bases de données comparables entre eux.

3. *Fusion des résultats.* Pour les comparaisons tous-conte-tous, il est nécessaire de considérer l'intégralité des résultats de BLAST pour chaque séquence. Nous avons donc mis en place une procédure permettant de fusionner les fichiers de résultats individuels obtenus sur des partitions de la banque et d'obtenir, en définitive, un fichier résultat unique.

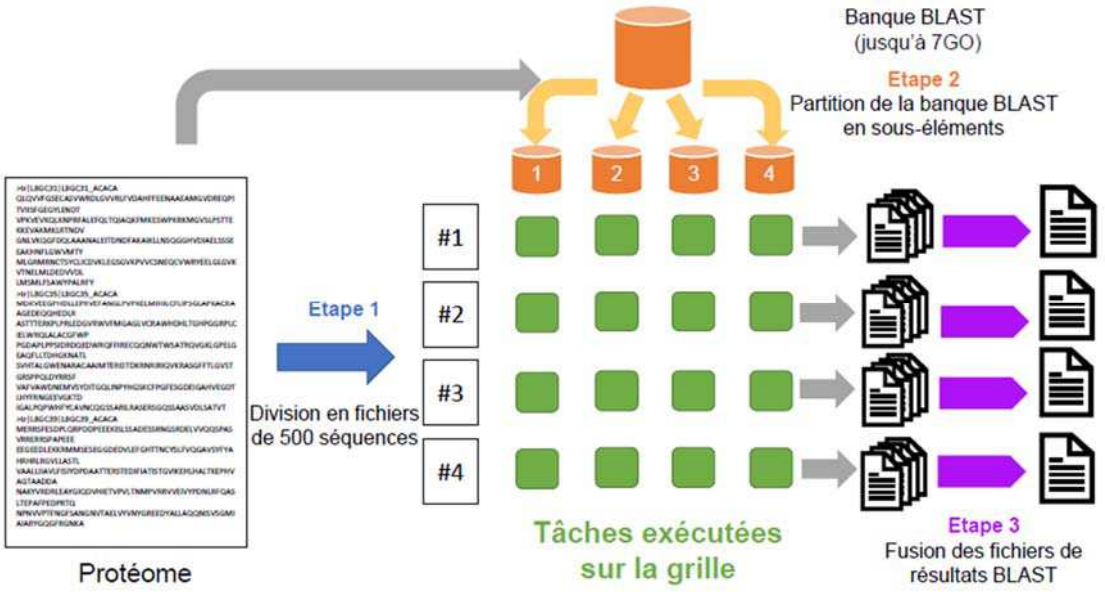


Figure 4-8 **Protocole des comparaisons tous-contre-tous**. Les protéomes sont divisés en fichiers de 500 séquences et les requêtes sont effectuées sur la grille sur des partitions de la base BLAST conçue à partir de l'ensemble des protéomes. Les fichiers issus de ces comparaisons multiples sont ensuite fusionnés.

Les étapes de partition et de fusion des données ont été effectuées localement et la répartition des différentes tâches sur les nœuds (machines) de la grille de calculs a été gérée avec l'orchestrateur Grilladin, développé au laboratoire (Kress, *non publié*). En utilisant ce protocole, l'exécution des BLAST tous-contre-tous nécessaire à la construction des 4 bases de données a été réalisée en deux semaines en utilisant jusqu'à 3 000 cœurs simultanément. Pour comparaison, le temps de calcul cumulé de l'ensemble de cette opération équivaut à 57 ans de calcul en utilisant une seule unité de calcul.

Pour ces problèmes de comparaisons tous contre tous, les solutions de calculs massivement parallélisés sont efficaces pour faciliter la gestion des données massives. Les étapes suivantes de la prédiction des relations d'orthologie ayant d'autres contraintes, elles ont nécessité d'autres adaptations notamment au niveau du programme d'inférence.

4.4.2 Stabiliser OrthoInspector face aux données massives

Les opérations d'inférence d'orthologie par OrthoInspector reposent sur une base de données SQL, dont les tables de données sont interrogées, puis remplies à chaque étape. Il est possible de paralléliser les calculs, ceux-ci étant essentiellement indépendants les uns des autres, à condition toutefois d'avoir accès à la base de données. Aussi, il est impossible d'utiliser des infrastructures de grilles de calcul, la base de données étant trop massive pour être transférée. Nous avons donc décidé de réaliser ces calculs sur notre infrastructure locale.

OrthoInspector ayant été initialement conçus pour des jeux de données de plusieurs centaines de protéomes, nous avons effectué une revue de son code source afin d'identifier les points

pouvant être optimisés et l'adapter au traitement de volumes supérieurs de données. Suite à cette analyse, nous avons effectué plusieurs modifications pour améliorer ses performances :

- *Rationnaliser les insertions dans la base de données SQL*. Dans son mode par défaut, OrthoInspector insère automatiquement, à chaque étape, les données dans la base de données au fur et à mesure de l'exécution. Auparavant, les insertions étaient réalisées pour chaque ligne insérée dans la base, ce qui pouvait ralentir le programme à cause d'un grand nombre d'accès à la base. Nous avons modifié cette procédure de façon à effectuer les insertions par groupe de 500 lignes et donc réduire le nombre d'accès à la base de données.
- *Changer la structure de données.* Le typage explicite des variables utilisées pour enregistrer les relations d'inparalogie, à la fois dans le programme et dans la base de données, reposait sur des nombres entiers dont la valeur maximale est 2 147 483 647. Cette limite est problématique lors de l'analyse de dizaines de millions de protéines répartis entre plusieurs milliers d'espèces, nous avons donc redéfinit le typage des variables correspondantes dans l'ensemble du programme pour autoriser des valeurs plus élevées.
- *Optimisation de l'étape de calcul des relations d'orthologie*. L'étape de calcul des relations d'orthologie utilise les données issues de l'évaluation de relations d'inparalogie et de meilleurs hits. Cette étape nécessite de parcourir les relations pour chaque paire d'organismes. Dans la version précédente, ces relations étaient parcourues une fois de plus que nécessaire avec des accès à la base de données excessifs, nous l'avons donc modifié afin d'éviter des opérations inutiles.

Pour le calcul des quatre bases de données d'orthologie, nous avons utilisé cette version adaptée du programme d'OrthoInspector en utilisant la procédure d'installation parallélisée sur notre infrastructure informatique locale.

4.4.3 Les données d'orthologie : analyse globale

A l'issue de l'installation, les bases de données d'orthologie que nous avons obtenues couvrent presque 6 milliards de relations d'orthologie, pour un volume cumulé de 5,4 To de données (*Tableau 4-5*). Comme l'on peut s'y attendre, étant donné le nombre d'espèces qu'elle représente, la base de données dédiée aux bactéries est la plus massive de l'ensemble, avec un volume de 4,3 To.

Tableau 4-5 **Contenu des bases de données finales d'OrthoInspector 3.0** La taille des bases des données croit de façon non linéaire avec le nombre de protéines et d'espèces.

	Inter-domaines	Eucaryotes	Bactéries	Archées	Total
Espèces	317	711	3 863	179	4753
Protéines	2 874 537	10 290 183	12 408 741	394 539	23 093 463
Relations	50 983 688	716 721 302	5 167 119 581	9 855 468	5 944 680 039
Inparalogues	19 679 837	227 482 373	998 167 303	1 533 447	1 246 862 960
Taille	251 Go	814 Go	4,3 To	9 Go	5,4 To

A partir de cette masse de données, il est possible d'effectuer des comparaisons des répertoires de gènes à l'échelle des Domaines ou même du Vivant pour avoir une vision globale des données générées. La Figure 4-9 présente la proportion de protéines de chaque espèce présentant au moins un orthologue dans toutes les autres. Les espèces dont le répertoire de gènes a peu varié, que ce soit en gain ou en perte, depuis leur ancêtre commun ont ainsi une correspondance proche de 100%.

Sans surprise, les 3 domaines du Vivant apparaissent distinctement, leurs représentants ayant plus de gènes orthologues entre eux qu'avec les espèces d'autres domaines. Ainsi les Bactéries partagent, en moyenne environ 30% de gènes orthologues avec les Bactéries contre 13,7% avec les Archées et 17% avec les Eucaryotes. Les Archées présentent des résultats assez similaires, 33% entre elles, 19% avec les Bactéries et 20% avec les Eucaryotes. Les Eucaryotes se démarquent de cette tendance avec 25,9% de leurs gènes présentant des orthologues entre eux et seulement 5% avec les Bactéries et 4% avec les Archées. Cette différence est à mettre en relation avec la taille importante des répertoires de gènes des Eucaryotes par rapport aux deux autres domaines.

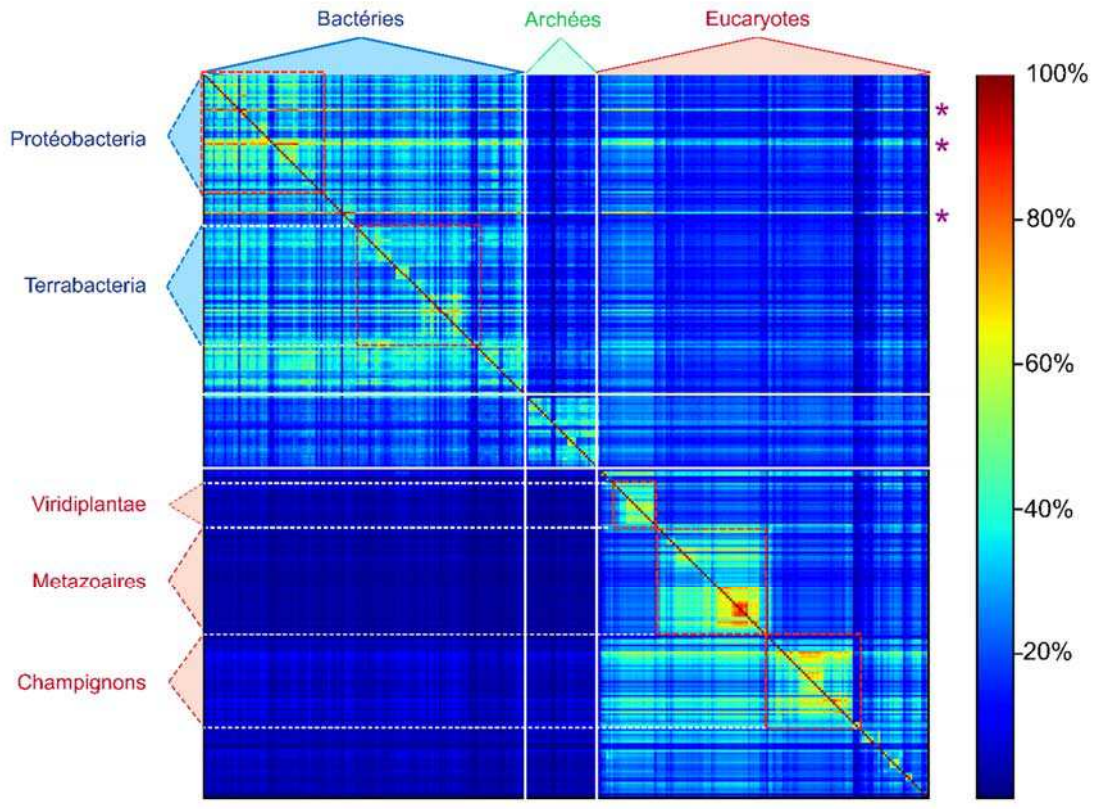


Figure 4-9 **Comparaison des répertoires de gènes des espèces de la base inter-domaine**. Chaque ligne de cette *heatmap* indique la proportion des gènes d'une espèce ayant un orthologue dans chacune des autres espèces (en colonne) selon le code couleur indiqué en légende. Les annotations indiquent les domaines et les grandes divisions taxonomiques auxquels chaque espèce appartient. Les astérisques violets, à droite de la figure, indiquent les lignes dont une forte proportion de gènes ont des orthologues dans toutes les autres espèces.

Cette tendance à partager une plus grande proportion de son répertoire de gènes avec les espèces proches se retrouvent également au niveau des grandes divisions à l'intérieur des domaines, dont les plus importantes en termes de nombre d'espèces sont également indiquées sur la Figure 4-9. Elle est d'autant plus accentuée aux bas niveaux taxonomiques. Ainsi les 6 espèces de Mammifères que nous avons sélectionnées comme espèces modèles partagent plus de 80% de leur répertoire de gènes. Ce clade ressort particulièrement sur la figure (zone rouge au niveau des Métazaoires), ce qui illustre la surreprésentation relative des clades proches de l'Homme dans notre échantillonnage d'espèces modèles.

Certaines espèces de Bactéries présentent une forte proportion de protéines homologues avec les autres Bactéries et d'une façon générale, avec l'ensemble des autres espèces. Ces Bactéries représentées par un astérisque sur la Figure 4-9 sont, de haut en bas, *Buchnera aphidicola* (65,7% avec les Bactéries en moyenne, 49,9% avec toutes les espèces), *Kinietoplastibacterium galati* (65,1%, 49,1%) et *Sulcia muelleri* (69,1%,54%). Ces trois espèces sont des endosymbiontes reconnus et ont la particularité d'avoir un répertoire de gènes particulièrement réduit. Ainsi leurs protéomes ont respectivement une taille de 572, 726 et 226 protéines, ce qui résulte de nombreuses pertes de gènes. Les gènes retenus chez ce genre d'espèces correspondent à un ensemble minimal de gènes domestiques, essentiels à la survie, ce qui explique qu'une part considérable d'entre eux ait des orthologues à travers l'ensemble du Vivant.

Cette figure illustrant la masse de données générées et la diversité des espèces présentes souligne le besoin d'outils de visualisation et de contextualisation pour pouvoir en extraire des connaissances. C'est ce constat qui a influencé la conception de l'interface d'OrthoInspector 3.0, visant à permettre l'accès à l'ensemble des données d'orthologie.

4.5 OrthoInspector 3.0 : un portail d'accès à l'orthologie

Le portail web d'OrthoInspector (disponible à https://lbgi.fr/orthoinspectorv3/) obéit à la même logique que la conception de l'organisation des données : il s'agit de fournir des données exhaustives, tout en mettant en avant les plus pertinentes et en proposant plusieurs niveaux de granularité. Ces considérations ont orienté notre conception du portail web tant au niveau des modalités d'accès aux données que dans leur représentation.

4.5.1 Parcourir les relations d'orthologie : différents niveaux de granularité

Le portail d'OrthoInspector permet d'accéder aux relations d'orthologie selon plusieurs modalités d'accès, disponibles directement sur la page d'accueil. Deux de ces modalités, la recherche par profil et le profilage fonctionnel (*GO profile*) sont des outils de génomique comparative que je décrirai en détail dans le chapitre suivant. Je décrirai donc ici les deux modalités principales, qui sont plutôt classiques pour une ressource d'orthologie :

- *Par identifiants UniProt*. Une barre de recherche, disponible sur la page d'accueil, mais surtout sur le bandeau visible sur l'ensemble du site, permet de chercher une protéine d'intérêt en utilisant son identifiant ou son numéro d'accès UniProt. Le domaine du

Vivant auquel appartient la protéine doit être précisé à l'avance. La recherche est facilitée par un outil d'auto-complétion qui propose des résultats pertinents en fonction de ce que l'utilisateur écrit.

- *Par similarité de séquences*. Un cadre sur la page d'accueil permet de retrouver les protéines de la ressource présentant la plus grande similarité avec une séquence d'intérêt. Cette recherche lance une recherche de similarité par BLAST dans les bases de données d'OrthoInspector et affiche les 100 meilleurs résultats. Cette option permet ainsi de retrouver des protéines sans connaître leur identifiant UniProt ou d'identifier des homologues proches de protéines absentes des bases OrthoInspector. Outre la page d'accueil, on peut accéder à cette recherche par le menu *Access>Blast search.*

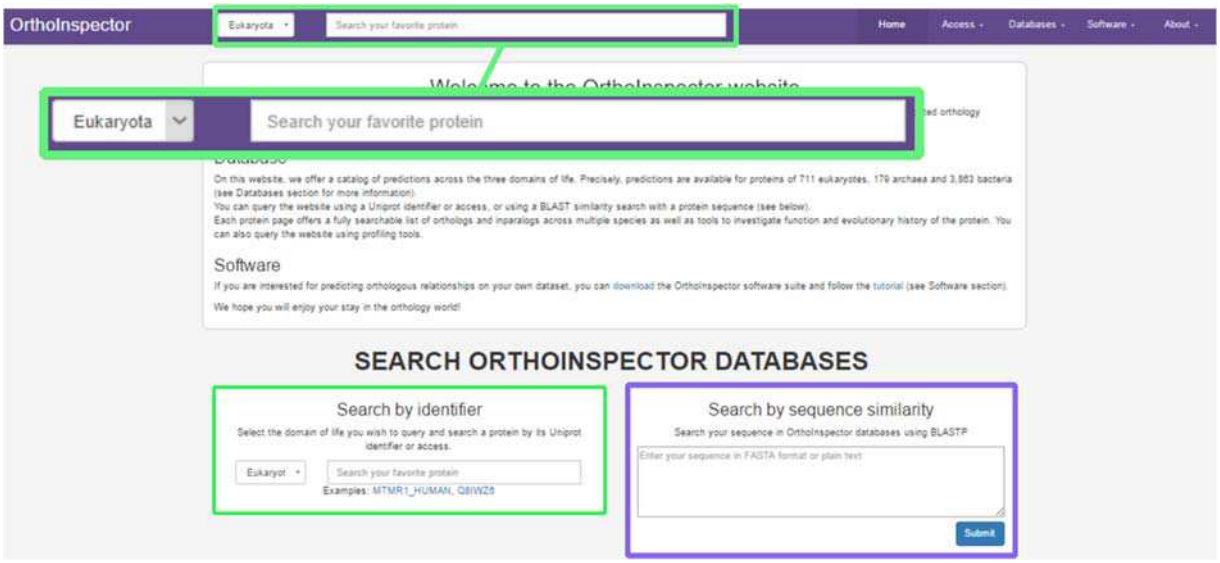


Figure 4-10 **Modalités d'accès à OrthoInspector.** Les deux modes principaux de recherche d'OrthoInspector sont la recherche par identifiant (en vert) et la recherche par similarité de séquences (en bleu), tous deux disponibles sur la page d'accueil.

Ces deux types de recherches donnent accès à une page dédiée à la protéine recherchée et à ses relations d'orthologie. Les relations d'orthologie sont alors accessibles selon trois niveaux de granularité, directement liés à notre architecture de données et donc conçus pour être utiles en fonction du contexte d'intérêt. Ces niveaux sont accessibles sous la forme de trois onglets dans la section « *Orthologs and taxonomic distribution* » : *Model organisms*, *Whole domain* et *Three domains.*

Model organisms

Cette section permet d'accéder aux relations d'orthologie d'une espèce avec les espèces 'modèles' que nous avons sélectionnées dans son domaine : les modèles expérimentaux et les protéomes de qualité couvrant toute la diversité du domaine considéré. Cette section permet ainsi de couvrir la part la plus grande des relations d'orthologie : l'orthologie d'une protéine avec celles d'autres espèces étudiées, relativement proches, et évite de compliquer l'analyse par

un trop grand volume de résultats. Pour cette raison, il s'agit également du point d'accès par défaut.

D'un point de vue technique, ce niveau d'accès repose sur la base d'orthologie complète du domaine étudié dont les résultats sont filtrés en amont en fonction de notre liste d'organismes modèles. Cela nous permet de proposer les relations d'orthologie avec les espèces modèles pour toutes les espèces du domaine considéré, même pour les organismes qui n'ont pas été retenues comme espèce modèle.

Whole domains

Le niveau s'intéressant au domaine complet permet d'accéder, à l'inverse, à l'ensemble des relations d'orthologie d'une protéine dans son domaine du Vivant. Il s'agit du niveau de granularité le plus fin, conçu pour permettre les analyses les plus spécifiques, par exemple celles visant l'ensemble des organismes d'un clade donné. Ce niveau d'accès repose sur les bases intra-domaines.

Three domains

Ce troisième niveau est dédié aux recherches de relations d'orthologie inter-domaines et est donc uniquement accessible pour les protéines d'espèces modèles.

L'implémentation de ces trois niveaux d'accès, reposant sur différentes bases de données, est conçue pour qu'ils s'articulent les uns aux autres avec fluidité grâce à une architecture et des identifiants communs. Ainsi changer de bases de données revient seulement à changer une variable au niveau du serveur, les requêtes qui y sont réalisées sont les mêmes d'une base à l'autre ainsi que le format des résultats. Pour garantir la fluidité entre les bases, les interactions possibles sur le portail suivent des règles spécifiques, qui rendent impossible de solliciter une base de données inadaptée (par exemple interroger la base de données Eucaryotes pour une protéine bactérienne) par erreur.

Premièrement, les options principales de recherche du portail reposent sur des requêtes à l'une des trois bases intra-domaines, qui doivent donc être sélectionnées à l'avance pour plus de rapidité. Dans certains cas, notamment la recherche par similarité de séquences, les trois bases sont interrogées simultanément pour couvrir un panel plus large de résultats possibles. La base de données inter-domaine ne correspondant qu'à un sous-ensemble des trois principales, nous ne la sollicitons jamais directement de cette façon.

Deuxièmement, les accès aux différents niveaux de relations d'orthologie se font par des spécifications directes au niveau de l'URL qui spécifient à la fois la base de données à interroger et le numéro d'accès de la protéine. Ainsi, passer d'une base intra-domaine à la base inter-domaine demande un simple changement de la partie de l'URL relative à la base de données, sans autre modification, les identifiants étant communs.

Troisièmement, les accès se faisant toujours par défaut sur des relations intra-domaines, il est impossible de se connecter par inadvertance à la base de données inter-domaines pour des protéines n'appartenant pas à une espèce modèle. Les seuls liens possibles vers cette base de données n'apparaissent que dans les cas où l'appartenance aux espèces modèles est vérifiée en amont.

Cette organisation spécifique permet ainsi de naviguer dans toutes les relations d'orthologie du portail OrthoInspector 3.0, à différents niveaux de précisions, sans subir de limitation due à la séparation des données en quatre bases.

4.5.2 La représentation de l'information

La représentation des données conditionne les connaissances que l'on peut extraire de l'analyse des relations d'orthologie. Cela comprend la façon dont elles sont remises dans leur contexte biologique et la façon dont on les visualise. Dans OrthoInspector 3.0, les relations d'orthologie de chaque protéine sont visualisables sur la page dédiée à cette protéine. La page se divise en deux volets principaux (Figure 4-11) le premier vise à donner des indications fonctionnelles sur la protéine et contextualiser les données d'orthologie ; le second à représenter celles-ci pour en faciliter l'interprétation. Entre ces deux volets, on retrouve deux sections permettant de retrouver d'autres protéines avec une histoire évolutive similaire. S'agissant d'un outil plus avancé de génomique comparative, je reviendrai en détail sur cette partie au chapitre suivant.

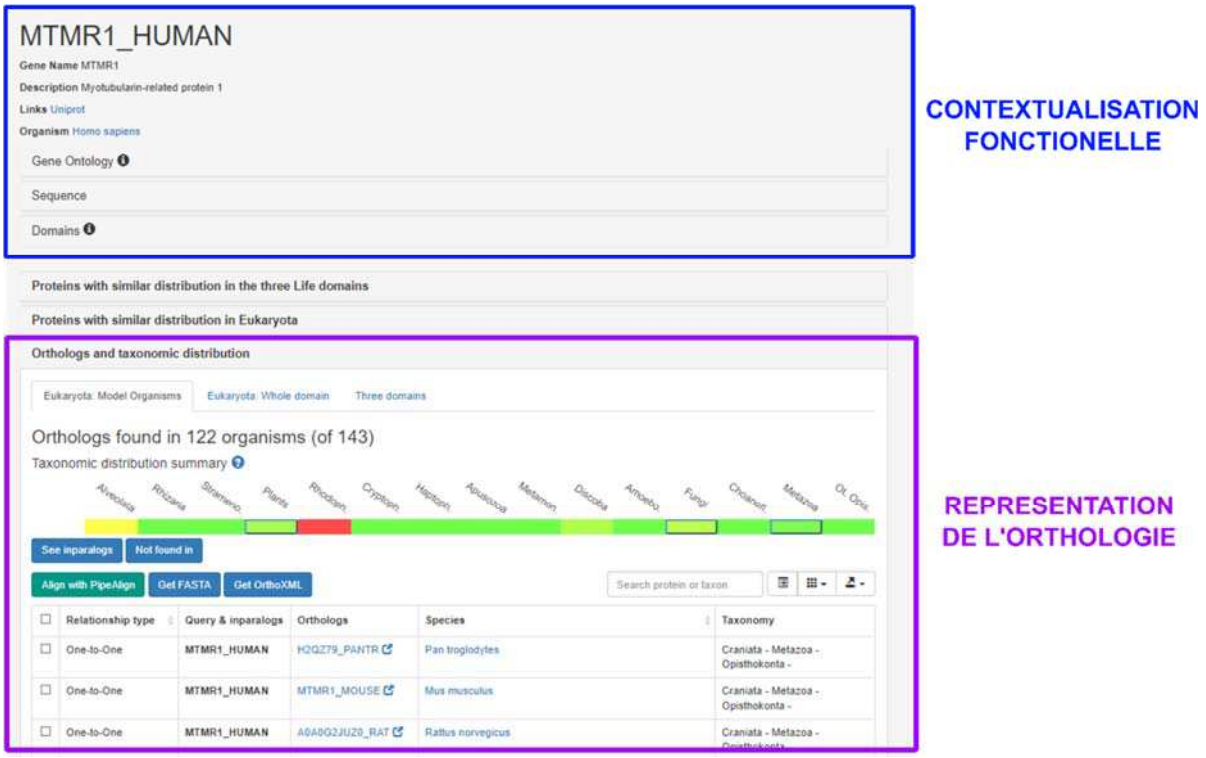


Figure 4-11 **Page de protéine dans OrthoInspector**. Elle se divise en deux parties principales, la contextualisation fonctionnelle de la protéine et une partie dédiée à l'analyse des relations d'orthologie. Les deux volets supplémentaires (non encadrés) permettent de retrouver les protéines présentant une histoire évolutive similaire.

4.5.2.1 La contextualisation fonctionnelle

La contextualisation permet d'analyser les relations d'orthologie d'une protéine au regard d'autres informations disponibles sur celle-ci. Dans OrthoInspector, les informations que nous avons choisi d'intégrer sont principalement fonctionnelles. Elles reposent sur des sources de données externes à OrthoInspector et y sont intégrées dynamiquement.

Les informations fonctionnelles proprement dites se retrouvent dans la description de la protéine, mais surtout dans une section dédiée aux termes *Gene Ontology*. L'ensemble des termes GO des trois catégories (*Molecular function, Biological process, Cellular component*) associés à cette protéine sont affichés avec un lien direct vers la description du terme sur le site QuickGO (https://ebi.ac.uk/QuickGO) (Binns et al., 2009). Ces informations sont extraites en temps réel d'une copie locale de la base de données *Gene Ontology*, dont on prévoit une mise à jour trimestrielle.

Le second type d'informations de contextualisation se réfère à la séquence de la protéine (directement affichable au format FASTA) et à l'architecture en domaines de la protéine. Cette architecture est affichée graphiquement dans la section *Domains* (Figure 4-12), selon la numération en acides aminés de la séquence. La couleur de chaque domaine indique la source des prédictions. On peut visualiser la description fonctionnelle de ces domaines ainsi qu'un lien vers la base de données InterPro en cliquant sur ces représentations schématiques. Toutes ces données sont récupérées en temps réel à partir d'un *webservice* de la base de données InterPro.

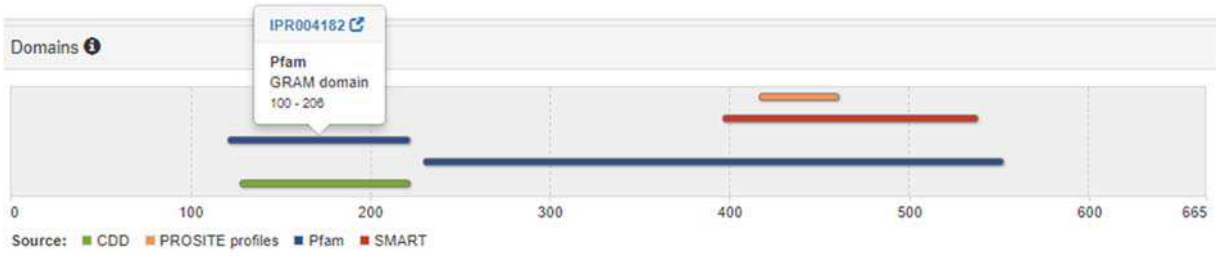


Figure 4-12 **Visualisation de l'architecture en domaines**. Les domaines sont représentés schématiquement selon leur position sur la séquence.

Ces informations permettent d'envisager les relations d'orthologie et leur distribution sous l'angle de la fonction. Selon le contexte dans lesquelles ces protéines sont étudiées, d'autres types d'informations peuvent être pertinentes. OrthoInspector n'est pas conçu comme un *hub* exhaustif d'informations concernant une protéine, aussi nous utilisons l'identifiant UniProt pour permettre un accès vers la page UniProt correspondante où des données plus complètes sont accessibles.

La Figure 4-13 schématise la façon dont OrthoInspector 3.0 utilise d'autres types de ressources pour contextualiser son information. Cet aspect de références croisées illustre l'importance de l'interopérabilité entre ressources pour profiter de la complémentarité des différents types de

données. Dans cette logique, nous avons pris contact avec le consortium UniProt pour que les données d'OrthoInspector y soient également référencées pour ajouter ces relations d'orthologie au paysage des données disponibles.

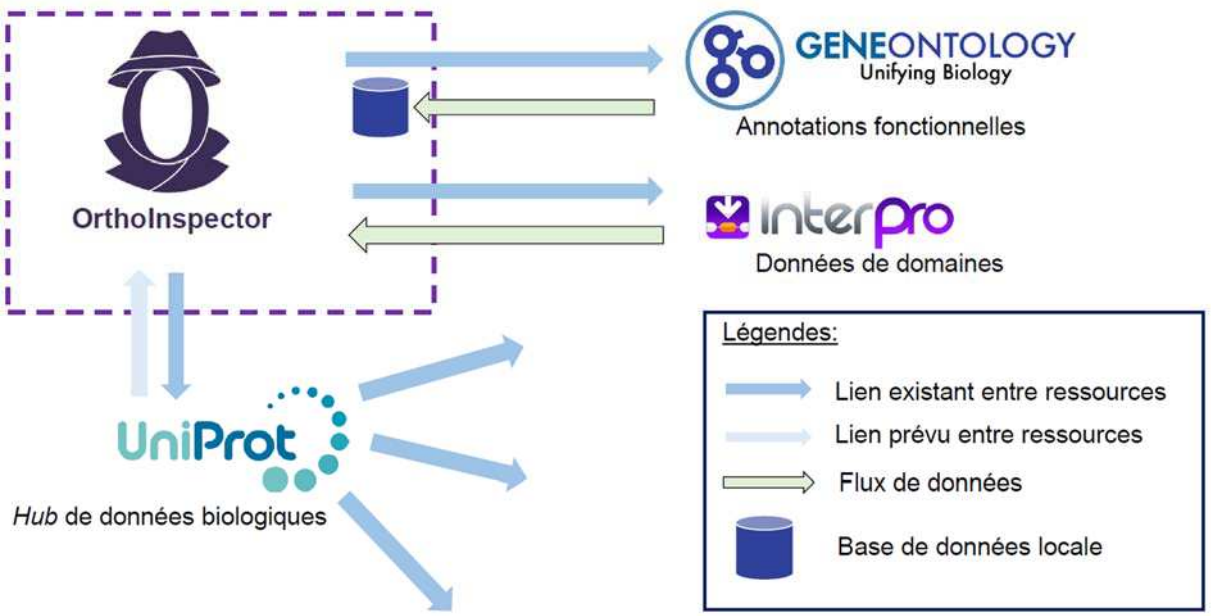


Figure 4-13 **Sources et références de données biologiques dans OrthoInspector**. *Les données fonctionnelles et de domaines protéiques sont récupérées des bases de données GO et InterPro, qui sont également référencées dans OrthoInspector. Les références d'OrthoInspector aiguillent vers UniProt, qui constitue lui-même un hub vers de multiples ressources.*

4.5.2.2 Visualiser l'orthologie

Dans OrthoInspector 3.0, les modalités de visualisation sont majoritairement orientées dans une logique de génomique comparative et sont au nombre de trois : un tableau recensant les relations d'orthologie, l'alignement multiple et la distribution taxonomique. Ce dernier type rentre également dans le cadre de la synthèse d'information, nous l'aborderons donc pas ici mais plus en détail dans la prochaine section.

Le tableau d'orthologie (Figure 4-14) est la méthode principale d'accès aux relations d'orthologie proprement dites. Dans le principe, il s'agit de la liste des relations d'orthologie impliquant la protéine requête. Chaque ligne du tableau correspond à une relation d'orthologie, soit un-à-un, un-à-plusieurs ou plusieurs-à-plusieurs vers une espèce 'cible'. Par défaut, les informations disponibles décrivant la relation comprennent le type de relation, un rappel de l'identifiant de la protéine requête et de ses inparalogues par rapport à l'espèce 'cible' et ses orthologues dans l'espèce cible ainsi que des éléments de classification taxonomique de l'espèce cible (les choix des niveaux taxonomiques seront abordés en détail dans la section suivante).
Malgré l'aspect classique de ce mode de représentation, il est directement conçu comme un outil de génomique comparative, notamment en reposant sur ces notions de classifications taxonomiques. Ainsi, les relations d'orthologie sont, par défaut, ordonnés selon la proximité

taxonomique avec l'espèce 'requête' : par exemple, les premières relations d'orthologie affichées pour l'Homme seront celles avec le gorille ou le chimpanzé. Mis à part les évènements de transfert horizontal, les relations d'orthologie sont classées dans l'ordre des événements de spéciation dont elles sont issues (du plus récent au plus éloigné) et, en utilisant les relations d'inparalogie, peuvent aider à placer les relations de duplications. De plus, nous avons intégré implicitement au tableau l'ensemble des données taxonomiques du NCBI afin que l'option de recherche automatique prenne en considération tous les niveaux taxonomiques. Comme le montre la Figure 4-14, on peut ainsi isoler uniquement les relations d'orthologie concernant un clade donné, ce qui facilite le parcours des données.

Relationship type	Query & inparalogs	Orthologs	Orthologs description	Orthologs length	Species	Taxonomy
One-to-One	AQR_HUMAN	G3RDX8_GORGO	Uncharacterized protein	1480 aa	Gorilla gorilla gorilla	Craniata - Metazoa - Opisthokonta -
One-to-One	AQR_HUMAN	H2Q944_PANTR	Aquarius homolog	1485 aa	Pan troglodytes	Craniata - Metazoa - Opisthokonta -
One-to-One	AQR_HUMAN	H2NMR5_PONAB	Uncharacterized protein	1484 aa	Pongo abelii	Craniata - Metazoa - Opisthokonta -
One-to-One	AQR_HUMAN	G1RK35_NOMLE	Uncharacterized protein	1485 aa	Nomascus leucogenys	Craniata - Metazoa - Opisthokonta -
One-to-One	AQR_HUMAN	F7C652_MACMU	Uncharacterized protein	1527 aa	Macaca mulatta	Craniata - Metazoa - Opisthokonta -
One-to-One	AQR_HUMAN	A0A096NTB5_PAPAN	Uncharacterized protein	1492 aa	Papio anubis	Craniata - Metazoa - Opisthokonta -
One-to-One	AQR_HUMAN	A0A0D9R489_CHLSB	Uncharacterized protein	1492 aa	Chlorocebus sabaeus	Craniata - Metazoa - Opisthokonta -
One-to-One	AQR_HUMAN	G7PAP9_MACFA	Intron-binding protein of 160 kDa	1482 aa	Macaca fascicularis	Craniata - Metazoa - Opisthokonta -
One-to-One	AQR_HUMAN	F7H585_CALJA	Uncharacterized protein (Fragment)	1437 aa	Callithrix jacchus	Craniata - Metazoa - Opisthokonta -
One-to-One	AQR_HUMAN	H0XEG4_OTOGA	Uncharacterized protein	1500 aa	Otolemur garnettii	Craniata - Metazoa - Opisthokonta -

Figure 4-14 **Tableau d'orthologie dans OrthoInspector**. Chaque relation d'orthologie est représentée par une ligne du tableau auxquelles sont associées des informations sur la relation. La description fonctionnelle ainsi que la longueur des protéines orthologues (en bleu) sont des données optionnelles pouvant être ajoutées en configurant l'affichage du tableau. Au-dessus des tableaux, des options permettent d'exporter les relations pour des analyses plus poussées ou les représenter à l'aide d'un alignement multiple dans PipeAlign. Le filtre du tableau permet, entre autres, de configurer l'affichage des relations en fonction de la taxonomie. Ici les relations représentées ne concernent que les Primates.

En plus des informations indiquées par défaut, il est possible d'afficher des informations supplémentaires, à savoir une courte description des protéines orthologues ainsi que leur taille en acide aminés. Ces options apportent du contexte aux relations d'orthologie et peuvent par exemple aider à évaluer la proximité de fonctions entre deux orthologues. Un orthologue ayant une taille sensiblement différente à la protéine requête pourrait, par exemple, avoir un domaine protéique manquant ou supplémentaire et avoir des spécificités fonctionnelles propres.

Le tableau d'orthologie est donc conçu pour permettre un accès flexible à l'ensemble des relations d'orthologie et à en faire une première analyse. Pour une analyse plus poussée, il est

également possible de sélectionner les relations d'orthologie et de les exporter pour permettre leur analyse par d'autres méthodes. Il est ainsi possible de récupérer les relations (l'ensemble ou une partie) au format OrthoXML ou les séquences au format FASTA, en utilisant des boutons situés au-dessus de la table (Figure 4-14). Une dernière possibilité correspond à notre seconde thématique de visualisation et consiste à en réaliser un alignement multiple en utilisant le dernier bouton : « *Align with Pipealign* ».

La fonction d'alignement multiple n'est pas, à proprement parler intégrée au portail d'OrthoInspector, mais passe par celui de PipeAlign2, dédié à la génération d'alignement multiple, également développé au laboratoire. Je n'ai pas été directement impliqué dans le développement de ce site, mais j'ai travaillé sur la compatibilité entre les deux portails, qui rentre dans la logique de conception d'OrthoInspector 3.0. Brièvement, PipeAlign2 repose sur un *workflow* d'alignement intégrant différents outils pour réaliser un alignement multiple à partir d'une ou plusieurs séquences. Les outils utilisés sont :

- DbClustal (Thompson et al., 2000) ou MAFFT (Katoh et al., 2002) : ces programmes construisent l'alignement multiple proprement dit. MAFFT étant plus rapide, il est utilisé par défaut pour les alignements de plus de 100 séquences.
- RASCAL (Thompson et al., 2003) : ce programme détecte les erreurs potentielles dans l'alignement multiple et procède au réalignement des séquences concernées.
- LEON-BIS (Vanhoutreve et al., 2016) : ce programme identifie les régions conservées dans l'alignement et les utilise pour éliminer de l'alignements des séquences ne présentant pas de similarité avec les protéines présentes.
- MASCSIMS (Thompson et al., 2006) : ce programme réalise une annotation automatique de l'alignement à partir des informations des séquences, dont leur architecture en domaines protéiques.

A l'issue de ces différentes étapes, PipeAlign2, permet donc d'obtenir un alignement de haute qualité, qu'il est possible de télécharger. En outre, celui-ci peut également être visualisé dans une interface graphique web permettant une analyse directe des séquences alignées. Grâce à cet outil intégré, il est aisé de passer des tableaux d'orthologie à des comparaisons de séquences et même d'utiliser ces deux outils de visualisations de façon complémentaire.

Le dernier outil de visualisation, original à OrthoInspector 3.0 complète cette offre de visualisation en représentant de façon synthétique l'histoire évolutive des protéines.

4.5.2.3 Synthétiser l'information : la distribution taxonomique

Une représentation synthétique de la distribution des orthologues à travers le Vivant ou d'un domaine doit permettre d'en comprendre rapidement l'histoire évolutive : l'apparition d'un gène, ses duplications et ses pertes. Ces critères ont orienté la création de la bannière de distribution taxonomique qui apparait directement au-dessus de la table de relations d'orthologie.

La façon la plus précise de représenter la distribution d'orthologues est le profil phylogénétique, qui permet de visualiser directement la présence ou l'absence d'un gène espèce par espèce. Cependant, pour plus d'une centaine d'espèces, ce profil devient rapidement illisible et n'est donc pas adapté à une représentation synthétique. Pour notre représentation, nous avons choisi d'utiliser une version synthétique du profil phylogénétique en ne le représentant pas espèce par espèce, mais uniquement par grands groupes taxonomiques (Figure 4-15). Dans ce cadre, ce n'est plus la simple présence ou absence d'orthologues qui est indiquée, mais la proportion des espèces du clade pour lesquelles un orthologue est présent. Celle-ci est représentée par une couleur variant du rouge (absence totale) au vert (présence chez tous les représentants du clade) de façon à être interprétable rapidement de façon visuelle (Figure 4-15).

Pour ce mode de représentation, le choix des clades à afficher est important. En fonction de la protéine étudiée et de l'échelle à laquelle on l'étudie, les divisions taxonomiques pertinentes diffèrent. Nous avons donc choisi différents niveaux de granularité capables, là encore, de s'adapter à plusieurs questions. Dans la logique d'OrthoInspector 3.0, cette problématique de granularité est indissociable de la question de l'architecture de la base de données et des différents niveaux d'accessibilité à l'orthologie. En se basant sur ceux-ci, nous avons opté pour une visualisation synthétique sur 3 niveaux, plus ou moins précis.

Vue inter-domaines

La visualisation inter-Domaines (Figure 4-15) représente le niveau de visualisation le plus large et donc le moins précis. Il s'agit de représenter la distribution au niveau des trois domaines du Vivant avec uniquement les divisions taxonomiques majeures. Cette représentation apparaît, logiquement, lorsque les relations d'orthologie sont visualisées pour l'ensemble des trois Domaines (onglet *Three domains*).

Figure 4-15 **Distribution taxonomique en vue inter-domaines**. Les clades majeurs sont représentés par une tuile annotée dont la couleur représente le nombre d'espèces ayant un orthologue pour la protéine considérée (ici F6HTI7_VITVI, une protéine photosynthétique). Les noms de clades sont colorés en fonction du domaine auquel ils appartiennent.

De fait, nous avons choisi des clades comprenant une proportion importante d'espèces, allant du superembranchement à l'embranchement, et que le consensus de classification taxonomique place en dessous du domaine. Il est à noter que certaines divisions taxonomiques que nous avons sélectionnées pour les Eucaryotes, à savoir les supergroupes SAR et Excavates et le

groupe Archaeplastidae ne correspondent pas à un niveau de la taxonomie du NCBI, qui répertorie pourtant leurs clades enfants. La stabilité de ces clades faisant consensus dans la communauté (Adl et al., 2012), nous les avons intégrés pour disposer des divisions les plus larges possibles. Chez les Eucaryotes et les Bactéries, certaines divisions taxonomiques peu représentées qui ne pouvaient pas être affichées dans le détail à ce niveau, sont indiquées dans des catégories **Autres.**

Vue du domaine

La distribution par Domaine (Figure 4-16) est employée lorsque les relations d'orthologie sont affichées au niveau intra-domaine, que ce soit pour les organismes modèles ou pour l'ensemble des relations d'orthologie (onglets *Model organismes* et *Whole domains*).

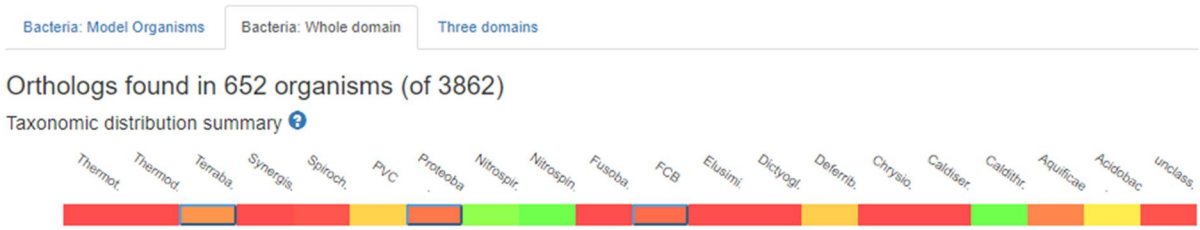


Figure 4-16 **Distribution taxonomique en vue du domaine**. L'exemple donné correspond aux Bactéries. Les modalités de représentation sont les mêmes que pour la vision inter-domaine, mais avec davantage de clades. Les tuiles entourées par un liseré bleu peuvent être cliquées pour accéder à la visualisation de distribution au niveau de ce clade.

Le niveau de granularité étant plus fin, nous avons choisi de présenter des divisions taxonomiques à des niveaux inférieurs. Pour les Eucaryotes, les divisions sélectionnées ne correspondent pas à un rang particulier, mais correspondent généralement aux sous-divisions de celles choisies pour les représentations à Trois Domaines. Nous y conservons cependant le clade des Amœbozoaires, le faible nombre d'espèces ne justifiant pas de le diviser davantage. Là encore, nous avons choisi de représenter des clades absents de la taxonomie du NCBI, mais faisant consensus dans la littérature, nommément Metamonada et Discoba qui représentent deux sous-catégories effectives des Excavés.

Pour les Archées, la logique suit celle adoptée pour les Eucaryotes avec les sous-divisions de celles présentées au niveau des Trois Domaines, ce qui correspond majoritairement au niveau de l'embranchement et de la classe pour les Euryarchées. A ce niveau, les Archées dont la classification n'est pas définie sont représentées dans une dernière catégorie, **Autres Archaea**.

Pour les Bactéries, le niveau du domaine permet de représenter séparément l'ensemble des nombreux embranchements que l'on retrouve au niveau du domaine, regroupés sous la section **Autres Bactéries** au niveau précédant. Ce faisant, on obtient une représentation plus claire. Vue la diversité de chacune des grandes divisions taxonomiques (8 classes dans l'embranchement des Protéobacteries ; 7 embranchements dans le superembranchement des Terrabactéries), il n'est cependant pas possible, à ce niveau, d'en afficher les clades constitutifs.

Les niveaux que nous venons de voir ont été choisis pour représenter équitablement les divisions de chaque domaine de façon à faciliter les analyses de génomique comparative à ce niveau ou au niveau du Vivant. Il est cependant courant d'effectuer des études à des niveaux taxonomiques plus fins, qui ne sont pas couverts dans nos représentations. Nous avons donc prévu un dernier niveau de distribution (Figure 4-17). Ces niveaux supplémentaires sont accessibles pour une sélection de clades, en cliquant sur leurs tuiles (entourées de bleu) de distribution au niveau du domaine (Figure 4-16).

Vue du clade

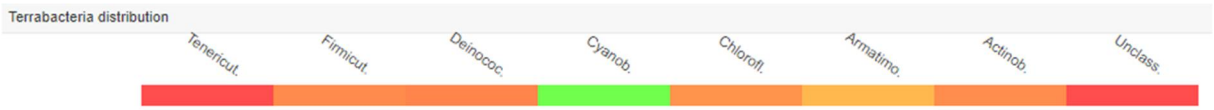


Figure 4-17 **Distribution taxonomique au niveau du clade des Terrabactéries**. Les modalités de représentation sont identiques à celles des niveaux précédents.

Cette action permet de visualiser des divisions taxonomiques inférieures, pour des clades très représentés et particulièrement étudiés. Pour les Eucaryotes, cela correspond aux clades ayant le rang de royaume dans la taxonomie : Champignons (384 espèces, 38 modèles), Viridiplantae (63 espèces, 18 modèles), Métazoaires (158 espèces, 48 modèles). Pour les Bactéries, cela correspond aux grandes divisions que nous n'avons pas pu étendre au niveau précédent : les groupes des Terrabactéries (1681 espèces, 55 modèles) et FCB (367 espèces, 11 modèles) et l'embranchement des Protéobactéries (1341 espèces, 49 modèles).

La Figure 4-18 indique les clades que nous avons choisis de représenter à chaque niveau de visualisation, ainsi que le nombre d'espèces qu'ils incluent.

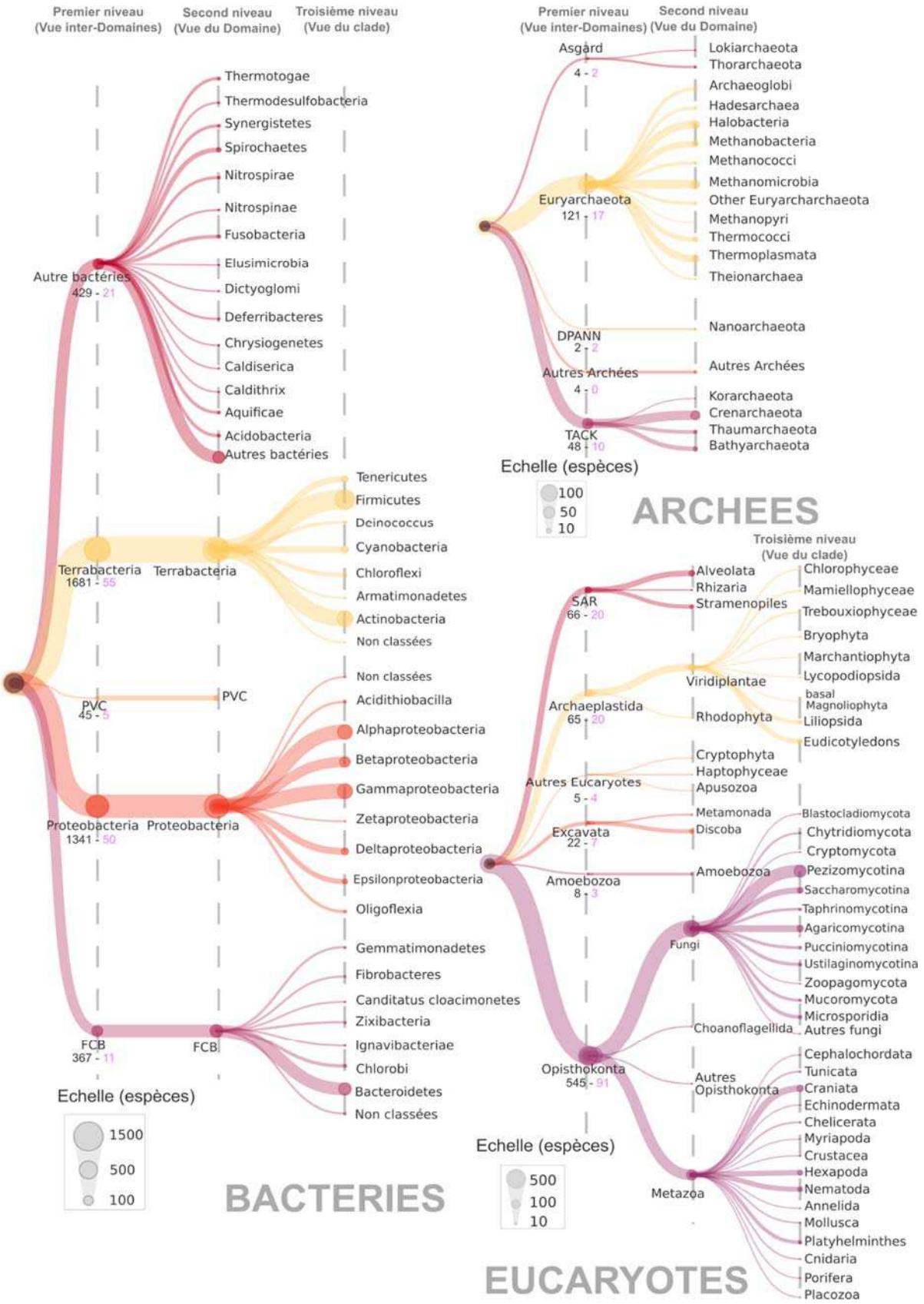


Figure 4-18 **Clades utilisés pour la représentation synthétique de la distribution taxonomique**. Les arbres représentent les clades choisis aux 3 niveaux dans les domaines du Vivant. Le nombre d'espèces pour chaque clade du premier niveau est indiqué sous le nom du clade, le nombre en rose correspond au nombre d'organismes modèles. La taille des nœuds est proportionnelle au nombre d'espèces dans chaque clade dans le jeu de protéomes complets.

La présence ou l'absence relative d'orthologues, accessible à différents niveaux taxonomiques, permet d'évaluer rapidement les événements de gain et de perte de gènes dans l'histoire évolutive d'une protéine. Pour que cette information soit complète, nous avons également ajouté la possibilité d'y retrouver les événements de duplication en ajoutant les données d'inparalogie à cette visualisation synthétique. En cliquant sur le bouton « *See inparalogs* », il est possible de voir les inparalogues de la protéine requête détectés par rapport à chaque clade. La Figure 4-19 donne l'exemple de la protéine DLC1_ARATH (*Endoribonuclease Dicer homolog 1*) d'*Arabidopsis thaliana*. La visualisation à l'échelle des Eucaryotes permet d'examiner la distribution de 4 inparalogues. RTL2_ARATH est une autre ribonuclease considérée comme inparalogue par rapport aux seuls Straménopiles, il s'agit probablement ici d'un homologue distant inféré à tort comme inparalogue. Les 3 inparalogues DCL2_ARATH, DCL3_ARATH et DCL4_ARATH le sont par rapport à l'ensemble des autres clades, ce qui laisse supposer une duplication postérieure à l'ancêtre commun des Plantes. La visualisation à un niveau de granularité plus fin permet de déterminer que ces gènes sont inparalogues par rapport aux clades des Chlorophyceae et des Trebouxiophyceae seulement. On peut donc remonter la duplication à un ancêtre commun des taxa restants, celui du clade des Streptophytes dont ils sont tous des sous-divisions.

Figure 4-19 **Visualisation des inparalogues dans OrthoInspector**. Les inparalogues de la protéine DCL1_ARATH sont représentés par des coches sous les clades correspondants et pour lequel une relation d'orthologie existe. Le profil du haut donne la distribution chez les Eucaryotes, celle du bas, plus précise, chez les Plantes.

Si la représentation synthétique indique les distributions jusqu'au niveau du clade, il est bien sûr possible de retrouver les espèces correspondantes à ces clades dans le tableau d'orthologie. C'est d'ailleurs tous le sens de la colonne *Taxonomy*, qui indique l'appartenance au clade de chacun des trois niveaux de visualisation. Cependant, les espèces pour lesquelles aucun orthologue n'a été détecté n'apparaissent pas. Considérant qu'il s'agit d'une information importante lorsque l'on s'intéresse spécifiquement aux événements de perte de gènes, nous rendons aussi disponible la liste des espèces qui sont dans ce cas de figure, par le biais d'un bouton '*Not found in*' accessible sous la distribution.

4.5.3 Les accès programmatiques

A travers cette description de l'interface d'OrthoInspector, je souhaitais montrer qu'elle est spécifiquement conçue pour faciliter l'analyse manuelle des données d'orthologie. Il est toutefois important de rappeler, surtout à l'ère des données massives, que certaines études de génomique se font à l'échelle de plusieurs centaines de gènes ou que les relations d'orthologie peuvent être utilisées dans le cadre de traitements automatisés.

Pour répondre à ces besoins, nous avons mis en place une interface programmatique REST permettant d'accéder aux relations d'orthologie de façon massive et automatisée, sans passer par l'interface web. L'API ainsi que sa documentation complète, basées sur le *framework* Swagger est disponible sur http://www.lbgi.fr/orthoinspectorv3/API.

Dans la même logique que le portail web, chaque requête nécessite de préciser à laquelle des quatre bases de données principales d'OrthoInspector l'on se réfère. Sans rentrer dans les détails, qui sont présentés dans la documentation, il est possible d'interroger trois types de données des bases OrthoInspector : les espèces présentes dans la base de données, les protéines et leur description et évidemment les relations d'orthologie.
Les relations d'orthologie sont mises à disposition selon des modalités correspondant à différents cas d'étude :

- *L'ensemble des relations d'orthologie d'une protéine*. Il s'agit de l'information disponible sur le portail web. Elle est utile lorsque l'on souhaite étudier une famille de protéines.
- *Les orthologues d'une protéine requête dans une espèce donnée*. Ce cas de figure répond à la recherche de protéines de même fonction dans deux espèces différentes, selon la conjecture d'orthologie.
- *Tous les orthologues entre deux espèces*. Cette requête suit la logique de la précédente dans un contexte de données massives. Elle permet de trouver l'ensemble des correspondances entre deux génomes. Il s'agit de la seule option qui n'est pas accessible par le portail web, car elle prend tout son sens dans le cadre d'analyses automatisées.

Toutes les requêtes de l'API rendent un résultat dans le format standard JSON (*JavaScript Object Notation*), afin de faciliter leur interprétation programmatique.

4.6 Discussion et futures directions

Une plateforme robuste de génomique comparative
Pour conclure, OrthoInspector 3.0 a été conçu comme une ressource complète pour des analyses de génomique comparative. Cela s'est concrétisé par une augmentation importante du nombre de protéomes intégrés ainsi que par une architecture et des modalités de représentations des données permettant l'analyse à différents niveaux de granularité. L'ensemble de la plateforme, comprenant l'exploitation des relations d'orthologie que nous avons vues ici, ainsi que les outils de profilage phylogénétique présentés dans le prochain chapitre, est décrit dans un article à paraître en janvier 2019 dans le volume *Database* de *Nucleic Acids Research* et déjà accessible

dans une version électronique (Nevers et al., 2018). Une copie de cet article est également disponible en annexe de ce manuscrit.

Par rapport aux autres ressources d'orthologie disponibles, OrthoInspector propose, à notre connaissance, la couverture la plus importante en termes d'espèces, tout en respectant des standards de qualité et de diversité. Nos outils de visualisation, combinés aux options plus avancées de génomique comparative qui seront décrites en détail dans le chapitre suivant permettent d'explorer l'ensemble des données de génomique comparative avec une attention particulière portée à la distribution phylogénétique et à son implication en termes d'histoire évolutive. Couplés à l'équilibre de la méthode en termes de spécificité et sensibilité, ces particularités le positionnent comme un outil unique pour réaliser des analyses de génomique comparative à petite comme à grande échelle.

Diversité, qualité et évolution des protéomes

La conception de bases de données et la sélection des protéomes et des espèces modèles illustrent les problématiques posées par le flux important de données en génomique comparative, au niveau de la qualité comme au niveau de la diversité taxonomique.

Certaines divisions taxonomiques sont encore sous-représentées dans les données, le domaine des Archées par exemple, regroupe toujours peu de protéomes comparés aux autres domaines. A un autre niveau, des clades d'intérêt pour des études ciblées comme les Gymnospermes chez les plantes, ne sont pas représentés dans nos bases en raison de l'absence de protéomes de bonne qualité. Ces manques de couverture de certaines divisions taxonomiques sont d'autant plus visibles que d'autres sont surreprésentées. Chez les Bactéries, l'embranchement des Protéobactéries regroupe plus d'un tiers du total des protéomes (1341) alors que d'autres embranchements sont représentés par une seule espèce. Ces différentiels peuvent bien sûr être la conséquence de la diversité réelle de chaque clade mais ils peuvent aussi être le reflet d'un intérêt plus important de la communauté scientifique pour tel ou tel embranchement, il est cependant difficile d'évaluer la part des différentes causes dans cet état de fait.

Dans tous les cas, le flux important de nouveaux protéomes devrait dans les prochaines années permettre d'intégrer de nouveaux embranchements. Ainsi, depuis le début de la conception de nos bases de données en octobre 2016, le nombre de protéomes de Référence UniProt disponibles pour les trois domaines du Vivant est passé de 5443 à 10 123, soit une augmentation de 86%, avec plus du double d'Archées. En conservant nos critères de qualité, cela correspond à 9 110 protéomes de qualité *Tableau 4-6*). On remarque d'ailleurs que 7 protéomes d'Eucaryotes et 6 protéomes de Bactéries que nous avions éliminés sur des critères de qualités ont subi des mises à jour et sont maintenant acceptables selon nos critères. Les améliorations sont particulièrement importantes pour deux protéomes d'intérêt, celui du blé *Triticum aestivum (*3% de petites protéines contre 22% précédemment) et celui du pigeon *Columbia livia* (13% de protéines ne commençant pas par une méthionine contre 56% précédemment, 8932 protéines contre 3892). Il faut noter cependant que 4 des 13 protéomes précédemment éliminés passent les filtres uniquement en raison de modifications mineures d'un des indicateurs de qualité. Cela souligne les limites des filtres basés sur des seuils, qui rendent quelque peu arbitraire la

définition des protéomes de bonne et mauvaise qualité. Les bases de données d'orthologie étant terminées, nous pourrons à l'avenir revisiter l'évaluation de la qualité des protéomes en se basant sur les relations d'orthologie, et le cas échéant, adapter ces seuils.

Tableau 4-6 **Evolution du nombre de protéomes de référence UniProt.**

	Novembre 2016		Octobre 2018	
	Protéomes de référence UniProt	Protéomes filtrés	Protéomes de référence UniProt	Protéomes filtrés
Eucaryotes	830	711	1141	1021
Archées	213	178	447	338
Bactéries	4400	3863	8535	7751
Total	5443	4752	10123	9110

Vers un protocole de mise à jour de la ressource

Pour suivre le rythme de ce flux de données considérable, que ce soit avec l'apparition de nouveaux protéomes de référence ou les modifications de composition en protéines, nous prévoyons des mises à jour régulières des bases de données OrthoInspector. Cela demande la mise en place d'un protocole de réévaluation des relations pour les protéines ayant été modifiées et de prédiction de relations pour les nouvelles protéines, sans pour autant remettre en cause l'ensemble des données générées. Il s'agir donc de procéder en deux temps pour les comparaisons de similarité tous-contre-tous : d'abord une comparaison de toutes les nouvelles protéines et des protéines modifiées avec l'ensemble des protéomes de la base, puis une comparaison ciblée de toutes les protéines détectées lors de cette comparaison avec uniquement les nouvelles protéines et protéines modifiées. Cela permettra d'obtenir les informations de réciprocité pertinentes pour les inférences d'orthologie sans avoir à réaliser l'ensemble des comparaisons. A partir de là, nous adapterons le logiciel OrthoInspector pour permettre des modifications de la base de données à partir de ces données partielles. Il s'agira, pour les nouvelles protéines, de réutiliser l'algorithme principal et pour les protéines modifiées ou supprimées, de commencer par éliminer les relations les concernant.

La mise à jour de la ressource demande bien sûr de prendre en compte l'architecture particulière d'OrthoInspector avec des organismes modèles simultanément présents dans deux bases de données distinctes, la base de données intra-domaine du protéome considéré ainsi que la base de données inter-domaines. Pour ce faire, nous prévoyons des mises à jour concernant spécifiquement les protéomes modèles à un rythme mensuel contre un rythme trimestriel pour les protéomes non modèles des bases de données intra-domaines. Découpler ces mises à jour permettra d'assurer la correspondance entre ces différentes bases de données ainsi que la qualité des données disponibles pour les espèces modèles. En outre, cela facilitera l'ajout de nouvelles espèces modèles en fonction des nouvelles données disponibles. Pour le reste, rien dans l'implémentation du portail d'OrthoInspector 3.0 ne dépend des protéomes actuels et l'ensemble des outils de représentation que nous avons vus sont conçus pour s'ajuster automatiquement à l'ajout de nouvelles espèces.

Développer l'interopérabilité d'OrthoInspector

Comme ce chapitre a pu le montrer, une grande partie de la conception d'OrthoInspector 3.0 s'appuie sur une réflexion portant sur la représentation des données et leur contextualisation avec d'autres données. Ce dernier aspect est essentiel pour permettre l'intégration des données d'orthologie, un descripteur évolutif, avec d'autres données biologiques. Nous fournissons pour l'instant des données issues d'autres bases de données publiques (Gene Ontology, InterPro, UniProt). Les liens directs, utiles pour les analyses manuelles ne sont pas pour autant adaptés à des analyses croisées des différentes ressources à grande échelle. Ce genre d'interrogation de ressources nécessite les technologies du web sémantique. A cet effet, nous prévoyons dans les prochains développements de représenter les données au format RDF, en s'appuyant notamment sur l'ontologie ORTH déjà utilisées par les ressources OMA et MGDB, afin de permettre la mise en place d'une interface SPARQL. Cela permettra les requêtes à travers plusieurs ressources et ainsi, d'utiliser les relations d'orthologie à notre disposition dans une analyse complète.

Ces futurs développements sont dans la continuité de notre volonté d'intégrer des données diverses pour décrire le plus efficacement les systèmes biologiques. Lors de la conception d'OrthoInspector 3.0, notre objectif était de permettre d'extraire de la masse de données que constituent les génomes des nombreuses espèces des marqueurs évolutifs. Sur le portail d'OrthoInspector 3.0, ces marqueurs prennent la forme de représentations synthétiques de l'histoire évolutive des protéines à plusieurs niveaux de granularité. Le soin particulier que nous avons apporté à la sélection des données visait à rendre ces marqueurs évolutifs les plus fiables et efficaces possibles. La base de données complète constitue ainsi un socle stable pour la mise en place d'une plateforme d'analyse de génomique comparative aidant à la compréhension des relations génotype-phénotype. Il s'agit du second axe principal de ma thèse que je décrirai en détail dans le prochain chapitre.

5 Lier l'histoire évolutive à la fonction

La ressource d'orthologie OrthoInspector 3.0 constitue un socle solide, en termes de données et d'outils, pour permettre des analyses de génomique comparative à plusieurs échelles. Ce chapitre couvre le second axe des contributions de cette thèse en présentant les outils basés sur le profilage phylogénétique que j'ai développés et qui sont intégrés au portail d'OrthoInspector 3.0. Le profilage phylogénétique permet d'étudier les relations génotype-phénotype en exploitant le lien entre histoire évolutive et fonction ou trait phénotypique. Nos outils explorent trois facettes du profilage phylogénétique : retrouver des gènes liés à un phénotype donné dont la distribution est connue, étudier un trait phénotypique sous l'angle de l'évolution et finalement, établir des liens entre gènes ou protéines d'une même espèce sur la base de leur histoire évolutive. Cette dernière application permet l'intégration des données évolutives à d'autres types de données, ce qui est illustré par leur contribution au réseau social MyGeneFriends, une plateforme d'analyse originale reliant gènes humains, maladies et chercheurs.

5.1 De l'évolution au phénotype : recherche par profil

L'hypothèse de base du profilage phylogénétique est qu'un gène responsable d'une fonction donnée est perdu ou acquis avec cette dernière au cours de l'évolution. On s'attend ainsi à trouver des orthologues de ce gène chez les espèces où cette fonction existe. En se basant sur cette hypothèse, nous avons conçu un outil permettant d'identifier les gènes d'une espèce requête présentant une distribution phylogénétique donnée (présence d'orthologues dans une sélection d'espèces et absence d'orthologue dans d'autres espèces). Cet outil s'appuie sur les bases de données OrthoInspector et est disponible sur le portail d'OrthoInspector 3.0 à l'adresse : http://www.lbgi.fr/orthoinspectorv3/profile_search.

5.1.1 Conception de l'outil de recherche par profil phylogénétique

La recherche de gènes en fonction d'un profil phylogénétique est basée sur des contraintes de présence ou d'absence pour une liste d'espèces. La façon la plus simple de l'implémenter serait d'identifier les gènes présents dans toutes les espèces présentant un caractère et absents de toutes les autres. Une telle implémentation pose cependant des difficultés lorsque l'on est confronté à une grande quantité de données comme c'est le cas dans OrthoInspector 3.0 :

- Une présence stricte dans toutes les espèces présentant un caractère donné ne prend pas en compte la diversité inhérente de ce caractère chez les espèces le présentant. Il n'est pas à exclure que quelques espèces n'aient pas un gène donné, par erreur de prédiction ou par exception biologique.
- Si assigner manuellement les espèces pour lesquelles un orthologue doit être présent ou absent est envisageable lorsque cette liste ne concerne qu'une dizaine d'espèces, la tâche

est bien moins évidente lorsqu'elle concerne un jeu de données comprenant des centaines, voire des milliers d'espèces.

D'un point de vue évolutif, la présence ou l'absence d'un gène est déterminé par des événements de gain et de perte de gènes, ayant eu lieu le plus souvent chez l'ancêtre commun de plusieurs espèces. Pour cette raison, nous avons choisi d'implémenter les contraintes de présence ou d'absence non pas uniquement, au niveau des espèces, mais au niveau de n'importe quel clade. Ainsi, la contrainte de présence au niveau d'un clade signifie que le gène recherché était présent chez le dernier ancêtre commun des espèces du clade, la présence d'un orthologue dans au moins une des espèces de ce clade suffit à valider cette contrainte. A l'inverse, l'absence imposée au niveau d'un clade signifiant que le gène a été perdu chez l'ancêtre commun des espèces de ce clade, le gène doit donc être absent de toutes les espèces de ce clade pour vérifier la contrainte.

Une fois les contraintes définies, il s'agit d'explorer les données d'orthologie pour retrouver l'ensemble des protéines qui les vérifient. Du fait du mode de représentation des relations d'orthologie dans les bases de données OrthoInspector, identifier les protéines présentant un orthologue dans une sélection d'espèces ne pose pas de difficultés techniques. A l'inverse, comme les seules relations définies dans la base correspondent à la présence d'un ou plusieurs orthologues, il n'existe pas de requête directe pour identifier les protéines répondant à un critère d'absence totale d'un clade. Le moyen de contourner cela est d'inverser les propositions : sont absentes d'un clade, toutes les protéines ne répondant pas à un critère de présence d'orthologue(s) dans ce clade. Ces propositions posées, la méthode de requête développée utilise le principe d'inférence par association soustractive, appliquée à des clades plutôt qu'à des espèces individuelles. On identifie séparément les protéines ayant une relation d'orthologie avec au moins une espèce de chaque clade de la sélection, puis on sélectionne celles répondant à l'ensemble des contraintes de présence. On soustrait ensuite l'ensemble des protéines ayant des orthologues avec des espèces des clades avec la contrainte d'absence (Figure 5-1). Ces différentes règles de combinaison sont implémentées en une seule requête SQL, structurée de manière à s'adapter à un nombre illimité de contraintes (voir Matériel et Méthodes pour les détails).

Cette implémentation peut poser des problèmes en termes de temps d'exécution pour des contraintes complexes incluant un grand nombre d'espèces. Pour limiter l'impact de ces problèmes de temps sur l'utilisation de l'outil, les résultats de chaque requête exécutée avec cet outil sont enregistrés en cache sur le serveur. Ainsi, toute requête déjà effectuée précédemment donnera le résultat de façon immédiate.

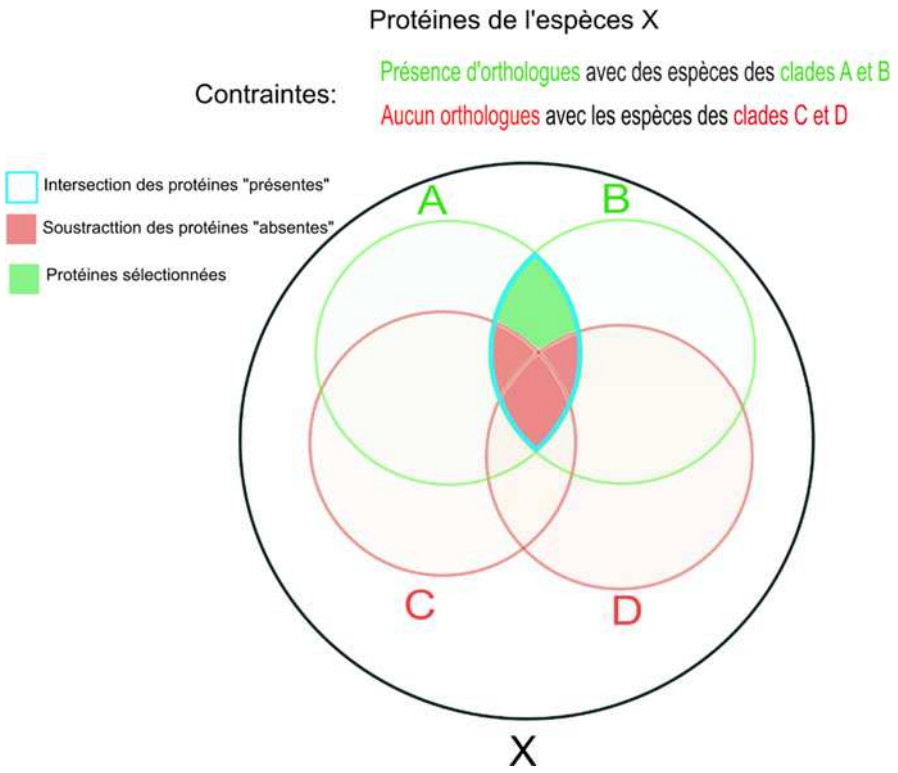


Figure 5-1 **Sélection des protéines respectant un profil phylogénétique.** Dans un premier temps, on retrouve toutes les protéines de l'espèce X ayant des orthologues dans les clades de contraintes de présence (représentées par des cercles verts ici). Les protéines d'intérêt sont à l'intersection des ensembles, en excluant celles appartenant aux clades de contraintes d'absence (représentés par des cercles rouges).

Le choix de l'interface de sélection des contraintes de présence/absence par clades devait répondre à plusieurs exigences : permettre un choix de clade le plus large possible d'une façon simple et facile à parcourir, qui ne soit pas alourdie par la diversité taxonomique disponible. Les clades étant, par définition, les différentes branches de l'arbre de la taxonomie du Vivant, nous avons opté pour un arbre basé sur la taxonomie du NCBI. Pour éviter une surcharge d'information, l'arbre taxonomique est réduit par défaut et il est possible d'accéder un clade d'intérêt par le biais d'une barre de recherche. La même interface suffit à sélectionner l'ensemble des contraintes : un clic sur un taxon définit une contrainte de présence, alors qu'un deuxième clic définit l'absence et un troisième annule toute contrainte. L'interface est conçue pour agir de façon 'intelligente' et éviter les sélections incohérentes : une contrainte d'absence dans un clade conduit à la même contrainte dans ses clades enfants et invalider la contrainte d'absence dans un clade l'invalide également dans ses clades parents (Figure 5-2).

L'idée principale derrière ces choix d'implémentation est de permettre la définition de critères de présence-absence à tous les niveaux taxonomiques. Selon cette même logique, la recherche par profil phylogénétique est disponible pour l'ensemble des bases de données d'OrthoInspector, inter-domaines comme intra-domaines. L'arbre taxonomique où l'on peut réaliser les sélections change selon la base de données d'intérêt. D'un point de vue pratique, la recherche par profil s'effectue en trois étapes :

1. Sélection d'une base de données.

2. Sélection de l'espèce requête pour lequel on cherche les protéines
3. Sélection des contraintes de présence et d'absence dans l'arbre taxonomique.

Cet outil de recherche permet d'obtenir la liste de l'ensemble des protéines de l'espèce ayant un profil donné. Il s'agit ensuite d'offrir les outils d'analyse de cette liste, en la contextualisant en termes fonctionnels.

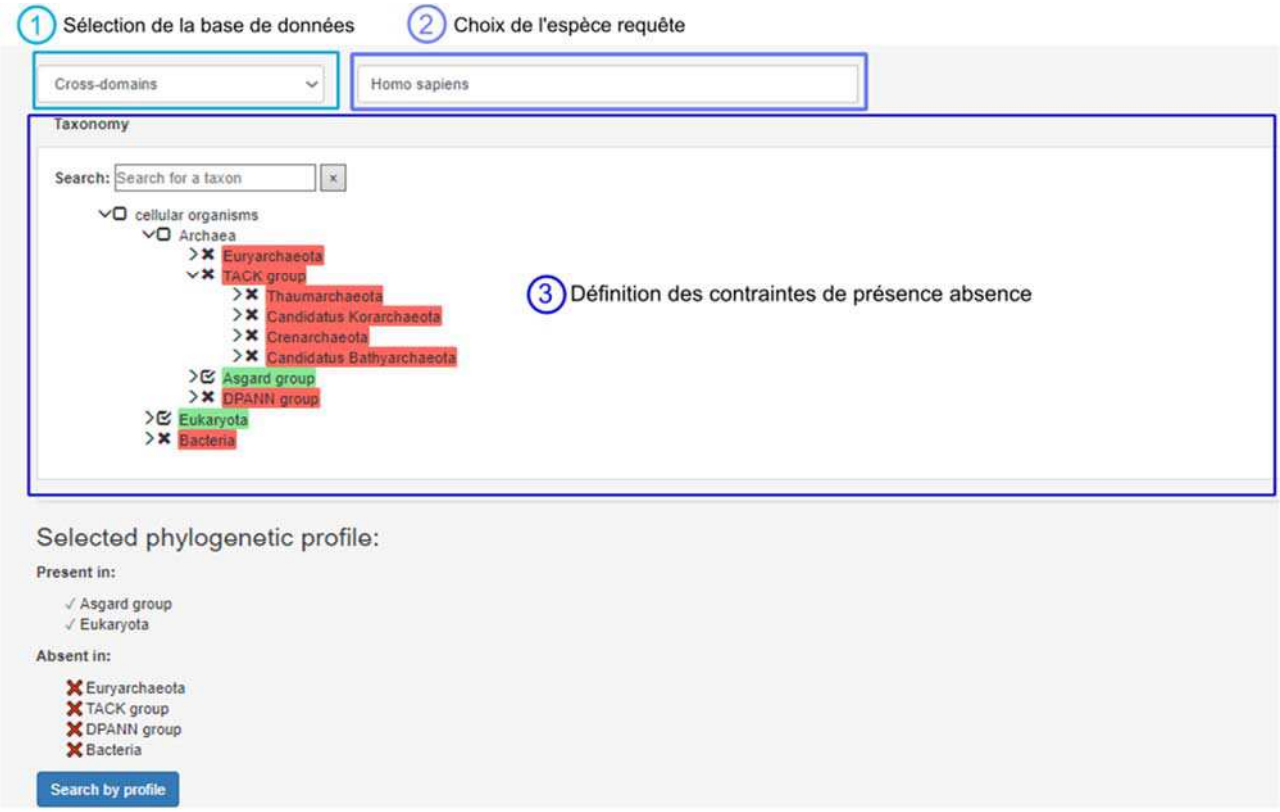


Figure 5-2 **Interface de recherche par profil phylogénétique**. Les paramètres se définissent en 3 étapes indiquées sur la figure. La sélection des contraintes de présence-absence se fait au niveau des clades dans un arbre taxonomique interactif.

5.1.2 Contextualisation des résultats

L'analyse des résultats d'une recherche par profil phylogénétique nécessite deux informations principales. La première est la distribution générale des protéines retrouvées, quels que soient les critères utilisés. Les protéines dont la distribution respecte un profil de présence-absence n'ont pas forcément une histoire évolutive identique. Il est donc essentiel de pouvoir visualiser chacune des histoires évolutives afin de discriminer des sous-catégories de profils, chacune potentiellement associée à des sous-fonctions différentes. La seconde information essentielle concerne les annotations fonctionnelles connues qui permettent de réaliser le lien entre les protéines identifiées par leur histoire évolutive et le phénotype recherché. Ce sont ces deux critères qui ont guidé la conception de l'interface de résultat et plus spécifiquement, la représentation des protéines identifiées.

5.1.2.1 Distribution taxonomique des protéines

La page de résultat, outre une section supérieure rappelant les critères de recherche utilisés affiche l'ensemble des protéines identifiées ordonnées en liste et représentées sous forme d'encarts (Figure 5-3). Ces encarts sont conçus pour intégrer les informations essentielles à l'analyse de façon compacte, afin de visualiser d'un seul coup d'œil un maximum de résultats. Pour afficher la distribution de chaque protéine, nous nous sommes appuyés sur le mode de représentation de distribution taxonomique présenté dans le chapitre précédent. Ce mode de visualisation synthétique résumant l'histoire évolutive d'une protéine est idéal pour mettre à disposition un panorama des histoires évolutives de plusieurs protéines. Sous la distribution, un volet dépliable permet d'avoir accès aux annotations fonctionnelles *Gene Ontology* de chaque protéine. Ce volet est par défaut replié dans un souci de concision, certains gènes pouvant avoir de très nombreuses annotations.

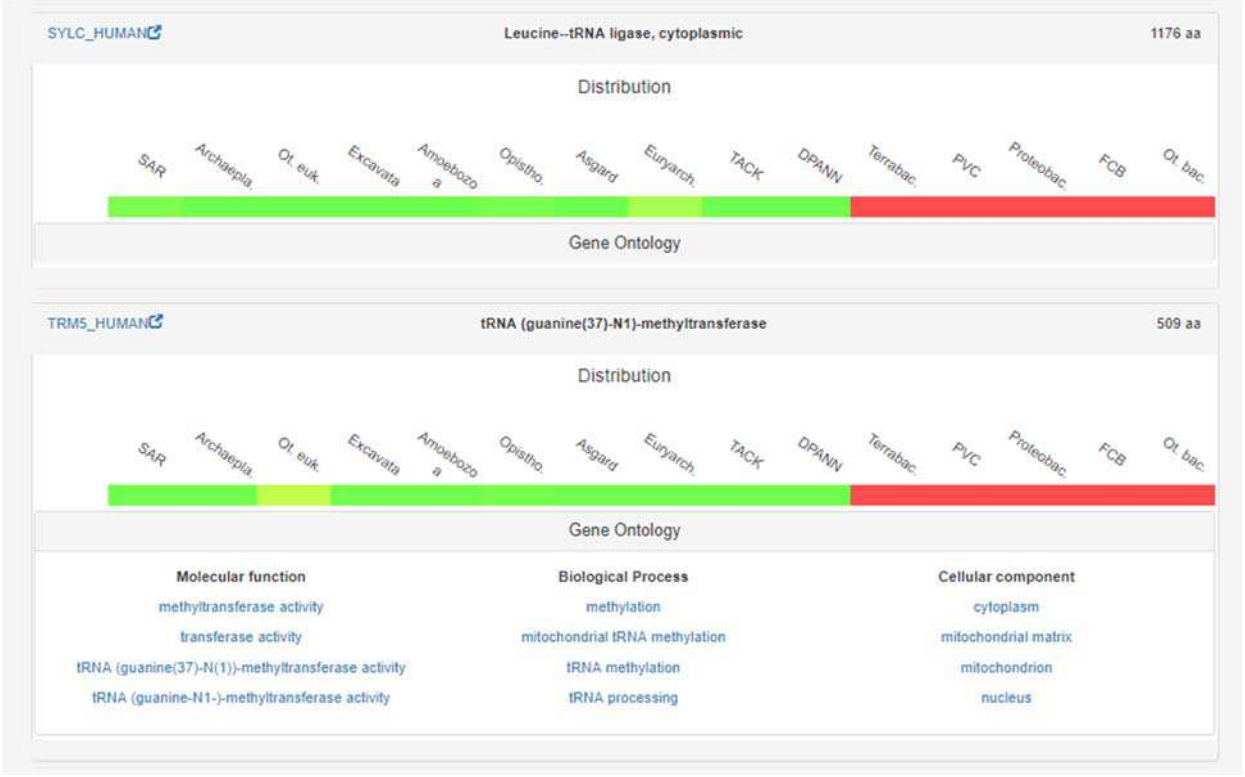


Figure 5-3 **Visualisation synthétique des protéines dans la recherche par profil**. Deux encarts de protéines sont représentés, chacun est décrit par l'identifiant, la description et la longueur de la protéine. On y retrouve la distribution synthétique de la protéine et un volet détaillant les annotations *Gene Ontology* associées.

5.1.2.2 Analyse fonctionnelle

Les encarts de protéines facilitent ainsi l'analyse des résultats, protéine par protéine. Pour analyser les liens entre histoires évolutives et fonctions de manière plus massive, nous proposons également des tests d'enrichissements fonctionnels sur la liste de protéines obtenues. Ces tests d'enrichissement permettent d'identifier dans les résultats, les termes fonctionnels surreprésentés dans les protéines identifiées, ce qui peut confirmer, infirmer ou affiner les liens

entre histoires évolutives et traits fonctionnels. Cette option exploite les termes *Gene Ontology* (bouton « *GO enrichment* ») et repose sur le *webservice* d'enrichissement en termes GO de la base de données Panther. De ce fait, l'option est seulement disponible pour les 112 espèces répertoriées sur la base de données, principalement des espèces couramment utilisées comme modèles expérimentaux.

Dans son principe, l'outil de profilage phylogénétique intègre une interface permettant de définir les profils phylogénétiques les plus divers et ainsi de s'adapter à de multiples questions biologiques. Les outils de visualisation des résultats et d'analyse fonctionnelle permettent de tirer de premières conclusions. Pour réaliser des analyses plus poussées, il est également possible d'exporter les résultats, sous forme d'une liste complète de protéines ou de séquences au format FASTA. Les potentialités de la recherche par profil sont illustrées ci-dessous avec la recherche des gènes liés à la mitochondrie chez les Eucaryotes.

5.1.3 L'étude d'un trait phénotypique iconique : la mitochondrie

L'application première des méthodes de recherche par profil est d'identifier les gènes liés à une fonction ou à un phénotype donné, en partant de sa distribution connue à l'avance. Ici, nous l'avons utilisé pour s'intéresser spécifiquement aux gènes liés à un trait phénotypique iconique des eucaryotes : la mitochondrie. Si la mitochondrie est présente chez la totalité des eucaryotes (à une exception près (Karnkowska et al., 2016)), elle se retrouve dans un état très réduit chez plusieurs clades.

Ces organelles, qui prennent alors le nom d'hydrogénosomes ou de métagénosomes, se caractérisent par leur incapacité à produire de l'énergie par respiration et une absence complète de génome propre (Müller et al., 2012). Ces structures sont observées notamment chez les Métamonades (une division d'Excavés qui inclue les groupes Fornicata et Parabasalia), l'embranchement des Microsporidies (champignons endoparasites) et chez les Archamoeba (une division des Amoebozoaires). En partant de ce constat, nous avons recherché les protéines présentes chez l'ancêtre commun des Eucaryotes et spécifiquement perdues dans ces divers clades pour évaluer leur correspondance fonctionnelle avec la mitochondrie.

Pour refléter l'histoire évolutive recherchée, nous avons donc mis en place deux ensembles de contraintes. Premièrement, nous avons imposé l'absence d'orthologues dans les clades susnommés, les Métamonades n'étant pas définis en tant que tels dans la taxonomie du NCBI, nous avons utilisé les deux clades enfants. Deuxièmement, les protéines recherchées étant *a priori* essentielles à la survie des autres eucaryotes, nous avons imposé une contrainte de présence d'orthologues dans l'ensemble des autres divisions des Eucaryotes. Afin de disposer d'annotations de qualité pour les analyses fonctionnelles, nous avons défini *Homo sapiens* comme espèce requête.

Avec ces contraintes de présence/absence, nous avons identifié 162 gènes. Si une analyse rapide permet bien d'identifier des gènes mitochondriaux, avec 28 de ces gènes associés au terme « *mitochondria* », ils ne semblent pas pour autant majoritaires. L'enrichissement en termes GO

confirme cette tendance avec un enrichissement significatif pour le terme de localisation « *mitochondrion* » (2,71e-7) mais plus faiblement par exemple que le terme « *cytoplasmic part* » ou « *endoplasmic reticulum membrane* ». L'association à la mitochondrie, bien que significative reste donc relativement faible par rapport à ce à quoi on pourrait s'attendre avec un profil si spécifique.

Partant de l'hypothèse que ces résultats pouvaient être dus à des spécificités phénotypiques de l'un des quatre clades considérés, nous avons réitéré la recherche en relevant les contraintes successivement. Si les résultats d'enrichissement en termes associés à la mitochondrie sont similaires pour trois de ces essais, la relâche de la contrainte sur les Microsporidies aboutit à un nombre bien plus élevés de résultats (318 protéines) et à un enrichissement très significatif (3.28^{e}-54 pour le terme *mitochondrion*). La comparaison de cette liste à celle obtenue en premier lieu montre que ces différences sont principalement dues à la présence d'orthologues dans l'une des espèces de Microsporidies : *Mitosporidium daphniae.* Une recherche bibliographique confirme qu'il s'agit de l'unique représentant du clade possédant un génome mitochondrial (Haag et al., 2014). Les résultats obtenus lors de notre première recherche s'expliquent donc principalement par cette différence entre le profil que nous avons défini pour la recherche et la distribution réelle du phénotype d'intérêt. Nous avons donc redéfini la recherche en ne soumettant pas cette espèce à la contrainte d'absence englobant les Microsporidies (Figure 5-4).

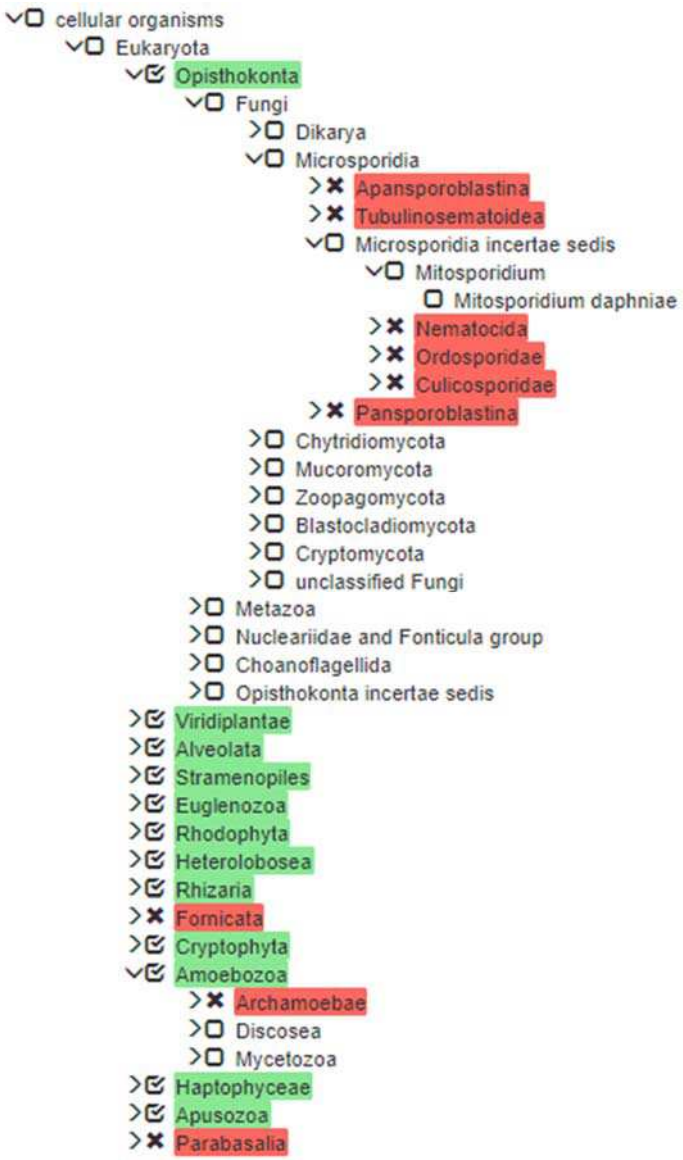


Figure 5-4 **Critères de recherche de gènes associés à la mitochondrie**. On recherche des gènes présents chez l'ensemble des clades eucaryotes (surlignés en vert), à l'exception des Fornicata, Parabasalia, Archamoebae et Microsporidies (surlignés en rouge) qui possédent une forme altérée de mitochondrie. La contrainte est levée pour *Mitosporidium daphniae*, une exception chez les Microsporidies.

En utilisant ces nouveaux critères de distribution, notre outil retrouve 257 protéines. Une première inspection visuelle des résultats permet d'estimer une forte proportion de gènes associés à la mitochondrie, avec 75 de ces protéines ayant le terme '*mitochondrial*' dans sa description. Les annotations *Gene Ontology* individuelles vont également dans ce sens, et on recense ainsi 99 annotations liées au composant cellulaire, '*mitochondrion'*. L'analyse d'enrichissement en termes *Gene Ontology* permet de formaliser ces observations : le terme *mitochondrion* est l'un des enrichissements les plus significatifs avec une valeur P de 3,07^{e}-45. On note que le nombre de protéines identifiées avec ce terme est seulement de 85 pour l'analyse d'enrichissement effectué par Panther, cette différence mineure peut s'expliquer par un échec

de *mapping* entre Panther et OrthoInspector ou par une prise en compte différente des termes *Gene Ontology*. Cela affecte cependant peu les résultats et on observe bien une surreprésentation conséquente de gènes liés à la mitochondrie.

L'analyse des distributions permet d'observer une certaine hétérogénéité entre les protéines, malgré les fortes contraintes imposées. Les différences se retrouvent majoritairement dans la proportion d'espèces de chaque clade dans lesquelles on retrouve des orthologues. La Figure 5-5 présente deux cas extrêmes. Les protéines absentes de nombreuses espèces comme FUT10_HUMAN ne présentent pas, en première analyse de rapport avec la mitochondrie et peuvent constituer des faux positifs.

Figure 5-5 **Hétérogénéité des distributions correspondant aux mêmes contraintes.** Deux exemples de protéine retrouvée dans la plupart des clades eucaryotes. SCO1_HUMAN est présente chez de nombreuses espèces non exclues (647 sur 705) quand FUT10_HUMAN est complétement absente de certains clades comme les Champignons (15 sur 384) ou les Alvéolés (1 sur 43).

La mitochondrie, et *a priori* ses gènes, dérivent selon toute vraisemblance d'un ancêtre bactérien. Il peut être intéressant d'analyser la distribution des gènes retrouvés non seulement chez les espèces eucaryotes, mais aussi dans les trois Domaines du Vivant. Pour cela, nous avons effectué une requête avec les mêmes contraintes sur les clades eucaryotes dans la base inter-domaines. 228 protéines sont identifiées en résultat de cette recherche. La différence avec la recherche précédente s'explique par le nombre réduit d'espèces dans la base de données inter-domaines qui, de fait, vérifie la contrainte de présence sur moins d'espèces dans les clades considérés. Cela a pour effet d'éliminer quelques faux positifs dont FUT10_HUMAN et la proportion de gènes mitochondriaux est légèrement supérieure, avec 96 gènes (42% contre 38% sur la base Eucaryote) associés au terme *mitochondrion* d'après Panther et un enrichissement plus significatif avec une P-valeur de $2.02*10^{e}$-48. Les distributions se répartissent en plusieurs catégories : une majorité de protéines présentes uniquement chez les Eucaryotes, certaines présentes dans les trois Domaines et finalement, les protéines présentes uniquement chez les Eucaryotes et les Bactéries. Ces différentes distributions révèlent quelques associations fonctionnelles, les protéines ayant des orthologues dans une forte proportion de bactéries sont associées à la traduction mitochondriale (ribosome, maturation de l'ARNt). A

l'inverse, parmi les protéines associées à la mitochondrie présentes uniquement chez les Eucaryotes, on retrouve de nombreuses protéines impliquées dans l'import de molécules vers la mitochondrie.

Ici, la recherche par profil phylogénétique nous a permis d'isoler des protéines associées à différentes fonctions et de déterminer des catégories distinctes de modules fonctionnels par une première inspection visuelle. Il est important de noter que si une telle recherche permet de découvrir les gènes associés à une fonction sous l'angle évolutif, elle ne permet pas d'isoler l'ensemble des gènes associés à ce processus. En effet, tous les gènes associés à un phénotype ne sont pas systématiquement perdus avec un phénotype donné, qu'ils aient un rôle important dans d'autres processus ou qu'ils aient subi une néofonctionnalisation. La mitochondrie est ici un cas particulier car elle n'est pas totalement perdue chez les espèces considérées et l'on peut considérer qu'il en est de même pour certains gènes associés. De plus, certains gènes mitochondriaux, même s'ils sont associés à un phénotype ancestral, peuvent avoir émergé plus récemment au cours de l'évolution et ne peuvent donc pas être identifié par une telle approche.

Pour autant, nous l'avons vu, les différences d'histoires évolutives entre les gènes identifiés de cette façon peuvent être associées à des sous-catégories fonctionnelles et donc, aider à en décrire le rôle biologique. Caractériser l'ensemble des gènes associés à une fonction selon leur distribution est le rôle du second outil que nous avons développé, qui prend le contrepied de la recherche par profil, en proposant de partir du phénotype pour étudier les profils évolutifs associés.

5.2 Du phénotype à l'évolution

La recherche par profilage phylogénétique permet de retrouver des gènes liés à une fonction ou un phénotype donné, en se basant sur la connaissance préalable de la distribution de ce phénotype. Cette distribution n'est pas forcément connue à l'avance et l'étude d'une fonction sous l'angle de l'évolution peut nécessiter d'analyser, dans un premier temps, la distribution des gènes qui y sont associés. Les gènes impliqués dans une même fonction principale pouvant être issus d'histoires évolutives différentes (comme l'a montré l'exemple de la mitochondrie), étudier les profils peut aider à identifier des sous-catégories fonctionnelles et les gènes associés. Pour faciliter ces analyses, nous avons mis au point un outil de profilage fonctionnel, permettant la visualisation synthétique des protéines associées à une fonction sous forme de distribution. Cet outil est également intégré au portail d'OrthoInspector, accessible depuis sa page d'accueil sous la section « *GO profile* ». On peut également y accéder à l'adresse suivante : http://www.lbgi.fr/orthoinspectorv3/go_profile.

5.2.1 Description de l'outil

Dans sa conception, l'outil de profilage fonctionnel fonctionne de façon symétrique à l'outil de recherche par profilage phylogénétique ; l'interface de recherche est donc similaire, mais au lieu de sélectionner des contraintes de présence/absence, il s'agit de sélectionner un terme

fonctionnel. Comme le reste du portail OrthoInspector 3.0, ces termes fonctionnels sont les annotations *Gene Ontology*. Tous les termes GO n'étant pas adaptés à toutes espèces (par exemple le terme *photosynthesis* ne s'applique pas aux gènes humains), on ne peut accéder qu'aux termes associés à l'espèce d'intérêt dans notre base de données GO locale *via* la recherche par auto-complétion. Il est ainsi impossible de lancer ces requêtes pour des espèces pour lesquels une annotation n'existe pas. Le paramétrage de la requête se sépare en trois étapes, dépendantes les unes des autres :

1. Sélection de la base
2. Sélection de l'espèce requête
3. Choix du terme fonctionnel

Les résultats sont ensuite obtenus automatiquement, le programme accédant dans un premier temps, aux identifiants UniProt des protéines associées au terme GO sélectionné *via* notre base de données locale et ces identifiants sont ensuite utilisés pour extraire les informations de distribution de la base de données OrthoInspector choisie.

Le but de l'outil étant de faire apparaître synthétiquement les distributions des gènes liés au terme sélectionné, nous avons choisi de représenter les résultats de la même façon que pour la recherche par profil, par des encarts synthétisant l'information de chaque protéine (décrits à l'étape précédente). Techniquement, nous avons donc basé l'affichage de ces deux parties du portail d'OrthoInspector 3.0 sur un même module programmatique qui permet de retranscrire sous cette modalité toutes listes de protéines issues d'une requête quelle qu'elle soit. En plus de permettre de garder une cohérence entre les outils, cette organisation modulaire nous laissera la possibilité par la suite d'intégrer aisément de nouveaux outils d'analyse basés sur cette représentation.

Ce mode de représentation étant le point névralgique de ces deux outils, il convient de noter que leur implémentation est là encore prévue pour s'adapter à des contraintes techniques, notamment en ce qui concerne la représentation des distributions. L'outil de visualisation de distribution disponible sur les pages de chaque protéine intègre les relations d'orthologie déjà récupérées dans la table d'orthologie. Accéder à l'ensemble de ces informations pour des centaines, voire des milliers de protéines et les afficher simultanément consomme des ressources de calcul et peut par conséquent, ralentir le processus, notamment pour les bases de données les plus volumineuses. Pour éviter ce problème, nous avons construit une base de données PostgreSQL dédiée aux distributions et enregistrant spécifiquement les informations de présence d'orthologues, sous la forme de relations protéine-espèce. Cette organisation nous permet ainsi d'avoir accès à l'ensemble des distributions en une requête simple et de rendre l'exploration des données sous cette vue, plus ergonomique.

Pour illustrer les possibilités de l'outil, nous avons utilisé ce dernier pour caractériser évolutivement les protéines associées à un phénotype bien étudié : la photosynthèse.

5.2.2 Analyse de la photosynthèse du point de vue évolutif

La photosynthèse est la capacité de produire des molécules organiques à partir du dioxyde de carbone en utilisant l'énergie lumineuse. On retrouve cette capacité chez plusieurs groupes de bactéries, notamment les Cyanobactéries, ainsi que dans plusieurs divisions taxonomiques eucaryotes : les Archaeplastides, les Straménopyles, les Haptophycae, les Cryptophytes et les Chromerida. Chez les Eucaryotes, cette fonction est assurée par le chloroplaste, une organelle tenant probablement son origine de l'endosymbiose d'une Cyanobactérie chez l'ancêtre des Archaeplastides (Criscuolo et Gribaldo, 2011) et d'endosymbiose d'autres eucaryotes photosynthétiques pour les ancêtres des autres taxons. Dans ce contexte, nous avons utilisé l'outil de profilage fonctionnel pour étudier les signatures évolutives des protéines photosynthétiques eucaryotes.

Dans un premier temps, nous avons utilisé *Chlamydomonas reinhardtii* comme espèce requête pour visualiser la distribution générale des protéines annotées par le terme GO « *photosynthesis* » dans la base de données inter-domaines. Cette requête donne 99 résultats que l'on peut séparer en plusieurs catégories en fonction de leur distribution à travers les domaines du Vivant. Une majorité de protéines (51) est retrouvée uniquement chez des eucaryotes. Parmi celles présentes dans les autres Domaines, 10 sont présentes chez des représentants des trois domaines et 5 chez plusieurs clades de Bactéries. Une dernière catégorie, plus importante, regroupe 33 protéines présentes à la fois chez les eucaryotes et quelques Terrabacteria (Figure 5-6).

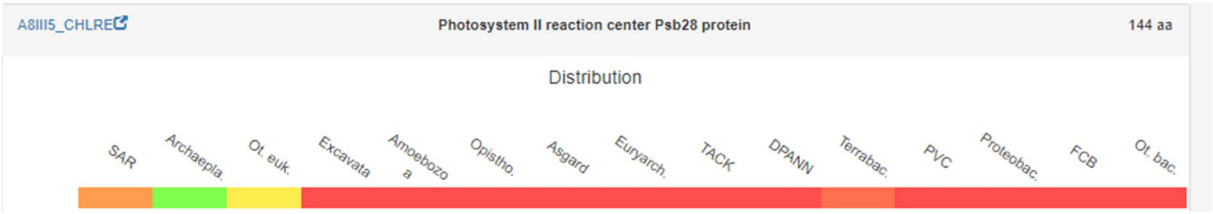


Figure 5-6 **Profil phylogénétique d'une protéine photosynthétique**. La protéine est présente chez les espèces eucaryotes photosynthétiques (SAR, Archaeplastidae et Other Eukaryota (Cryptophytes et Chromerida)) et certaines Terrabactéria (6/55 ; rouge-orangé sur la figure).

Cette dernière catégorie est homogène à la fois en termes de distribution et de fonction puisque la majorité des protéines (23/34) sont associées aux photosystèmes I et II, centraux pour la photosynthèse. La distribution synthétique montre que ces protéines ont 1 à 7 orthologues chez les Terrabactéries. Une analyse plus fine passant par les pages de ces protéines dans OrthoInspector révèle que toutes les Terrabactéries concernées sont des Cyanobactéries. On peut donc faire l'hypothèse que l'ensemble des gènes de cette catégorie sont des constituants ancestraux de l'appareil photosynthétique de *Chlamydomonas*.

Pour les protéines de *Chlamydomonas* reliées à la photosynthèse et trouvées exclusivement chez les Eucaryotes, nous avons effectué la même requête dans la base Eucaryote de façon à avoir une résolution plus fine des clades conservés. Comme l'on pouvait s'y attendre, les protéines associées à la photosynthèse se retrouvent, à deux exceptions près, uniquement dans

les 6 grandes divisions capables de photosynthèse : les Straménopiles, les Viridiplantae, les Rhodophytes, les Cryptophytes, les Haptophyceae et les Alvéolés (Figure 5-7). Une majorité de protéines ont des orthologues dans plusieurs de ces clades, bien qu'une trentaine d'entre elles semblent être apparues récemment et n'ont d'orthologues que chez une partie des plantes. A ce niveau, il est moins aisé de définir des catégories à partir de ces distributions, qui restent relativement similaires.

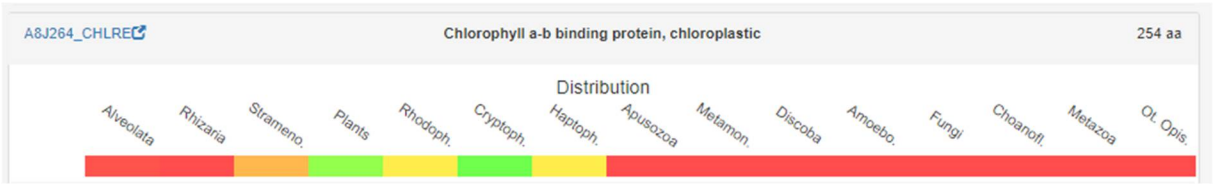


Figure 5-7 **Profil évolutif d'une protéine photosynthétique chez les Eucaryotes**. La protéine est présente dans la majorité des Plantes (56/63) une partie des Straménopiles (7/21), le seul représentant des Cryptophytes, un Haptophyceae (1/2) et Alveolés (1/55, peu visible) ici. Cette distribution correspond à la capacité de photosynthèse chez les Eucaryotes.

Les profils de distributions des clades corrélant, d'une façon générale, avec la distribution de la photosynthèse, nous avons cherché à remonter cette correspondance plus loin, jusqu'au niveau de l'espèce. En analysant finement les distributions, il est apparu qu'une plante et 14 espèces de Straménopiles n'ont aucun des orthologues de protéines photosynthétiques. En utilisant les pages des protéines correspondantes, nous avons identifié, parmi les Straménopiles, deux clades (les Oomycètes et le genre *Blastocystis*) dépourvus de chloroplastes et donc incapables de photosynthèse (Lamour et al., 2007). La plante n'ayant pas d'orthologue de protéines photosynthétiques est l'algue parasitaire *Helicosporidium,* chez qui la perte de la photosynthèse est également documentée (Pombert et al., 2014). En outre, nous remarquons que l'unique représentant des Cryptophytes, bien que photosynthétique, ne présente aucun orthologue pour une part des protéines entraînant *de facto* l'absence de tout le clade.

Les observations par inspection visuelle des distributions de protéines associées à la photosynthèse permettent, on le voit, de mieux caractériser la distribution du phénotype. Cette connaissance peut ensuite être exploitée pour préciser le profil et rechercher de nouveaux gènes. Nous en avons tiré profit pour initier une recherche par profil, en intégrant des contraintes sur les espèces non-photosynthétiques identifiées sur la base des distributions et en évitant les contraintes sur les Cryptophytes représentés par une seule espèce n'ayant pas d'orthologue pour une partie de ces protéines (Figure 5-8).

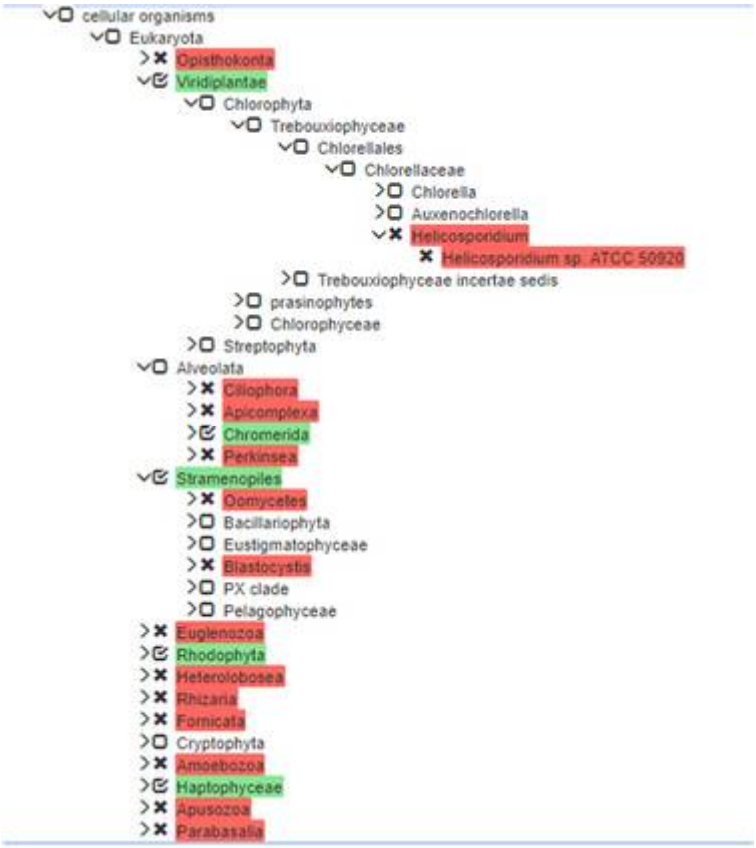


Figure 5-8 **Recherche par profil de gènes chloroplastiques**. Contraintes utilisées pour la recherche de gènes associés à la photosynthèse exploitant les informations issues des distributions. Sont ajoutées notamment les contraintes d'absence (en rouge) sur *Helicosporidium*, les Oomycètes et *Blastocystis*. La contrainte de présence est relâchée sur les Cryptophytes.

Cette recherche par profil permet de retrouver 99 protéines de *Chlamydomonas reinhardtii*, dont 31 sont annotées comme associées à la photosynthèse et donc, communes avec la liste identifiée par la recherche précédente. On trouve donc, sans surprise, un enrichissement significatif en gènes associés au terme *photosynthesis* ($1{,}56^{e}$-46). Parmi les nouveaux gènes détectés, 12 gènes n'ont aucune annotation. Sur la base de l'analyse approfondie de leur distribution, nous faisons l'hypothèse qu'il s'agit de gènes potentiellement associés à la photosynthèse.

Comme le montre cette étude, la représentation synthétique des gènes associés à une fonction peut servir deux objectifs : identifier des signatures évolutives spécifiques d'un sous-ensemble de ces gènes et permettre une meilleure compréhension de la distribution du phénotype étudié ainsi que des gènes qui y correspondent. Utilisée en complément de la recherche par profil, elle permet de faciliter l'identification d'autres gènes potentiellement associés à la fonction choisie.

5.3 L'histoire évolutive pour relier les gènes entre eux : un chemin vers l'intégration

Le dernier axe d'exploitation des profils phylogénétiques que nous avons mis en place s'intéresse non pas à relier directement des protéines à un phénotype, mais à identifier des liens

entre gènes, en fonction de la similarité de leurs profil phylogénétiques et donc, de leur histoire évolutive. Selon l'hypothèse de base du profilage phylogénétique, les gènes interagissant entre eux ou ayant des liens fonctionnels sont conservés ensemble au cours de l'évolution, on peut donc supposer un lien fonctionnel entre les gènes présentant une histoire similaire. En conséquence, ces liens facilitent aussi l'analyse des relations entre histoire évolutive et fonction : on peut partir d'un gène connu comme étant impliqué dans une fonction pour retrouver d'autres gènes qui y sont liés. Dans cet axe d'exploitation, nous avons défini un protocole pour construire des profils phylogénétiques complets à partir de relations d'orthologie et les exploiter pour examiner les similarités d'histoire évolutive entre gènes.

5.3.1 Génération de profils phylogénétiques

La recherche par profil et le profilage fonctionnel présentés dans les sections précédentes reposent sur des requêtes directes aux bases de données d'orthologie, ils ne nécessitent donc pas une représentation complète des profils en tant que tels. Il s'agit en revanche d'un prérequis pour mesurer les similarités entre histoire évolutive à l'échelle de la totalité des gènes d'une espèce. Nous avons pour cela développé un programme, Phyligrane, permettant la génération de matrices phylogénétiques complètes pour chaque espèce à partir des bases de données d'orthologie.

Brièvement, ce programme prend comme paramètres une espèce requête, le nom de la base de données OrthoInspector requise et une liste d'espèces cibles. Pour chaque protéine *Pi* de l'espèce requête, un profil phylogénétique est généré dans l'ensemble des espèces cibles *ej* sous la forme d'un vecteur *Vi*. Pour cela, l'existence d'un des trois types de relation d'orthologie (un-à-un, un-à-plusieurs, plusieurs-à-plusieurs) entre *Pi* et une protéine de *ej* dans la base de données est évaluée, si une relation existe on considère la protéine comme présente et la position du vecteur *Vij* prend la valeur 1, dans le cas contraire elle prend la valeur 0. Une fois l'ensemble des profils phylogénétiques calculés, Phyligrane créé un fichier CSV décrivant la matrice phylogénétique totale, dont les lignes sont les protéines de l'espèce requête et les colonnes les espèces. Pour faciliter l'analyse visuelle de ces fichiers, les colonnes d'espèces sont automatiquement classées en fonction de leur taxonomie.

En utilisant ce programme, nous avons généré les matrices phylogénétiques de toutes les espèces pour chacune des bases de données d'OrthoInspector. Le choix des espèces cibles à prendre en compte est un paramètre essentiel pour le profilage phylogénétique, dont la précision est tributaire du nombre et de la diversité des espèces considérées. Les Archées étant relativement peu nombreuses dans nos bases de données, elles ont toutes été utilisées comme espèces cibles. A l'inverse, les Bactéries et les Eucaryotes étant bien représentés, nous avons privilégié la diversité au nombre et utilisé uniquement comme cibles les espèces 'modèles' définies plus tôt. Dans la logique de la définition des espèces 'modèles' (voir section 4.3.1), cela permet d'éviter un biais trop important vers des clades surreprésentés lors de la comparaison de profils.

5.3.2 Comparaisons de profils

Les distances entre histoires évolutives de ces protéines ont été estimées sur la base des profils phylogénétiques. Dans le cadre de cette application, nous avons testé deux mesures distinctes : l'indice de dissimilarité de Jaccard, dont la valeur évolue selon la proportion d'indicateurs de présence communs entre deux profils phylogénétiques et une distance basée sur le paramètre de corrélation de Pearson. Nous avons évalué ces deux mesures de distance en les appliquant à la matrice phylogénétique de l'homme générée à partir de la base inter-domaines. Les résultats d'une telle comparaison sont représentés sous la forme d'une matrice carrée, dont les lignes et les colonnes sont les protéines humaines et dont les cellules indiquent les distances entre paires de protéines. La Figure 5-9 montre la distribution des distances obtenues pour chacune des deux mesures. La distance de corrélation s'étend, contrairement à la dissimilarité de Jaccard au-delà de la valeur 1, pour rendre compte des protéines ayant un profil inversement corrélé. Outre ce point majeur, les mesures suivent une distribution différente : on retrouve une proportion croissante de relations vers les mesures de dissimilarité importante pour la dissimilarité de Jaccard, avec peu de mesures en dessous de 0.4 et une majorité à la valeur extrême de 1, alors que la distance de corrélation de Pearson suit une distribution relativement homogène, centrée autour d'une valeur de 0,6.

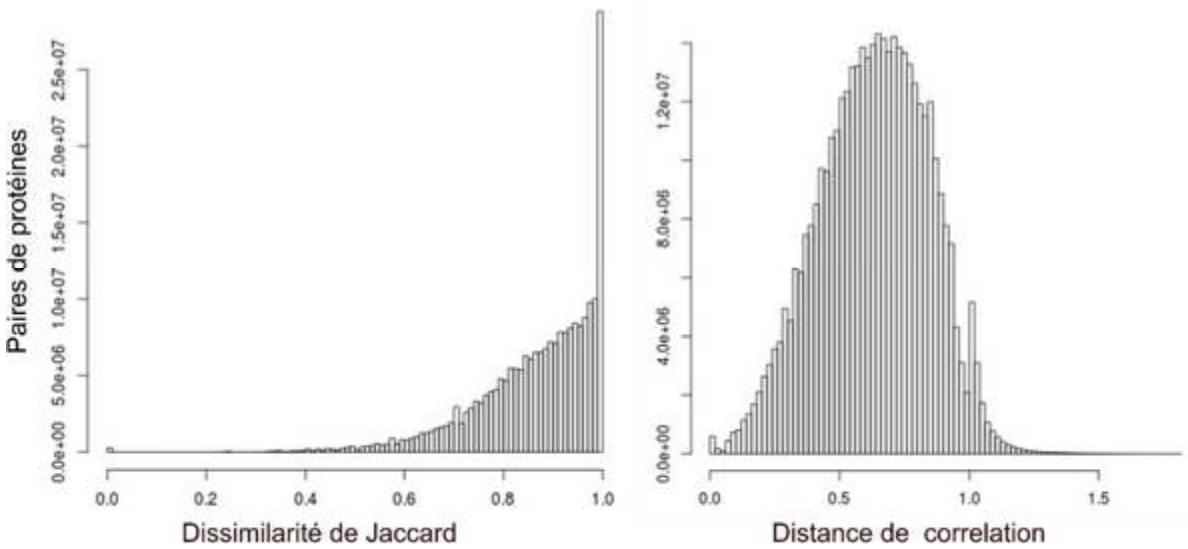


Figure 5-9 **Distributions des valeurs de distance pour une matrice phylogénétique**. Les deux mesures de distances ont été appliquées sur la matrice phylogénétique de l'homme pour la base inter-domaines.

Notre objectif, dans le cadre de cet outil, étant de s'intéresser exclusivement à une fraction de protéines ayant un certain degré de proximité entre elles, nous avons choisi d'utiliser la dissimilarité de Jaccard dont la distribution facilite la discrimination basée sur un seuil. Nous avons donc calculé les distances entre profils selon cette métrique pour l'ensemble des matrices phylogénétiques de chaque base de données, pour générer autant de matrices de distances.

Pour chacune des espèces, le nombre de mesures étant égal au carré du nombre de protéines, les résultats de cette opération sont particulièrement volumineux. A titre d'exemple, pour le colza (*Brassica napus)*, l'espèce d'OrthoInspector dont le protéome est le plus volumineux avec

98 750 protéines, cela équivaut à 9 milliards de mesures de distance stockées dans une matrice de 18 Go. L'ensemble des mesures de distances représentent plus de 2,4 To de données pour plusieurs centaines de milliards de couples protéine-protéine. Les principaux contributeurs à ce volume sont les Eucaryotes dont les protéomes de grande taille induisent de nombreuses comparaisons protéine-protéine par espèce (Tableau 5-1).

Tableau 5-1 **Volume des données de distances entre profils**

	Eucaryotes	Bactéries	Archées	Inter-domaines	Total
Volume des matrices	1,5 To	522 Go	13 Go	402 Go	2,4 To
Nombre de mesures (en milliards)	189,5	51,5	1,0	44,8	286,7
Nombre de mesures <0,5 (en milliards)	1,7	0,3	0,3	0,2	2,5

Une telle quantité de données posant des problèmes de stockage et d'accessibilité, nous avons décidé d'exploiter uniquement les cas les plus pertinents. La plupart des profils étant éloignés, la majeure partie des valeurs de dissimilarité est proche de 1 (47,5% des mesures sont supérieures à 0,9) et n'apporte aucune information sur les liens fonctionnels entre protéines. Sur l'ensemble de ces matrices de distances, nous avons extrait uniquement les paires de protéines dont la dissimilarité de Jaccard est inférieure ou égale au seuil de 0.5, soit 1% des relations totales. A cette valeur, les deux profils considérés ont en commun plus de trois quarts des espèces présentes pour l'ensemble des deux profils. Pour faciliter leur exploitation automatique, les données sélectionnées sur ces critères ont été enregistrées dans une base de données SQL dédiée, dont les tables indexées comprennent les identifiants des deux protéines de la paire et la mesure de distance exacte. Cette base est considérablement réduite par rapport aux données initiales avec un volume de 270 Go.

Ces mesures de distance sont directement exploitées dans OrthoInspector 3.0 comme nous allons le voir, mais pourraient aussi être intégrées à d'autres plateformes comme le montre l'exemple de MyGeneFriends décrit dans la dernière section du chapitre.

5.3.3 Les distances dans OrthoInspector 3.0

Les mesures de similarités entre profils évolutifs de deux protéines ont des implications à la fois évolutives et fonctionnelles, et permettent d'apporter une nouvelle source de contextualisation entre protéines. Dans cette logique, nous avons intégré les similarités entre histoires évolutives à la page décrivant chaque protéine dans deux sections dédiées « *Proteins with similar distributions* ». Ces sections permettent de retrouver les protéines ayant une distribution similaire soit dans les espèces du domaine considéré, soit dans les espèces des trois Domaines du Vivant (Figure 5-10). Cette dernière section repose sur les calculs de similarité réalisés sur les matrices phylogénétiques de la base inter-domaines et n'est donc disponible que

pour les protéines d'organismes 'modèles' dans OrthoInspector. Pour les protéines spécifiques des Bactéries ou des Eucaryotes, les deux sections fournissent des résultats identiques, la dissimilarité étant calculée sur le même jeu d'espèces, les espèces 'modèles'.

Dans ces sections sont répertoriées les protéines dont le profil est le plus similaire à la protéine requête. Le seuil de distance retenu à l'heure actuelle pour l'affichage est de 0.4, ce qui correspond à 84% d'espèces présentes en commun entre deux profils. Ce seuil déterminé à partir d'exemples de référence pourra par la suite être adapté entre 0 et 0,5 sur la base d'évaluations objectives.

Les protéines similaires sont présentées sous la forme d'un tableau comprenant l'identifiant de chaque protéine (avec un lien vers la page correspondante d'OrthoInspector et vers UniProt), sa description et la valeur exacte de la dissimilarité entre les deux protéines. Par souci de clarté, seules les 5 protéines les plus proches sont affichées par défaut. Un lien permet d'avoir accès, le cas échéant, aux autres protéines dont la dissimilarité est inférieure au seuil de 0.4.

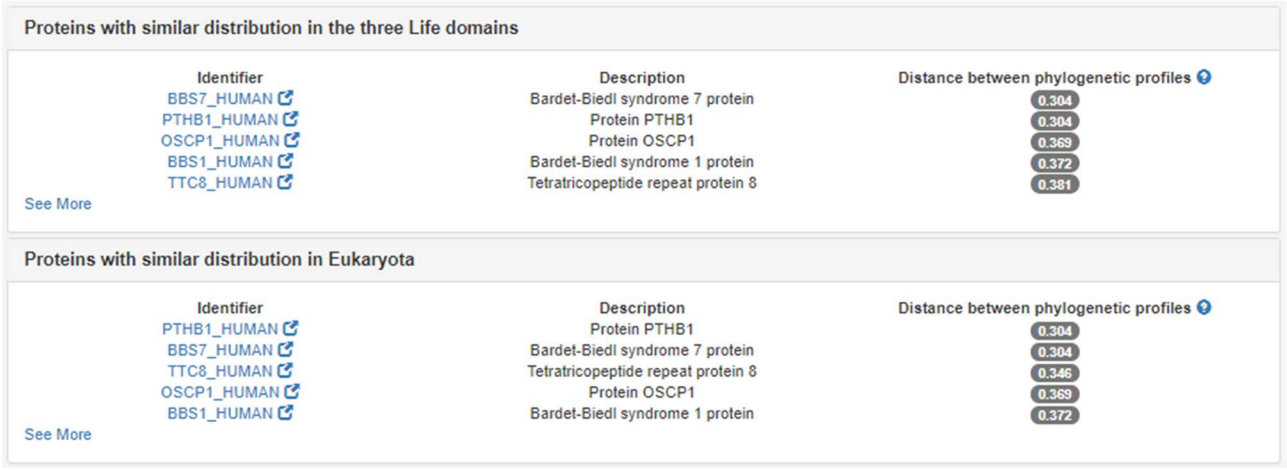


Figure 5-10 **Protéines avec un profil phylogénétique similaire dans OrthoInspector 3.0** Protéines présentant un profil phylogénétique similaire à la protéine humaine BBS2 (BBS2_HUMAN) dans les trois domaines (en haut) et chez les Eucaryotes (en bas). Pour les protéines spécifiques des Eucaryotes, les distances sont les mêmes dans les deux cas. En revanche, TTC8_HUMAN possède des orthologues chez certains procaryotes contrairement à BBS2_HUMAN et voit donc sa distance augmenter lorsque l'on prend en compte la distribution dans les trois domaines.

5.3.4 Une exploration par similarité : la méthanogenèse

La similarité entre profils permet d'estimer les liens évolutifs entre protéines d'une même espèce et par extension, leurs liens fonctionnels (protéines interagissant entre elles ou faisant partie d'un même complexe ou d'un même processus biologique). Pour illustrer ce type d'application et les potentialités offertes par OrtoInspector, nous avons recherché les protéines présentant un profil similaire à la protéine mcrA (*Methyl-coenzyme M reductase I subunit alpha*) de l'archée *Methanopyrus kandleri*. Cette protéine associée à deux autres sous-unités mcrB (*Methyl coenzyme M reductase, beta subunit*) et mcrG (*Methyl-coenzyme M reductase*

subunit gamma) a un rôle dans une voie métabolique spécifique de certaines Archées, la méthanogenèse.

Dans la section dédiée d'OrthoInspector, on retrouve 19 protéines avec un profil similaire à cette protéine (identifiant MCRA_METKA) (Figure 5-11). Parmi ces protéines figurent les deux autres sous-unités associées à notre protéine d'intérêt. Ainsi, les mesures de similarité permettent de retrouver les protéines d'un même complexe lorsque leur histoire évolutive est similaire. On remarque qu'une partie des autres protéines proches de la protéine d'intérêt sont également associées, d'après leurs descriptions, à la même unité catalytique. Cependant, une partie des protéines trouvées dans cette liste manquent d'annotations (*Uncharacterized proteins*). On peut faire l'hypothèse qu'elles sont impliquées dans la même fonction.

Proteins with similar distribution in Archaea

Identifier	Description	Distance between phylogenetic profiles
Q8TYG7_METKA	Uncharacterized protein conserved in archaea	0.187
Q8TXL1_METKA	Methyl-coenzyme M reductase operon protein C	0.187
F1SVE8_METKA	Methyl coenzyme M reductase, subunit D	0.187
Q8TXT1_METKA	Uncharacterized protein conserved in archaea	0.187
Y796_METKA	UPF0288 protein MK0796	0.189
Q8TXS7_METKA	Uncharacterized protein conserved in archaea, related to methyl coenzyme M reductase II, operon protein C (MtrC)	0.226
F1SVF3_METKA	Methyl coenzyme M reductase, beta subunit	0.228
Q8TXT0_METKA	Activator of 2-hydroxyglutaryl-CoA dehydratase (HSP70-class ATPase domain)	0.263
Q8TXL0_METKA	Methyl coenzyme M reductase, gamma subunit	0.265
Q8TYZ1_METKA	Uncharacterized protein conserved in archaea	0.289
Y941_METKA	Uncharacterized methyltransferase MK0941	0.291
Q8TVH3_METKA	Nitrogenase subunit NifH (ATPase)	0.291
Q8TVK4_METKA	Nitrogenase molybdenum-iron subunit	0.294
Q8TYE9_METKA	Predicted DNA-binding protein containing a Zn-ribbon	0.316
Q8TXT5_METKA	Uncharacterized protein conserved in archaea	0.318
Q8TXS8_METKA	Predicted Fe-S oxidoreductase	0.381
Q8TUS3_METKA	Predicted phosphatase of the PHP family	0.392
Q8TVF7_METKA	Transcription factor homologous to NACalpha-BTF3 fused to metal-binding domain	0.397
Q8TXS9_METKA	Uncharacterized protein conserved in archaea	0.397

Figure 5-11 **Protéines avec une distribution similaire à MCRA_METKA**. Les protéines membres du même complexe sont surlignées en bleu, les autres protéines associées au même complexe d'après leur description en vert.

L'analyse des distributions de ces protéines révèle que, malgré quelques différences mineures elles partagent plusieurs points communs : une présence dans trois classes majeures de l'embranchement Euryarchaota (Methanobacteria, Methanococci et Methanomicrobia) et dans une espèce d'euryarchée non classée (Euryarchaeota archaeon 55_53). A cela s'ajoutent quelques clades pour lesquels on observe une présence dans seulement une partie des représentants, à savoir une partie (8/20) de la classe des Thermoplasmata (correspondant à l'ordre des Methanomassilicoccales et deux autres espèces non classées), le groupe des archées ASGARD, les classes des Hadesarchaea, des Theionarchaea, des Archaeoglobi et l'embranchement des Bathyarchaeota. A partir de ces observations, il est possible d'établir des contraintes pour un profilage phylogénétique afin d'associer des protéines non caractérisées à une fonction potentielle (Figure 5-12).

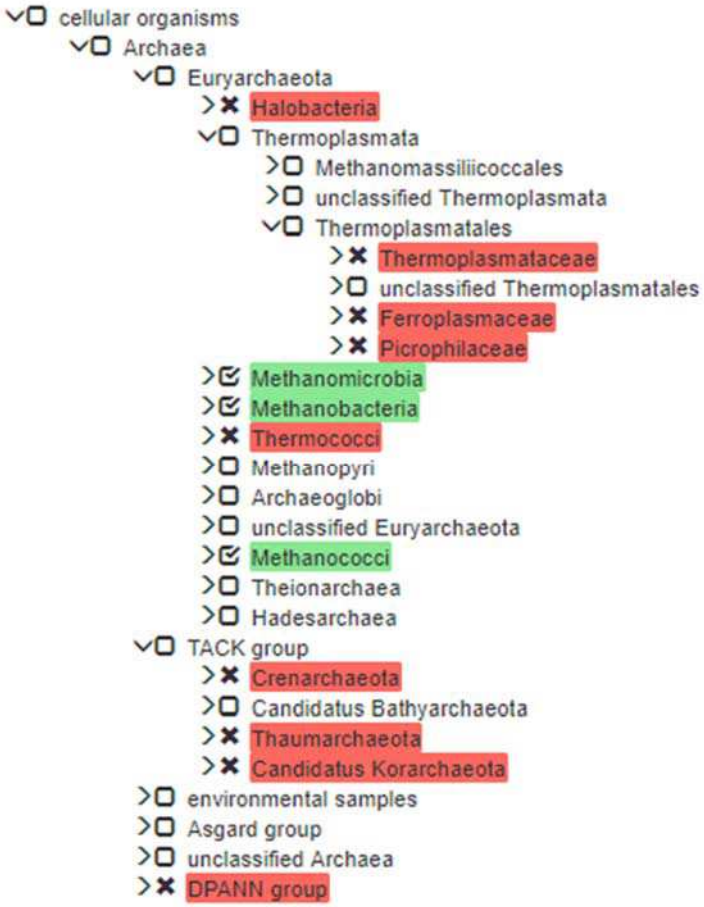


Figure 5-12 **Recherche de profils associés à mrcA**. Recherche de gènes utilisant les contraintes identifiées par l'analyse des profils proches de la protéine MCRA_METKA (en vert, présence d'orthologues, en rouge, absence). Les autres clades, ainsi que les groupes sans réalité taxonomique (*unclassified, environmental samples*) sont ignorés.

La recherche effectuée avec ces critères donne au total 57 résultats : les 20 protéines ayant servi à définir les contraintes et 37 autres. Les analyses d'enrichissement fonctionnel n'étant pas disponible pour *Methanopyrus kandleri*, on ne peut pas faire reposer les associations fonctionnelles sur des bases statistiques. On remarque cependant que le terme *methanogenesis* est le plus représenté dans les résultats, avec 12 occurrences dont 7 parmi les nouvelles protéines identifiées par la recherche par profil. Cela nous permet de faire l'hypothèse que le profil recherché est associé au sens large aux voies métaboliques de la méthanogenèse.

Pour finir, le profilage fonctionnel des protéines associées à ce terme GO nous permet de conclure cette analyse. Les 31 protéines associées à la méthanogenèse ont des profils relativement hétérogènes mais 24 sont présentes dans l'ensemble des clades identifiés plus tôt qui apparaissent donc comme des espèces capables de méthanogenèse. Une recherche rapide dans la littérature nous permet de confirmer cela. On peut ainsi considérer que les protéines non caractérisées identifiées précédemment sont bien des candidats comme gènes associés au phénotype de méthanogènes.

Comme l'illustre cet exemple, les mesures de similarité basées sur les profils constituent un outil efficace pour non seulement identifier les protéines liées fonctionnellement, mais également pour initier l'étude des liens entre histoire évolutive et fonction à partir d'une seule protéine. En ce sens, cette méthode d'analyse des profils phylogénétiques est complémentaire des autres outils de génomique comparative disponibles sur OrthoInspector 3.0 pour exploiter le profilage phylogénétique, un marqueur évolutif, dans l'étude des phénotypes.

Par leur nature, les mesures de similarité entre gènes permettent de construire un réseau entre les protéines, cette caractéristique est essentielle pour la dernière partie de mes travaux de thèse sur cet axe, qui implique l'intégration des marqueurs évolutifs avec d'autres données biologiques.

5.4 MyGeneFriends

MyGeneFriends est une plateforme d'analyse biologique des relations entre gènes humains et maladies génétiques qui prend le parti de formaliser ces relations sous forme d'un réseau social, en s'appropriant les codes du Web 2.0. Dans le principe, MyGeneFriends est un *hub* d'intégration de données hétérogènes par leur nature et leur source. Ces données sont utilisées pour construire des liens, dits d'amitié entre les différents acteurs (entre gènes, entre maladies et entre gènes et maladies). Dans MyGeneFriends, l'utilisateur (chercheur, médecin…) est lui-même un acteur du réseau social, ce qui ouvre des possibilités intéressantes en termes de contextualisation et de visualisation de données. MyGeneFriends est disponible sur http://lbgi.fr/mygenefriends/ et a fait l'objet d'une publication (Allot et al., 2017) dans la revue *Journal of Medical Internet Research*. Cette publication est disponible en annexe de ce manuscrit.

Ma contribution à MyGeneFriends, dans le cadre de ces travaux de thèse, concerne l'intégration des données évolutives, sous différentes formes, à ce réseau social. Il s'agit principalement d'apporter un élément de contexte complémentaire pour l'analyse des relations génotype-phénotype dans le contexte des pathologies humaines. Ici, je décris rapidement l'architecture du réseau social ainsi que la façon dont les données ayant trait aux gènes y sont représentées, à la fois sur la page de profil du gène et par ses relations d'amitiés. Je décris plus spécifiquement comment j'ai contribué à y intégrer les données évolutives.

5.4.1 Les acteurs du réseau social

MyGeneFriends envisage la gestion des flux de données hétérogènes en s'inspirant des techniques mises en place dans les réseaux sociaux, acteur clés du secteur technologie aujourd'hui. Les réseaux sociaux reposent, essentiellement sur deux types d'entités, les acteurs du réseau, classiquement des utilisateurs humains, représentés par un profil personnel et les relations qui lient ces acteurs. MyGeneFriends étant prévu pour faciliter l'étude des relations entre gènes humains et maladies, il met sur un même niveau trois types d'acteurs : l'utilisateur humain, les gènes, et les maladies, chacun ayant des liens spécifiques entre eux.

5.4.1.1 L'humain

L'humain, qu'il soit un chercheur, un médecin ou un patient éclairé, est l'utilisateur de MyGeneFriends. Comme dans tout réseau social, l'humain est décrit par certaines informations personnelles qui lui sont propres et pertinentes dans le cadre du domaine de la recherche : sa description, ses affiliations et éventuellement, sa liste de publications. Il peut devenir 'ami' avec d'autres utilisateurs humains pour faciliter le partage de données dans le cadre de collaborations. L'utilisateur peut définir ses sujets de recherche (*Topics*) dans les contextes desquels il peut choisir de devenir ami avec des gènes ou des maladies. Les liens qu'il créé lui permettent d'influencer le réseau social et de faciliter son analyse des données. En fonction de ses liens avec les autres acteurs, le réseau lui fait en retour des suggestions de nouveaux liens pertinents avec de nouveaux gènes ou maladies ayant un lien avec son sujet d'étude et lui recommande des publications scientifiques.

5.4.1.2 La maladie

La maladie décrit une altération de santé, donnant lieu à plusieurs phénotypes anormaux ; MyGeneFriends s'intéresse spécifiquement aux maladies génétiques humaines pour faciliter l'étude de leurs bases génétiques. Toutes les informations les concernant sont directement extraites de deux bases de données dédiées, OMIM (Amberger et Hamosh, 2017) et Orphanet. Les maladies sont décrites par deux caractéristiques principales, les phénotypes qui y sont associés, issus de la base de données HPO (Köhler et al., 2017), ainsi que les variants génétiques connus qui en sont responsables. Les maladies créent automatiquement des relations d'« amitié » avec les autres maladies. Ces liens d'amitiés sont créées lorsque les maladies partagent des phénotypes communs, si elles sont causées par l'atteinte des mêmes gènes et finalement, en fonction du nombre d'utilisateurs humains qui les ont sélectionnées comme amies dans le cadre du même sujet d'intérêt. Elles sont également « amies » avec l'ensemble des gènes qui y sont associés.

5.4.1.3 Le gène

MyGeneFriends inclut l'ensemble des gènes humains, codant ou non pour des protéines, extrait de la base de données génomiques Ensembl. Le gène est décrit par des données hétérogènes, basées sur des données transcriptomiques, bibliographiques ou encore évolutives, et c'est surtout à ce niveau que se jouent les enjeux d'intégration des données. Pour la suite de cette section, je me concentrerai donc principalement sur cet acteur. Comme pour les autres acteurs, les données sont utilisées à deux niveaux, pour décrire le gène lui-même sur sa page de profil et par le biais des liens avec les autres acteurs.

5.4.2 Le profil du gène

Le profil de gènes, comme l'ensemble des profils des acteurs de MyGeneFriends, est basé sur un modèle commun fragmentant l'information en différentes catégories comme montré dans la

Figure 5-13. L'information se répartit en trois niveaux : les informations disponibles sur le gène sur le premier niveau, les relations avec les autres acteurs au second niveau et un fil d'actualité permettant de suivre les changements dans les données disponibles. Ici, je me concentrerai sur les catégories du premier niveau en montrant comment elles contextualisent le gène et ma contribution avec l'apport d'un indicateur évolutif.

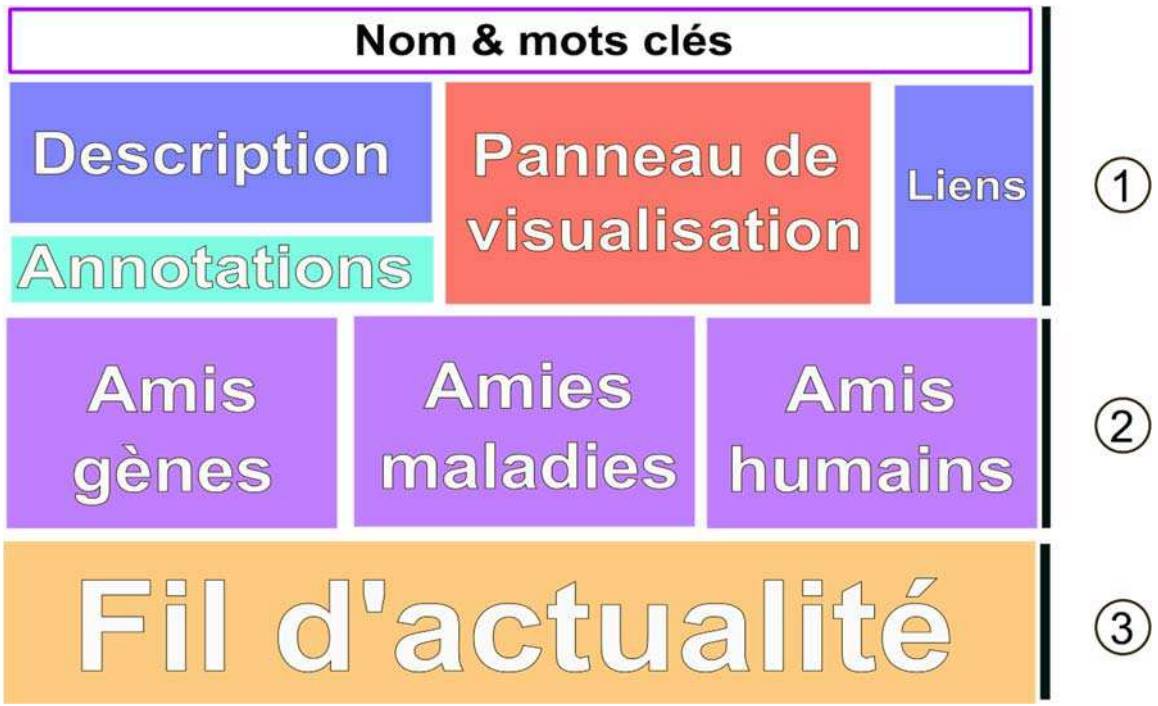


Figure 5-13 **Page de profil MyGeneFriends**. Modèle d'une page de profils et les différentes sections d'intégration des données. Les informations se stratifient en 3 niveaux. 1. Données générales de l'acteur. 2. Amitiés avec d'autres acteurs. 3. Fil d'actualité. Figure adaptée d'Allot (communication personnelle).

La section supérieure permet de définir le gène synthétiquement ; on y retrouve en premier lieu ses identifiants et ceux de ses produits (protéines) à travers plusieurs ressources publiques, ainsi que des mots-clés tirés de ses annotations. La description complète du gène, issue de la base Refseq, est disponible dans la section du même nom, image de la contextualisation permise par une structure de réseau social, les mots de cette description correspondant au sujet de recherche (*Topic*) actif de l'utilisateur y sont surlignés pour aider à identifier rapidement les informations d'intérêt. Dans la même logique d'interaction et de personnalisation, les utilisateurs peuvent ajouter des informations personnelles à chaque gène.

L'intégration de données se fait réellement dans le panneau de visualisation : quatre types de données y sont disponibles : les différents variants du gène, les tissus ou organes où le gène est exprimé d'après des données transcriptomiques, les publications qui sont rattachées à ce gène et, dans le cas de gènes protéiques, le compartiment cellulaire où il exprimé. Ces données permettent de contextualiser le gène selon les informations qui lui sont propres et surtout de les visualiser d'une manière qui permet d'en tirer tout le sens. A titre d'exemple, la localisation cellulaire dont les informations sont tirées des termes GO de la catégorie *cellular component* sont représentées dans un nuage de mots et leur taille est proportionnelle à la spécificité du mot

(inversement proportionnelle au nombre de fois où on le retrouve pour les gènes humains), ce qui permet de mettre en évidence le terme *a priori* le plus pertinent. Les valeurs des expressions des gènes sont elles représentées schématiquement sur une « carte » du corps humain, là encore de manière à exprimer synthétiquement les données et en faciliter l'exploitation.

Dans ce contexte, il s'agissait de présenter les données évolutives décrivant un gène d'une façon qui s'avère rapide à interpréter, et évidemment pertinente dans l'analyse des données. Le marqueur évolutif que nous avons choisi pour décrire le gène est également basé sur le profil phylogénétique, mais MyGeneFriends n'étant pas une ressource dédiée à la génomique comparative, une représentation de la distribution synthétique est moins pertinente que pour OrthoInspector 3.0. MyGeneFriends étant ciblé sur les gènes humains, nous avons opté pour une représentation prenant l'homme comme référence et classant les gènes en fonction de leur âge. Cette information peut servir d'indications sur la fonction des gènes. L'ensemble des catégories que nous avons définies sont présentées dans la Figure 5-14.

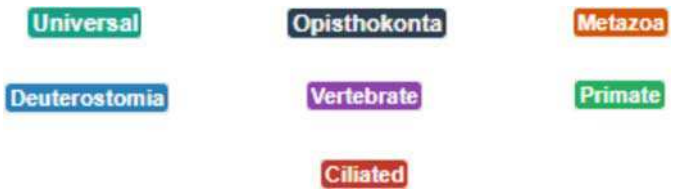


Figure 5-14 **Catégories évolutives utilisées dans MyGeneFriends**. Les 6 divisions taxonomiques catégorisant les gènes sur la base de leur profil phylogénétique, ordonnée (de droite à gauche, de haut en bas) en fonction de leur âge. La catégorie *ciliated* décrit une distribution phylogénétique correspondant au phénotype ciliaire.

La catégorie de chaque gène protéique humain a été déterminée sur la base de son profil de présence et d'absence dans 100 espèces eucaryotes. Elle prend en compte la division taxonomique comprenant l'ancêtre commun de l'homme et de l'espèce la plus éloignée chez laquelle on retrouve des orthologues. Ainsi, un gène humain avec des orthologues chez les animaux et les champignons se voit attribuer la catégorie Opisthokonte. La catégorie *Universal* est utilisée pour décrire les gènes présents chez une majorité d'Eucaryotes, et donc probablement apparu chez un de leurs ancêtres communs, voire chez le dernier ancêtre commun universel, et conservé chez une majorité d'entre eux, ce qui marque en toute hypothèse un rôle essentiel pour la cellule eucaryote.

La dernière des 7 catégories, *ciliated*, ne correspond pas à une division taxonomique. Nous l'avons définie, également à partir du profil phylogénétique sur la base de notre expertise sur l'étude des relations génotype-phénotype concernant le cil (voir chapitre suivant). Pour cela, nous y avons inclus les gènes présents dans tous les clades considérés comme étant ciliés et absents de toutes les espèces non ciliées. Cette distribution étant particulièrement spécifique, elle apporte également une information précise sur la fonction probable des gènes.

La proportion de gènes humains de MyGeneFriends annotés avec chaque catégorie est indiquée dans le Tableau 5-2. Sur 20193 protéines du protéome de l'homme, 14 628 ont pu être attribuées

à une de ces catégories, les autres présentant des distributions trop complexes pour pouvoir les classer efficacement. Ces annotations ont été transférées vers MyGeneFriends, on note cependant le nombre d'annotations basées sur les profils et le nombre d'annotations disponibles sur MyGeneFriends ne sont pas identiques. Cela s'explique principalement par la difficulté de faire correspondre les identifiants de gènes et de protéines à travers plusieurs ressources. Pour autant, un total de 13 510 gènes de MyGeneFriends bénéficient d'une catégorisation évolutive.

Tableau 5-2 **Nombres de gènes par catégorie dans MyGeneFriends**. Nombre de protéines humaines (SwissProt) identifiées par catégorie sur la base de leur profil et nombre d'annotations transférées dans MyGeneFriends.

	Universal	Opisthokonta	Deuterostomia	Metazoa	Vertebrates	Primates	Ciliated
Uniprot	2476	968	788	4267	4528	1601	209
MyGeneFriends	2480	958	768	4154	4195	759	196

Les indicateurs de catégorie, une représentation minimaliste de l'histoire évolutive, trouvent leur place dans la section supérieure du site « Nom et mots clé », sous la forme d'un tag coloré (Figure 5-14). Cliquer sur le tag permet, en outre, d'avoir accès à l'ensemble des gènes répertoriés dans la catégorie. Ce tag évolutif permet ainsi un accès synthétique à une information évolutive facilement interprétable par l'utilisateur humain. Cependant, ma contribution essentielle au réseau MyGeneFriends réside dans la génération de liens d'amitié entre gènes.

5.4.3 Les relations d'amitié entre gènes : quand l'évolution fait des amis

Les relations d'amitié sont au centre de MyGeneFriends car elles permettent de proposer aux utilisateurs de s'intéresser à d'autres acteurs que ses amis actuels, en se basant sur le réseau d'amitié de ceux-ci. Les relations d'amitié entre gènes sont de quatre types : les associations fonctionnelles, les données d'interaction protéine-protéine, les liens sociaux et finalement, les informations évolutives. Je vais dans un premier temps décrire ces types de relation avant de préciser comment les informations évolutives, basées sur les distances entre profils les complètent.

Les relations d'amitié fonctionnelle sont basées sur les termes GO selon deux modalités. La première est une comparaison simple basée sur le nombre de termes que deux gènes ont en commun : l'hypothèse étant que plus ce nombre est élevé plus la similarité fonctionnelle est importante. Pour éviter des biais dus à la fois à la structure hiérarchique de GO (un gène annoté avec un terme l'est également avec ses termes parents) et à la surreprésentation de certaines annotations, la deuxième modalité utilise une mesure de Similarité Fonctionnelle Sémantique (Reyes-Palomares et al., 2013), prenant en compte la spécificité de chaque terme (plus élevée si peu de gènes sont annotés avec ce terme). Sur le principe des liens fonctionnels, des relations sont également définies entre les gènes responsables de même maladie, le poids étant là encore proportionnel au nombre de pathologies en commun entre les gènes. Basées sur des annotations fonctionnelles ou pathologiques connues, ces relations se veulent une façon d'associer les gènes impliqués dans les mêmes processus fonctionnels.

De façon à décrire les gènes sous un autre angle, le second type d'amitié est basé sur des interactions protéine-protéine de la base de données STRING. De fait, ces prédictions d'interaction reposent déjà sur l'intégration de plusieurs données (génomique comparative, fouille de données), mais également sur des données d'interaction expérimentales. MyGeneFriends considère uniquement les interactions les plus fiables de STRING (dépassant un score de 700 sur 1000). Les données d'interaction permettent d'avoir accès aux protéines faisant partie des mêmes systèmes biologiques ou de systèmes proches, sans reposer directement sur les annotations fonctionnelles.

Le lien d'amitié social exploite directement l'aspect « réseau social » de MyGeneFriends, en reliant les gènes qui sont « amis » d'un même utilisateur humain, pour un sujet de recherche donné. Ainsi, les utilisateurs, experts, agissent sur la structure du réseau et permettent par leur comportement d'ajouter de l'information sur les relations entre les gènes.

Ici, les liens d'amitié entre gènes reposent sur plusieurs sources de données, dont les informations peuvent se recouper pour renforcer la pertinence du lien d'amitié, tout en apportant leurs éclairages spécifiques. Un marqueur évolutif, à travers les proximités entre profils phylogénétiques, apporte dans cette logique une nouvelle sensibilité, complémentaire aux précédentes. Les relations évolutives entre gènes, disponibles sous la section 'Orthology' des liens d'amitiés reposent sur les distances de Jaccard entre les profils phylogénétiques des protéines humaines dans un panel de 100 espèces représentatives de la diversité des Eucaryotes. Le seuil de similarité servant à définir les relations d'amitié sur cette base, est comme dans OrthoInspector, fixé à une valeur de 0.4.

Dans MyGeneFriends, les relations d'amitié d'un gène peuvent, comme je l'ai mentionné plus tôt, être consultées sur la page de profil du gène, mais également être visualisées dans le contexte du sujet de recherche (*Topic*) de l'utilisateur. L'apport des différentes sources y est alors clairement visible. Sur mon profil personnel MyGeneFriends, l'un des Topics regroupe les gènes associés à une classe de ciliopathies, les Dyskinésies Ciliaires Primitives, La Figure 5-15 représente les liens d'amitiés entre ces gènes.

Les 33 gènes de ce *Topic* sont tous responsables de la même classe de maladies, qui est également au cœur des thématiques de notre laboratoire. Ils sont donc tous liés par des relations de type « maladie en commun » (bleues) et fortement connectés par des relations sociale (rouges). En filtrant ces relations, on isole un groupe de 14 gènes associés par des relations STRING et des relations évolutives, avec peu de recouvrement entre les deux catégories. Sur les 33 gènes, 19 n'ont aucune relation sur cette base (*singletons*). Partant de ce constat, j'ai étendu le réseau des gènes associés à des relations évolutives, en ajoutant au *Topic* tous les « amis évolutifs » des gènes déjà reliés par un lien d'amitié de ce genre.

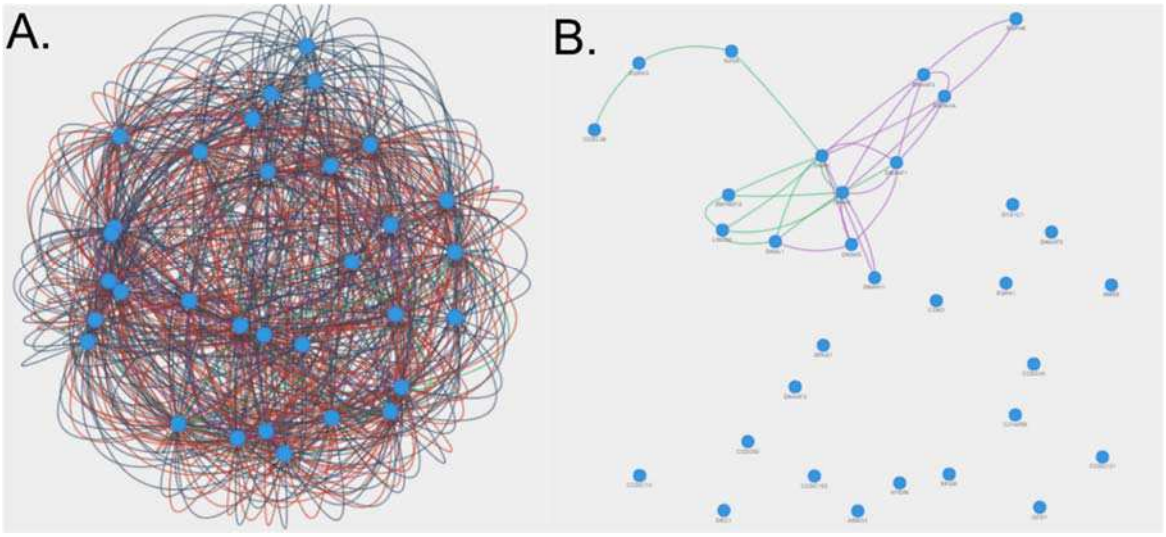


Figure 5-15 **Relations entre les gènes associés aux Dyskinésies Ciliaires Primitives**. Les nœuds représentent les gènes et les arrêtes les liens entre gènes. A. Tous les liens sont représentés, dont les liens sociaux (en rouge) et de maladies (en bleu). B. Seuls les liens évolutifs (en vert) et les liens d'interaction protéine-protéine (STRING, violets) sont représentés.

Cet ajout de 15 nouveaux gènes augmente sensiblement les connexions du réseau (Figure 5-16). Ainsi, 4 des gènes *singletons* du premier réseau se retrouvent également associés au réseau principal. L'augmentation de la connexion repose évidemment sur les liens évolutifs, mais aussi sur les liens issus de STRING pour 4 des nouveaux gènes, preuve supplémentaire qu'ils participent bien d'un même système biologique et d'un certain recouvrement entre les relations issues de différentes sources.

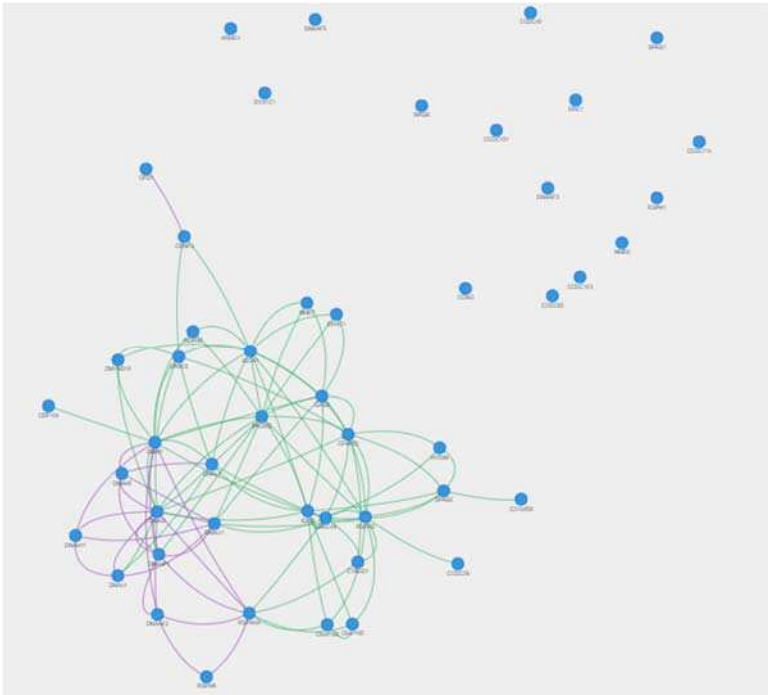


Figure 5-16 **Réseau étendu des gènes associés aux Dyskinésies Ciliaires Primitives**. Réseau de gènes, étendu par le biais des données évolutives. Les relations d'amitié basées sur les interactions protéine-protéine (STRING, en violet) et l'évolution (en vert) sont représentées.

Les nouveaux gènes sont fonctionnellement pertinents pour l'étude de ce processus. A titre d'exemple, le gène *Cilia and flagellated protein 52*, comme son nom l'indique est lié fonctionnellement au cil et est annoté avec le tag évolutif *ciliated* dans MyGeneFriends. Les annotations qui le concernent impliquent un rôle dans la motilité. A ce titre, ce gène, comme les autres retrouvés par la similarité évolutive, peut devenir une cible à prioriser dans la recherche de nouveaux gènes associés à cette maladie.

Comme le montre cet exemple, les similarités de profils intégrées à MyGeneFriends permettent de décrire le paysage des processus biologiques à l'origine de maladies génétiques. Ici, le concept de marqueur évolutif prend tout son sens, il est un descripteur parmi d'autres pour comprendre le lien entre génotype et phénotype.

5.5 Discussion et futures directions

Les profils phylogénétiques, nous l'avons vu, sont une représentation simple mais synthétique permettant d'extraire de la connaissance de la masse de données génomiques disponibles et d'explorer le Vivant. Les outils que j'ai développés sont un moyen d'exploiter ces marqueurs évolutifs pour mieux comprendre le rapport entre l'évolution des gènes et leurs fonctions, et en définitive le phénotype auquel ils aboutissent. Cette exploitation s'est faite selon plusieurs modalités complémentaires, pour passer de l'histoire évolutive à la fonction et *vice-versa* dans le cadre d'OrthoInspector et pour complémenter d'autres données biologiques illustrant d'autres aspects de la réalité biologique, comme dans le cas de MyGeneFriends.

Vers des profils phylogénétiques continus
Ces exploitations reposent directement ou indirectement sur des profils phylogénétiques binaires. Ceux-ci reflètent l'historique de gain et de perte de gènes, deux éléments moteurs de l'histoire évolutive. Il est possible de générer des profils phylogénétiques en principe plus informatifs, en remplaçant l'indicateur de présence par des données quantitatives. Cela rendrait possible d'identifier des gènes associés à un génotype particulier dont la perte ne s'est pas traduite par une perte complète du gène mais, par exemple, par une néofonctionnalisation accompagnée par une forte variation de séquence. On peut ainsi envisager d'intégrer aux profils phylogénétiques, les données de similarité de séquences entre orthologues, pour détecter des relâchements ou des renforcements de pression de sélection sur un gène. L'utilisation de profils continus est envisageable dans OrthoInspector en adaptant les outils, notamment en remplaçant la contrainte d'absence stricte par un seuil de conservation et en utilisant des mesures de distances adaptées.

La conception de profils phylogénétiques continus soulève cependant d'autres problèmes en termes de qualité de données et de définition des similarités de séquences. Les profils reposant sur la similarité de séquences nécessitent de disposer de séquences de gènes et de protéines exactes, là où les profils phylogénétiques simples ne nécessitent « que » des protéomes complets et des relations d'orthologie correctes, et tolèrent donc les erreurs de séquences. Cette problématique se pose surtout pour les séquences protéiques d'une majorité d'eucaryotes, dont la prédiction des bornes intron-exon constitue une source commune d'erreur de détermination

de la séquence. L'expérience que j'ai acquise au cours de ma thèse me pousse à la prudence quant à l'automatisation des comparaisons de séquences à grande échelle et à considérer l'hypothèse du profilage continu seulement dans le cadre d'un contrôle renforcé de la qualité des séquences.

Outre la qualité des séquences, l'autre problème associé aux profils continus est la mesure de similarité de séquences. Comme nous l'avons vu dans l'introduction, les profils continus reposent habituellement sur les scores de BLAST entre deux protéines, rapides à obtenir et reflétant la similarité de séquences de deux protéines sur un alignement local. Il me parait important de prendre en compte l'ensemble de la protéine afin de comparer les protéines sur une base équivalente. Je propose donc de mesurer la similarité de séquences sur des alignements multiples globaux. Ces mesures de similarité devront ensuite être normalisées afin de prendre en compte les différences de vitesses d'évolution entre chaque espèce. Ces points de protocole sont envisageables en utilisant comme base OrthoInspector tel qu'il est conçu actuellement. La construction des alignements multiples étant couteuse en termes de calculs, les travaux prévus dans ce cadre seront à envisager, dans un premier temps, sur un nombre d'espèces réduit.

Sur les profils binaires en tant que tels, il est également possible d'affiner les développements que j'ai réalisés. Les distances intégrées dans OrthoInspector 3.0 comme dans MyGeneFriends sont basées sur des méthodes naïves (voir Introduction) qui ne prennent pas en compte les relations taxonomiques entre chaque espèce considérée. Une sélection équilibrée d'espèces permet en principe de réduire ces biais et les comparaisons basées sur cette distance donnent des résultats satisfaisants. Il sera, cependant, intéressant par la suite de proposer d'autres mesures de distances comme alternatives à la distance de Jaccard et plus spécifiquement, d'adopter des mesures prenant implicitement en compte les relations taxonomiques telles que les mesures sur des vecteurs de transitions. Dans tous les cas, il ne s'agit pas de remplacer les mesures existantes, l'utilisation de relations taxonomiques comprenant également des incertitudes, mais de complémenter les mesures déjà existantes.

Renforcer la complémentarité

Les outils que j'ai développés autour des profils analysent la fonction des gènes sous la forme de termes *Gene Ontology*. Il s'agit d'un outil efficace pour étudier les différents aspects fonctionnels du Vivant, mais d'autres outils existent pour s'intéresser à des aspects plus spécifiques de la fonction biologique. L'outil de visualisation des profils fonctionnels n'est, dans sa conception, pas attaché uniquement aux termes *Gene Ontology* et peut présenter la distribution de toutes protéines d'une liste donnée. Aussi, d'autres annotations pourront être intégrées à cet outil, je pense principalement aux descripteurs de voies métaboliques ou de signalisation que l'on peut trouver dans les bases de données KEGG et Réactome et, plus spécifiquement pour l'homme, les descripteurs de phénotypes HPO. L'objectif étant de généraliser l'analyse par génomique comparative à toute question biologique, l'essentiel sera en définitive de permettre à tout utilisateur d'utiliser sa propre liste de protéines, par exemple issues de ses propres expériences omiques, dans une logique affirmée de complémentarité.

De même, l'intégration des mesures de similarité de profils entre gènes à d'autres sources de données, est une voie prometteuse pour formaliser le marqueur évolutif dans un contexte de comparaison de données. Ces distances se prêtent particulièrement bien à la représentation sous forme de réseaux, tout comme les données de co-expression issues de transcriptomique, les données d'interaction protéine-protéine, ou encore des données de maladies. Les travaux réalisés ici, en particulier dans MyGeneFriends, posent la base pour l'exploitation de cette piste dans le but de décrire plus précisément les systèmes biologiques.

La complémentarité est *in fine*, un mot qui résume bien les perspectives ouvertes par le développement des différents outils conçus ici. Comme j'ai voulu le démontrer dans ce chapitre, les trois méthodes choisies pour explorer les profils phylogénétiques se complètent dans les analyses évolutives proposées par OrthoInspector 3.0. Les retours d'utilisation préliminaires de l'ensemble de ces outils, ainsi que ma propre expérience, soulignent le besoin d'une articulation plus efficace encore. Les futurs développements de ces outils viseront donc une complémentarité accrue : la visualisation des profils obtenus par les recherches par profil et par le profilage fonctionnel intégrera les mesures de distance, afin de permettre de regrouper automatiquement les profils très similaires. Sur le même principe, il est important qu'une recherche fonctionnelle ayant permis d'identifier un profil de distribution puisse générer directement les contraintes nécessaires pour une recherche par profil, sans passer par une énumération laborieuse des espèces ou clades présents et absents. Cet objectif de complémentarité guidera les évolutions futures des outils disponibles publiquement sur OrthoInspector 3.0.

En renforçant la complémentarité, on donne la possibilité d'étudier les relations entre marqueur évolutif et fonctions sur plusieurs niveaux. Ces études apportent de nombreuses opportunités pour l'étude des systèmes biologiques, ce que je démontre dans le prochain chapitre, avec l'application des concepts vus précédemment à l'étude des ciliopathies.

6 Applications aux ciliopathies

6.1 Introduction

Les ciliopathies, maladies génétiques liées à des atteintes des cils eucaryotes (voir Introduction, section 2.5), sont des pathologies très complexes sur le plan phénotypique. La compréhension des mécanismes impliqués est au centre d'efforts conséquents de la communauté. Les études de génomique comparative ont contribué à plusieurs reprises à l'identification de gènes ciliaires et à leur caractérisation plus complète. Ces succès s'expliquent par l'histoire évolutive très particulière de cette organelle eucaryote qui, au cours de l'évolution, a subi de nombreuses pertes qui se retrouvent au niveau de la distribution atypique des gènes ciliaires. Les études précédentes de génomique comparative prenant en compte une diversité relativement réduite d'espèces, je me suis intéressé à la façon dont un plus large panel d'espèces pouvait aider à une meilleure caractérisation des gènes ciliaires. On retrouve donc dans ce projet les deux grands axes de ma thèse, l'importance de la sélection et de la représentation des données et le potentiel apporté par la complémentarité de différentes approches.

Notre étude du cil a précédé et nourri le développement d'OrthoInspector 3.0. Elle reposait sur le panel des 259 eucaryotes d'OrthoInspector 2.0. La mise en place de l'étude par profilage phylogénétique a dans un premier temps nécessité une réflexion sur la sélection des espèces, de façon à éviter le biais de surreprésentation de certains clades, notamment celui des champignons. Dans le cadre de l'étude du cil, nous avons mis l'accent sur l'intégration, pour chaque clade, d'espèces ciliées et non ciliées en s'appuyant sur un travail d'analyse bibliographique préalable. La sélection des données est une chose, leur représentation est tout aussi importante. Dans ce projet, nous avons également apporté un soin particulier à regrouper les profils en fonction de divisions taxonomiques. Ces deux aspects, de sélection et de représentation des profils phylogénétiques, ont permis de faciliter l'exploitation des profils comme leur visualisation.

Comme vu précédemment (voir section 2.4.3), l'analyse des profils phylogénétiques peut être réalisée de différentes façons : par la mise en place de contraintes de présence et d'absence, par des mesures de distances et par des méthodes de *Machine Learning*. Chacune de ces méthodes permet l'étude d'un même objet selon plusieurs aspects et ont chacune leurs avantages et leurs inconvénients. Dans le cadre de l'étude du cil, nous avons choisi de combiner plusieurs modalités d'analyse de ces profils pour identifier les gènes ciliaires. La mise en œuvre de ces différentes approches a orienté les développements des outils généraux du portail OrthoInspector et en particulier, la complémentarité entre recherche par profils et similarité de profils de gènes.

Un aspect particulièrement important et novateur de ce projet est l'étude à différents niveaux de granularité de l'histoire évolutive des gènes ciliaires. Cela a conduit à l'identification de différentes classes d'espèces ciliées en fonction de leur répertoire de gènes et a permis, en définitive, de formaliser des sous-groupes d'histoires évolutives, chacun correspondant à des

modules fonctionnels et à des catégories particulières de gènes ciliaires. Cet aspect souligne l'importance des allers-retours entre la fonction des gènes et leur histoire évolutive, un autre aspect important d'OrthoInspector 3.0.

A plusieurs niveaux, l'étude des gènes ciliaires a donc orienté la conception des marqueurs évolutifs ainsi que le développement des différents outils qui permettent de les analyser. Cette application complète des différents concepts a abouti à une caractérisation détaillée du cil et des ciliopathies sous l'aspect évolutif, avec l'identification de trois grands groupes de gènes ciliaires, et de 87 gènes peu caractérisés potentiellement liés à la fonction ciliaire, dont 5 ont été validés expérimentalement par nos collaborateurs. L'ensemble des méthodes utilisées et des résultats sont décrits en détail dans une publication (Nevers et al., 2017), dans le journal *Molecular Biology Evolution*, intégrée dans ce chapitre.

6.2 Publication: Insights into Ciliary Genes and Evolution from Multi-Level Phylogenetic Profiling

Insights into Ciliary Genes and Evolution from Multi-Level Phylogenetic Profiling

Yannis Nevers,*[,1] Megana K. Prasad,[2] Laetitia Poidevin,[1] Kirsley Chennen,[1] Alexis Allot,[1] Arnaud Kress,[1] Raymond Ripp,[1] Julie D. Thompson,[1] Hélène Dollfus,[2,3] Olivier Poch,[1] and Odile Lecompte*[,1]

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

[2]Laboratoire de Génétique Médicale, Institut de Génétique Médicale d'Alsace, INSERM U1112, Université de Strasbourg, Fédération de Médecine Translationnelle de Strasbourg (FMTS), Strasbourg, France

[3]Centre de Référence pour les Affections Rares en Génétique Ophtalmologique, Service de Génétique Médicale, Hôpitaux Universitaires de Strasbourg, Strasbourg, France

***Corresponding authors:** E-mails: yannis.nevers@etu.unistra.fr; odile.lecompte@unistra.fr.

Associate editor: Joel Dudley

Abstract

Cilia (flagella) are important eukaryotic organelles, present in the Last Eukaryotic Common Ancestor, and are involved in cell motility and integration of extracellular signals. Ciliary dysfunction causes a class of genetic diseases, known as ciliopathies, however current knowledge of the underlying mechanisms is still limited and a better characterization of genes is needed. As cilia have been lost independently several times during evolution and they are subject to important functional variation between species, ciliary genes can be investigated through comparative genomics. We performed phylogenetic profiling by predicting orthologs of human protein-coding genes in 100 eukaryotic species. The analysis integrated three independent methods to predict a consensus set of 274 ciliary genes, including 87 new promising candidates. A fine-grained analysis of the phylogenetic profiles allowed a partitioning of ciliary genes into modules with distinct evolutionary histories and ciliary functions (assembly, movement, centriole, etc.) and thus propagation of potential annotations to previously undocumented genes. The cilia/basal body localization was experimentally confirmed for five of these previously unannotated proteins (LRRC23, LRRC34, TEX9, WDR27, and BIVM), validating the relevance of our approach. Furthermore, our multi-level analysis sheds light on the core gene sets retained in gamete-only flagellates or Ecdysozoa for instance. By combining gene-centric and species-oriented analyses, this work reveals new ciliary and ciliopathy gene candidates and provides clues about the evolution of ciliary processes in the eukaryotic domain. Additionally, the positive and negative reference gene sets and the phylogenetic profile of human genes constructed during this study can be exploited in future work.

Key words: **cilium, ciliopathies, evolution, comparative genomics, phylogenetic profiling.**

Article

Introduction

Cilia, or flagella, are membrane bounded organelles that protrude from the cell surface in many eukaryotes and are the most common movement effectors of eukaryotic cells, as well as important centers of detection and integration of extracellular signals. Eukaryotic cilia generally share the same structural basis (fig. 1): a membrane-covered extension of the microtubule cytoskeleton, an axoneme, extending from the basal body in a characteristic structure of nine microtubule doublets, encircling a pair of microtubules (9 + 2). As they are not enclosed by a membrane, cilia maintain their compartmentalization by means of a complex macromolecular structure, the transition zone (TZ), which regulates the in-and-out of the organelle (Reiter et al. 2012; Avidor-Reiss and Leroux 2015). Molecular machinery for bi-directional transport, the intraflagellar transport (IFT-A and IFT-B) complexes, allows trafficking of structural components and other factors along the length of the organelles (Lechtreck 2015). Although these general principles hold in most ciliated organisms, there are many exceptions and overall, cilia are subject to an important variability within and between eukaryotic phyla, either in terms of number, length, position on the cell surface or structural and molecular composition (Moran et al. 2014; Carvalho-Santos et al. 2011).

Cilia diversity exists not only between species but also within a single organism, depending on developmental stage, cell or tissue type. In vertebrates, cilia are historically divided into two categories: motile and primary cilia, on the basis of their functions and axonemal structure. Motile cilia exhibit the classical 9 + 2 structure and adopt a variety of functions and numbers, e.g., the sperm flagellum allows gamete movement or the multiciliated epithelia (airway epithelium,

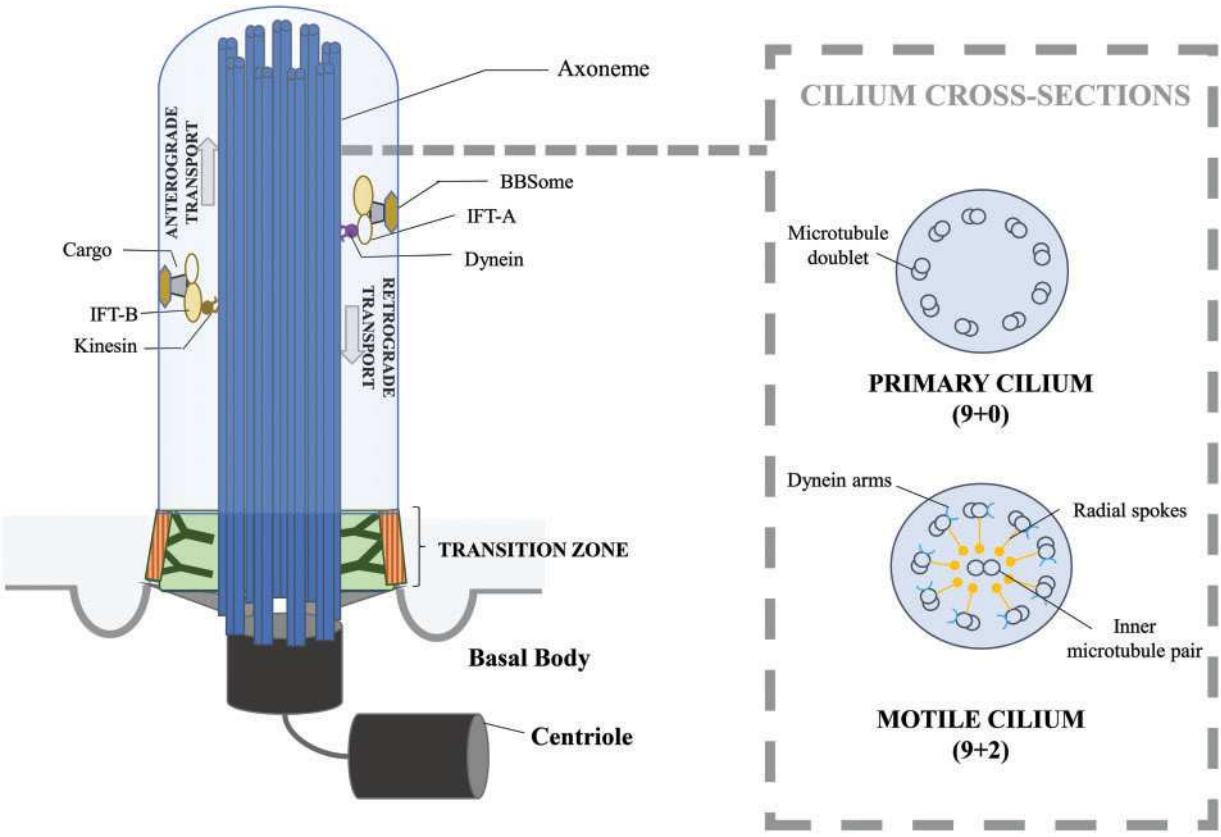


FIG. 1. Schematic representation of cilium. Structural components of a cilium and intraflagellar-transport machinery. The basal body is anchored at the membrane and is the basis of the axoneme. The transition zone filters the molecules that enter and leave the cilium, allowing maintenance of the organelle integrity. IFT-particles mediate the intracellular transport: IFT-B linked to kinesin in the anterograde direction, IFT-A and dynein in the retrograde direction. The cross-sections detail ultrastructural differences between primary and motile cilium, notably the molecular machinery allowing the movement for motile cilium.

epididymis, etc.) responsible for fluid flows along a tissue surface. Atypical structures of motile cilia have also been observed. For instance, in the embryonic node in the early stage of development, motile cilia are characterized by the absence of the central doublet (9 + 0 structure). This structure results in a particular circular beating pattern that is critical for the generation of left/right symmetry. Primary cilia are present in most cells and are nonmotile. Their axonemes adopt a 9 + 0 conformation (fig. 1) and lack all protein complexes responsible for movement. Because of its inability to perform movement, the primary cilium has long been overlooked as a vestigial organelle with no essential function. However, since the discovery of its implication in human diseases (Pazour et al. 2000), its functional roles have been deeply investigated and it is now considered as an essential extracellular sensor allowing integration of chemical, osmotic, mechanical or optical signals (Berbari et al. 2009). It is also critical for the regulation of important developmental pathways, such as Wnt (Lancaster et al. 2011) or Hedgehog (Breunig et al. 2008) signaling and is directly linked to cell cycle regulation (Ke and Yang 2014).

With such a diversity of ciliary functions, ciliary dysfunctions are linked to a wide range of human genetic diseases, known as ciliopathies, encompassing a large variety of symptoms (Badano et al. 2006). Notably, some classes of ciliopathies are linkable to defect in particular structures of cilium. A first class of ciliopathies, "primary ciliary dyskinesia," only affects motile cilia and leads to infertility, chronic sinopulmonary diseases and symmetry defects (Praveen et al. 2015). They are often caused by variation in genes coding for a particular molecular structure, that is, dynein arms and radial spokes (see fig. 1) that are at the origin of motile cilium beating, thus affecting motility. Other ciliopathies are linked to primary cilia dysfunction and affect specific organs and tissues, such as kidney (Kathem et al. 2014) in Polycystic Kidney Diseases or retina (Wheway et al. 2014) in Leber Congenital Amaurosis. The most pleiotropic cases, including Bardet-Biedl Syndrome (BBS) or Meckel Syndrome, lead to a wider range of symptoms and clinical signs, including those cited above but also obesity, diabetes, polydactyly, and other developmental defects. These pleiotropic ciliopathies can be associated with defects in particular molecular complexes and structures that are critical for cilium assembly and maintenance. Notably Jeune Syndrome and BBS are associated with defects in the intraflagellar-transport machinery (respectively, IFT-complexes and BBSome), whereas Meckel Syndrome and Joubert Syndrome are caused by dysfunction of TZ genes.

Although significant progress has been achieved towards understanding some general principles governing cilia and ciliopathies, there is still much to do to fully uncover all genes

and mechanisms linked to these diseases (e.g., causative genes are unknown for 50% of Joubert Syndrome cases [Parisi and Glass 2013]) and more generally, to understand ciliary processes at the molecular level. In this context, experimental studies have been undertaken to identify genes associated with cilia, notably a great number of transcriptomic and proteomic studies, in different species, tissues and conditions (indexed in the CilDb database [Arnaiz et al. 2014]). However, due to the inherent complexity of cilia and the functional variability between different biological contexts there is a limited overlap between the results of these experiments: for example, in the 55 cilia-centric high-throughput studies registered in CilDb (Arnaiz et al. 2014), only 21 genes are found in more than 20 studies.

Alternatively, *in silico* comparative genomics approaches have capitalized on the peculiar evolutionary history of the cilium. This organelle was present in the Last Eukaryotic Common Ancestor (LECA) and has experienced a profound diversification among eukaryotes, so much so that it is commonly used as a major determinant in eukaryotic classification (Adl et al. 2012). In addition, this ancestral organelle has been lost independently in many lineages, including most seed plants, most Fungi, or Amoebozoa. This particular distribution has motivated the application of phylogenetic profile approaches (Pellegrini et al. 1999) to discover ciliary genes present in ciliated species and absent in nonciliated ones. In 2004, a three-way comparison (Li et al. 2004) identified the essential ciliary gene BBS5 among genes present in human and a ciliated alga, but absent in a nonciliated land plant. Simultaneously, an *in silico* identification of *Drosophila melanogaster* genes conserved in five ciliated species and absent in three nonciliated ones (Avidor-Reiss et al. 2004) yielded ~200 candidate genes and 15 of them were experimentally validated. Since these pioneering studies, the number and diversity of available genomes have considerably increased, allowing a more accurate definition of phylogenetic profiles, critical for a better prediction of ciliary genes.

Extended profiles also yield the possibility to exploit the important structural and functional diversity of cilium between ciliated species, especially to find genes associated with cilia subfunctions that had been lost in a particular organism (e.g., Nematodes have lost the ability to construct motile cilia although they are able to develop sensory/primary cilium). Evolution of several iconic complexes, namely IFT-A, IFT-B, and BBSome (van Dam et al. 2013), involved in intraflagellar transport as well as Transition Zone proteins (Barker et al. 2014) have been extensively investigated in a wide panel of eukaryotic species resulting in significant advances in the understanding of cilia. However, this kind of work has, to date, never been done for all ciliary genes. Given that ciliopathies tend to be associated with particular complexes and submodules, and that phylogenetic profile analysis is suitable for the identification of a group of genes linked to diseases (Tabach et al. 2013), holistic studies leveraging the eukaryotic diversity of cilia should allow to identify evolutionary signatures of genes with links to the different classes of ciliopathies.

The intrinsic variability between evolutionary histories of ciliary genes means that it is necessary to distinguish biological heterogeneity from technical noise in both steps of the prediction protocol, that is, construction of the phylogenetic profiles and their subsequent analysis. In the construction step, involving the correct evaluation of the presence or absence of genes in diverse genomes, the choice of well covered and well annotated genome sequences is critical, as is the correct prediction of orthologs between divergent species. In the analysis step, the challenge lies in the definition of the protocol to accurately distinguish profiles corresponding to ciliary genes. Most previous studies (Avidor-Reiss et al. 2004; Li et al. 2004, p. 200; Merchant et al. 2007; Hodges et al. 2011) used scoring systems, which can be classified as knowledge-guided methods, to find genes with a phylogenetic profile correlating with the presence-absence of cilia. More recently, Dey and colleagues (Dey et al. 2015) used an automatic and agglomerative clustering method to identify, among others, modules enriched in known ciliary genes, whereas Li et al. proposed the CLIME algorithm (clustering by inferred models of evolution; Li et al. 2014) based on a training set of known ciliary genes and a phylogenetic tree to predict new ciliary genes via a Machine Learning approach. These different methods have all contributed to extending our knowledge about ciliary genes however it remains difficult to assess their respective strengths and weaknesses since they have been applied on different data sets, in terms of species, proteome versions and methods of orthology prediction.

To go further in terms of prediction and characterization of ciliary genes, we performed a comparative genomics study with optimized protocols and predicted a consensus set of 274 ciliary genes, including 87 new candidate genes with poorly defined function. We then proceeded to a species oriented analysis of the phylogenetic profiles, and identified categories of genes sharing identical evolutionary fate in Ecdysozoan (Nematodes and Arthropods) species. These categories correlate with both ciliary structures (IFT, TZ, and motility associated complexes) and known classes of ciliopathies. Furthermore, we identified a category covering an important number of genes that probably correspond to less well studied functional modules, and experimentally validated the ciliary localization for five of these genes.

Results

Phylogenetic Profiling of Human Genes

To predict the human genes implicated in ciliary processes, we established phylogenetic profiles for the 20,193 human protein-coding genes (supplementary table S1, Supplementary Material online) by searching for orthologs in 100 species chosen to represent a wide evolutionary diversity and to sample all major eukaryotic clades (Stramenopiles, Alveolata, Excavata, Archaeplastida, Amoebozoa, Fungi, and Holozoa). The panel includes 60 "ciliated species" (organisms that produce a cilium or a flagellum at some point in their life cycle) and 34 "nonciliated species." The six remaining species had no observed ciliated or flagellated stages, but presence of cilia cannot be excluded since the life cycle of these species is only partially deciphered. These six species will be referred to as "Potentially ciliated species" in the following sections.

Orthology relationships were predicted using OrthoInspector (Linard et al. 2015) and the presence/absence of orthologs recorded in a binary matrix where rows represent genes and columns represent species. To exploit the phylogenetic profiles and discriminate ancestral ciliary genes (i.e., genes present in a wide range of ciliated clades that were therefore likely to be present in the LECA) from the rest of the human genes, three independent protocols were tested and assessed using reference sets.

Definition of Reference Sets

Comparison of the independent approaches for ciliary gene identification requires objective criteria to estimate the sensitivity and specificity of each method. Obviously, these accuracy measures cannot be calculated directly in the absence of a prior exhaustive knowledge about ciliary genes. As a proxy, two nonexhaustive but high-confidence reference sets were designed: a positive set including genes of known ciliary function and a negative set of genes with known function and probably no implication in any ciliary process.

The set of positive ciliary genes was defined using the *Ciliary Gold Standard* (CGS) provided by the Syscilia consortium (van Dam et al. 2013). The CGS is an expert curated list of 302 genes for which the ciliary function is well documented and was designed to allow benchmarking of high-throughput and computational methodologies. We updated this list by adding 75 genes experimentally annotated with cilia-related Gene Ontology (GO) terms. The resulting positive gene set consists of 377 validated ciliary genes (supplementary table S2, Supplementary Material online).

Unfortunately, no established set of genes unrelated to cilia is currently available. One major reason for this is the difficulty associated with proving that a gene is not directly or indirectly involved in a particular process. Therefore, we created a list of genes unlikely to have any implication in cilia-related mechanisms. This list fulfills two criteria: i) selected genes were studied and functionally characterized and ii) they were not involved in a functional process linked to cilia. On this basis, we selected genes belonging to Reactome pathways in which there were no or few genes belonging to or interacting with members of our extended positive gene set (see Methods). This leads to a negative reference set of 971 genes that are likely to be unrelated to cilia and are functionally diverse, as they belong to 68 different Reactome pathways (supplementary table S3, Supplementary Material online). In the remaining sections, genes in the negative set that are detected by a ciliary gene prediction method will be referred to as "false positives."

To efficiently interpret the results obtained from these sets, it is important to note two elements. Firstly, the sets were defined using functional information without any input about the evolutionary histories of the genes. Thus, the 377 genes of the positive set did not necessarily exhibit a ciliary phylogenetic profile (presence in ciliated clades and absence in nonciliated). Indeed, 128 genes are recent innovations of Opisthokonts, Metazoa, or Vertebrates and 65 are involved in general processes and thus are not restricted to ciliated species (e.g., tubulins α and β; supplementary table S4, Supplementary Material online). Consequently, roughly 50% of the positive gene set (184 out of 377) are identifiable by comparative genomics. Secondly, as mentioned above, these two sets are far from being exhaustive, thus indicators derived from these reference sets cannot be used as absolute accuracy indicators, but rather as tools for comparing different methodologies.

Prediction of Ciliary Genes by Three Independent Methods

The first method used to predict ciliary genes was a knowledge-guided method aimed at identifying genes for which an ortholog is present in ciliated species and absent in nonciliated species. A strict binary analysis requiring presence in all ciliated species and absence in all nonciliated species is too stringent with regard to possible errors in genome annotations and the large number of studied organisms with a wide diversity of evolutionary histories. Therefore, genes were ranked according to a scoring metric that takes into account their distribution in ciliated/nonciliated species in each major eukaryotic clade (see Materials and Methods). Considering presence/absence information at the lineage level minimizes gene prediction inaccuracies, and in addition allows to attribute the same weight for each taxon in the prediction regardless of the number of genomes available (for example, only one representative of ciliated fungi was considered, but presence of genes in this species and not in nonciliated fungi should still be considered highly informative). Using this metric, 357 high scoring genes with a profile specific to ciliary species were selected among the 20,193 human protein-coding genes. As an indication of the accuracy of the method, 122 of these 357 genes (34.17%) belong to the positive data set, corresponding to a considerable enrichment in known ciliary genes (18.3-fold enrichment, $P = 8.8{*}10^{-126}$ one-tailed fisher exact test), whereas no gene belonging to the negative set (false positive) was detected.

The second prediction method relied on a hierarchical clustering of phylogenetic profiles for all human protein-coding genes. Using Pearson distances and the Ward algorithm, the profiles were automatically partitioned into 14 clusters (fig. 2) with cluster size ranging from 327 to 2,766 genes. The Gene Ontology term enrichment for each cluster was analyzed using Panther (Mi et al. 2016) to determine the associated biological processes, molecular functions, and cellular components (table 1).

Ten out of the 14 clusters correlate well with taxonomical divisions, ranging from Human or Primate specific genes to conserved eukaryotic genes (present in all studied species). Cluster 1 contains Human (or Primate) specific genes, probably originating from recent duplication events and accordingly, the best functional enrichment corresponds to keratinization, a fast-evolving process in Mammals and Primates (Gautam et al. 2015). Cluster 2 is composed of genes restricted to eutherian Mammals and is enriched in genes involved in olfactory reception. Cluster 3 encompasses genes ranging from Amniote-specific to Vertebrate-specific genes with an enrichment in genes linked to the immune system. Cluster 4 corresponds to Deuterostome-specific genes, with a

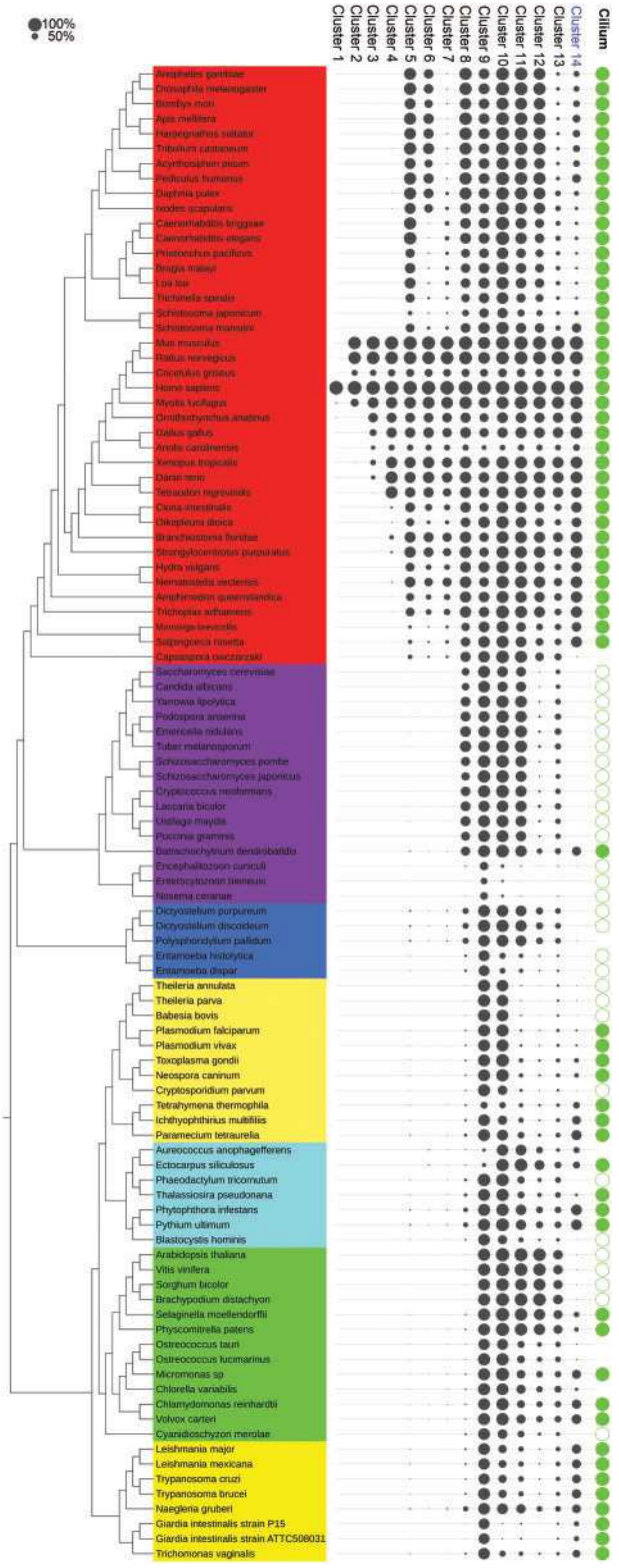


Fig. 2. Average phylogenetic profiles of each hierarchical cluster. Species names are colored according to major eukaryotic clades (from bottom to top: Excavata, Plants, Stramenopiles, Alveolata, Amoebozoa, Fungi, Holozoa, i.e., Metazoa, and their closest single-celled relatives). Each column corresponds to a cluster, ranked according to the taxonomical rank to which they correspond starting from human specific (cluster1) to "universal" eukaryotic genes (Cluster 9 and 10). Cluster from 11 to 14 do not correspond to a specific taxonomical division and are ordered by descending size. The

strong presence of Vertebrate-specific genes linked to cell–cell adhesion. Three clusters group Metazoa-specific genes (clusters 5, 6, and 7). Cluster 5 appears to be associated with multicellularity, an iconic particularity of Metazoa. Interestingly, clusters 6 and 7 differ from cluster 5 by gene losses in a subset of organisms: genes lost in nematodes for cluster 6 with an enrichment in transcription regulation and genes lost in Ecdysozoa (Nematodes and Arthropods) for cluster 7. Cluster 8 contains genes mainly present in Opisthokonts, without specific functional enrichment. Finally, clusters 9 and 10 correspond to genes detected in a majority of eukaryotic species, with cluster 10 exhibiting more losses in some amitochondrial parasitic organisms (Microsporidia, *Giardia* genus, *Entamoeba* genus). These clusters probably correspond to essential processes, as confirmed by enrichment in genes involved in ribosome-related roles and in nitrogen metabolic processes.

In contrast, four clusters exhibit a distribution uncorrelated to any specific taxonomical group, with genes presence in diverse organisms spread across the whole tree of life (clusters 11, 12, 13, and 14). Clusters 11, 12, and 13 appear heterogeneous and, accordingly, are not enriched in specific GO terms. Cluster 14 has a clearer distribution that correlates with the presence of a cilium. As expected, this atypical cluster of 327 genes is particularly enriched in GO terms related to the cilium, both in the Biological Process category with "cilium morphogenesis" (25.98-fold enrichment, *P* value: $2.85*10^{-91}$) and in the Cellular Component category with the term "cilium" (14.96-fold enrichment, *P* value: $1.84*10^{-98}$), with two of the most significant functional enrichments in all clusters. The detection of a cluster with these top-ranking enrichments outlines the outstanding evolutionary history of cilia-related genes compared with the rest of the human genes. This ciliary cluster of 327 genes includes 121 genes belonging to the positive reference set, and has a slightly better enrichment in known ciliary genes than that obtained with the first methodology (32.43%, 19.7-fold enrichment, *P*: $4.27*10^{-128}$ one-tailed fisher exact test), but includes two false positive genes (CHST15 and SEC31A) from the Negative Reference Set. Indeed, SEC31 (Core component of coat protein complex II involved in ER to Golgi transport) is conserved in six nonciliated species across different clades, whereas CHST15 (Carbohydrate sulfotransferase involved in chondroitin modification) is present in only four nonmetazoan ciliated species. Thus, while clustering exhibits a good sensitivity in predicting ciliary genes, the unsupervised nature of the method induces punctual losses of specificity.

Finally, the third method used to predict ciliary genes is based on the CLIME algorithm (Li et al. 2014). This program

Fig. 2 Continued
circle size is proportional to the percentage of genes from a given cluster with an ortholog found in a given species. The far right column "Cilium" indicates ciliated species (full green circle) and nonciliated species (empty circle). Species for which the existence of a ciliated state was unclear have no circle. Distribution of the Cluster 14 correlates to cilium distribution. This figure was generated using the iTol website (Letunic and Bork 2011).

Table 1. Gene Ontology Enrichment of the 14 Phylogenetic Profile Clusters.

Clusters	Genes	Biological Process		Molecular Function		Cellular component	
		GO terms	*P* value	GO terms	*P* value	GO terms	*P* value
1	2,151	Keratinization	$6.82*10^{-12}$	DNA binding	$3.52*10^{-14}$	Cornified envelope	$2.78*10^{-14}$
2	1,483	Detection of chemical stimulus involved in sensory perception	$1.39*10^{-115}$	Olfactory receptor activity	$1.32*10^{-108}$	Intrinsic component of membrane	$3.07*10^{-10}$
3	2,079	Immune response	$6.20*10^{-39}$	Transmembrane signaling receptor activity	$3.05*10^{-36}$	Intrinsic component of membrane	$4.60*10^{-32}$
4	2,766	Cell–cell adhesion via plasma-membrane adhesion molecules	$1.27*10^{-25}$	Glycosaminoglycan binding	$2.04*10^{-13}$	Proteinaceous Extracellular matrix	$7.45*10^{-23}$
5	2,614	Single-multicellular organism process	$5.96*10^{-49}$	Binding	$4.26*10^{-37}$	Intracellular	$4.69*10^{-47}$
6	2,302	Regulation of transcription from RNA polymerase II promoter	$5.16*10^{-19}$	Protein binding	$7.41*10^{-17}$	Intracellular part	$1.07*10^{-7}$
7	1,521	Sodium ion transport	$9.25*10^{-10}$	Sodium ion transmembrane transporter activity	$1.04*10^{-11}$	Transmembrane transporter complex	$2.35*10^{-7}$
8	491	Transport	$3.34*10^{-16}$	Catalytic activity	$1.06*10^{-33}$	Cytoplasmic part	$2.64*10^{-34}$
9	683	Ribosome biogenesis	$2.96*10^{-77}$	Organic cyclic compound binding	$2.60*10^{-72}$	Cytosol	$1.65*10^{-69}$
10	534	Cellular nitrogen compound metabolic process	$7.91*10^{-62}$	RNA binding	$1.69*10^{-71}$	Intracellular organelle part	$5.88*10^{-61}$
11	1,187	Cellular metabolic process	$2.26*10^{-77}$	Catalytic activity	$1.10*10^{-86}$	Intracellular part	$4.51*10^{-83}$
12	1,171	Metabolic process	$3.67*10^{-47}$	Catalytic activity	$6.63*10^{-34}$	Intracellular part	$8.79*10^{-37}$
13	884	Single-organism metabolic process	$2.90*10^{-44}$	Catalytic activity	$4.46*10^{-111}$	Cytoplasm	$5.59*10^{-22}$
14	327	Cilium morphogenesis	$2.85*10^{-91}$	Microtubule motor activity	$7.65*10^{-25}$	Cilium	$1.84*10^{-98}$

Fig. 3. Venn diagram of the genes predicted by three methods. Number of genes in each subcategory is indicated in black. True positives and false positives are indicated in green and red, respectively.

combines a training set of genes associated with the function of interest and a phylogenetic tree to generate a set of evolutionary models corresponding to phylogenetic profiles of the training set. The models are then used to search for other genes corresponding to these evolutionary models. The algorithm was run using the genes of the CGS as the training set. 178 genes were found in informative, extended, evolutionarily conserved modules. This set is considerably smaller than those obtained by the first two methods (357 and 327) but is also more enriched in ciliary genes, with 91 genes of the positive set (24.1%, 27.4-fold enrichment, $P = 3.6*10^{-113}$ one-tailed Fisher exact test). Importantly, no false positives were detected. With this absence of false positives and a strong enrichment in ciliary genes, CLIME appears to be a specific method, but it also produces a considerably smaller predicted set, and is thus the least sensitive of the three evaluated methods.

Consensus Approach to Define a Reliable Set of Candidate Ciliary Genes

For a more extensive comparison of the three prediction methods, we investigated the genes that were predicted to be cilia related by more than one method (fig. 3). The 431 genes (supplementary table S1, Supplementary Material online) predicted by at least one method include 131 true positives (out of the 184 identifiable by comparative genomics) and two false positives (found in the Negative Reference set). Of these 431 genes, a large core (157 genes) is predicted by all three methods, with more than half of them (87) being true positives. Additionally, 117 genes are predicted by two methods, including 29 genes in the positive set. A large majority of the detected true positives (116 out of 131, 89%) are thus predicted by at least two approaches. Almost all CLIME predictions, and all the confirmed positive predictions, are corroborated by at least another approach. Nevertheless, this method misses a significant portion of the true ciliary genes

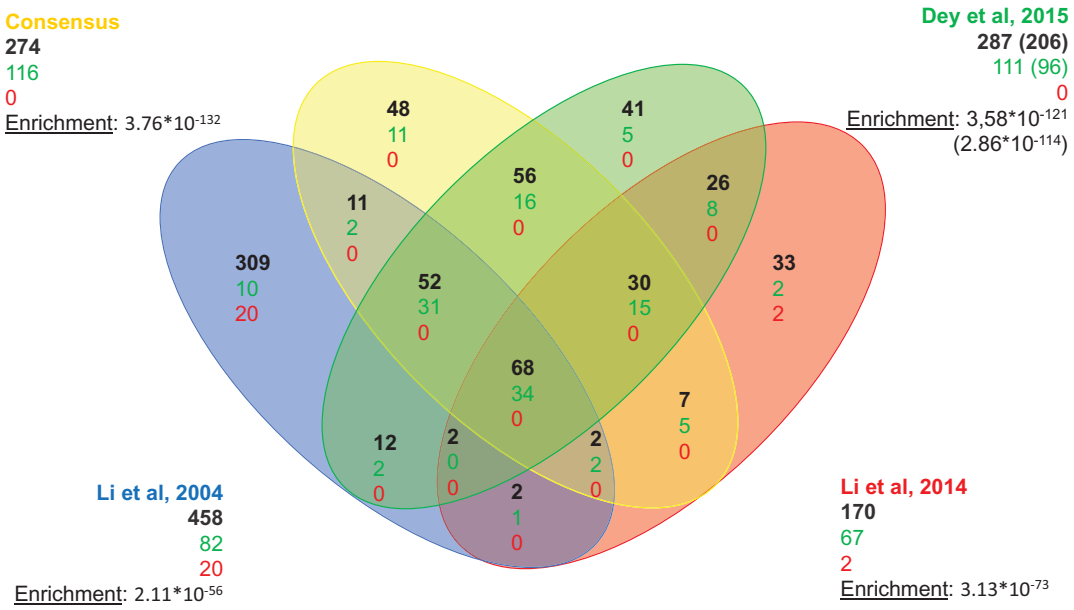


Fig. 4. Venn diagram of the genes predicted by this work and three previous studies. Number of genes in each subcategory is indicated in black. True positives and false positives are indicated in green and red, respectively. The corresponding enrichment is indicated by the hypergeometric *P* value. In the original work, genes predicted in Dey et al. (2015) were collapsed in orthology groups, with no distinction between paralogs. The numbers between parentheses correspond to the numbers presented in the original publications: the number of groups of genes predicted. The number of individual genes and true positives present in those groups is indicated with no parentheses and was used to determine overlap.

detected by other methods. In contrast, 94 and 59 genes are exclusively predicted by the knowledge-guided and clustering methods, respectively, with around 10% of true positives among both sets of specific predictions balanced by 3% of false positives in the clustering method. Generally, these two automatic methods, which do not need a training set, demonstrate a higher sensitivity than CLIME, balanced by a reduced specificity in the case of clustering.

The set of genes predicted by at least two methods, composed of 274 genes (supplementary table S1, Supplementary Material online), includes a large number of the positive genes (116/131), whereas excluding false positive containing sets. Moreover, the fact that a gene is predicted by two independent approaches adds reliability to the obtained prediction. Based on these criteria, we conclude that the combination of genes predicted by at least two methods in a consensus approach constitutes a robust and conservative prediction tool.

We compared this consensus set to other genomic comparative studies related to cilia in order to obtain a relative assessment of its quality (fig. 4) The list of genes found in the four studies and their overlap are available in supplementary table S4, Supplementary Material online). As expected, and apart from a core set of 68 genes predicted by all studies, recent studies outperform the founding study of Li et al, 2004 that predicted a reduced number of true ciliary genes (82) with regard to the larger number of predicted ciliary genes (458), and with a significant number of false positives (20). This study used comparisons of three organisms and orthology prediction based on BLAST hit analysis. Our comparison demonstrates the considerable gain obtained by more precise approaches and a wider range of genomes. Compared with more recent studies (Li et al. 2014; Dey et al. 2015), our consensus approach shows better results, in terms of both true positive prediction and enrichment in known ciliary genes, and as such, it constitutes an efficient basis for finding new candidates for ciliary function.

Scattered Distribution of Ciliary Genes among Ciliated Species

The 274 genes predicted by our consensus method were detected on the basis of their distribution in plant, protist, fungi, and metazoan taxa and thus correspond to ciliary genes specifically conserved in ciliated species and likely to have been present in the LECA. Thus, we will subsequently refer to them as "Ancestral Ciliary Genes" (ACG). However, as there is a wide variety of cilia, there is also diversity of the gene repertories of these ciliated species. Therefore, to gain insight into the evolution of ciliary genes and ciliary processes, the distribution of ACG was investigated in a species oriented manner and the studied species were hierarchically clustered on the basis of their presence–absence profiles for the 274 ACG. In the resulting hierarchical dendrogram, four major clusters can be distinguished (larger clusters with an Approximately Unbiased (AU)-criterion ≥ 0.90 after 10,000 bootstrap iterations, see Materials and Methods; fig. 5, supplementary fig. S1, Supplementary Material online).

As expected, most nonciliated species constitute a specific cluster (cluster α in fig. 5). These species are characterized by a quasi-absence of cilia-related genes, as they retain <16 genes from the predicted set. The species in this cluster include all representatives of seed plants and Rhodophyta (Archaeplastida), most nonciliated fungi, Amoebozoa, Stramenopiles, and one of the four nonciliated

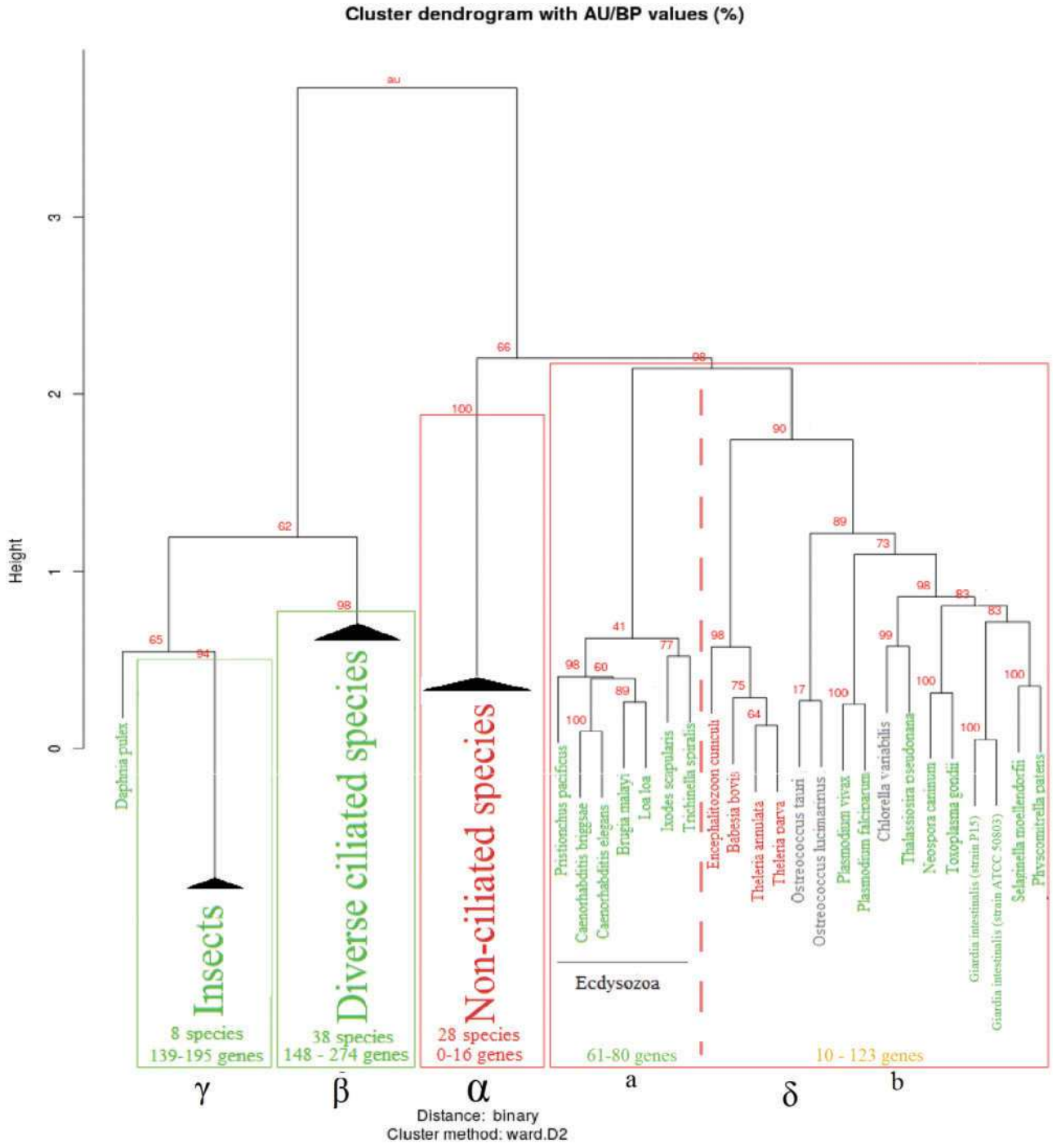


Fig. 5. Hierarchical clustering dendrogram of ciliary gene profiles in eukaryotic species. Homogeneous clusters are collapsed. The complete dendrogram is available in supplementary material, Supplementary Material online. Approximately Unbiased (AU) *P* value of each branch is annotated in red. Ciliated species and collapsed clusters with exclusively ciliated species are colored in green, nonciliated species and clusters in red. Species for which the ciliated state is unclear are in black. Significant clusters (AU $\geq$ 0.90) are framed by a red rectangle and annotated by a greek letter. Collapsed cluster are homogeneous in term of presence of cilia, and are annotated with the number of species they contain, and the minimal and maximal number of orthologs detected in these species. The cluster δ contains both ciliated and nonciliated species. It forks in two subclusters (separated by a dotted line): subcluster a corresponds to nematodes and ticks, representative of Ecdysozoa and subcluster b to a mixed group of eukaryotic species.

Apicomplexa (*Cryptosoporidum parvum*). However, they do not represent the entirety of nonciliated species in this study, as four nonciliated species are missing from the cluster: the microsporidia *Encephilitozoon cuniculi* and the three remaining nonciliated Apicomplexa: *Babesia bovis*, *Theileria parva*, and *Theileria annulata* (see below).

Two of the species identified as "Potentially Ciliated Species." *Capsaspora owczarzaki* and *Polysphondilium pallidum*, are also part of this cluster, with 16 and 12 genes, respectively. Such low gene counts and their position in the hierarchical clustering led us to hypothesize that a ciliated stage did not exist in these two species.

The three remaining clusters include representatives of ciliated species from all major taxa.

Cluster β contains species with an important number of ACG, ranging from 148 in the more extreme case (*Tetrahymena termophila*) up to 268 (*Mus musculus*) and 274 (*Homo sapiens*), with an average of more than 200 genes conserved in each species. The 38 species present in this cluster are representative of all major clades with ciliated species: Metazoa (with the notable exception of the Ecdysozoa taxon) and unicellular Choanoflagellate, Fungi (with *Batrachochitrium dendrobatridis*), Alveolata with all Ciliophora, most Stramenopiles, Archaepastidae with all ciliated Chlorophytes and a majority of Excavates. Notably, all the species can develop one or multiple cilia at some point during their vegetative life cycle. Finally, one species in the cluster is a "Potentially Ciliated Species," *Aureococcus anophagefferens*. Its genome conserves a large number of ACG (183), notably all components of IFT-A, BBSome and most components of IFT-B, and thus it is probable that this species also develops a flagellum during a yet to be described vegetative life stage.

Cluster γ is uniquely composed of ciliated insect species and is related to cluster β according to its position in the hierarchical clustering dendrogram. Additionally, *Daphnia pulex* is close to this cohesive cluster and could be considered as a member of this group, although the confidence value is lower, as the AU criterion of this extended cluster is only 65 (compared with 94 for insects only). This cluster of taxonomically related species is mainly characterized by a more reduced number of genes on average compared with the first cluster (from 139 to 195 ACG in insects, 118 in *Daphnia*), suggesting that insects, and more globally Pancrustacea, have been subject to important losses of ciliary genes during evolution.

The last cluster, δ, is characterized by an especially low number of genes related to cilia. According to the clustering dendrogram, it is partitioned into two subclusters. The first subcluster, δa, includes the four nonciliated species that were unexpectedly absent in the nonciliated cluster, as well as a mix of ciliated species representative of multiple clades with Embryophytes, ciliated Apicomplexa (Aconoidasida), Diplomonads and the Stramenopiles *Thalassiosira pseudonan*a. These species exhibit a number of ACG ranging from 48 for *Plasmodium falciparum* up to a maximum of 123 for *Selaginella moellendorffi*, which is a reduced set compared with other ciliated species representative of their clades that have approximately twice as many genes. The second subcluster, δb, is comprised of Ecdysozoa representatives (Nematodes and the tick *Ixodes scapularis*, an arthropod that belongs to the arachnid taxon) with between 61 and 80 ACG. Although the profiles of the insect and noninsect groups of Ecdysozoa are grouped separately in the hierarchical clustering dendrogram, they are related since nematodes lack most of the ACG absent in insects (see fig. 6).

As the species oriented clustering uncovers a divergence between ciliated species in terms of gene contents, we performed an in-depth investigation of the functional implications of these divergences.

Reduced Gene Set for Flagellum Restricted to Gametic Stage

Ciliated species in cluster δa possess few ciliary genes and come from a wide diversity of taxa. With the exception of *Giardia intestinalis*, comparison of the physiological characteristics shared by the ciliated species of the δa cluster revealed that they develop a flagellum (equivalent to the motile cilium) only during their gamete life stage and not during vegetative states of their life cycle, in contrast to their relatives belonging to cluster β and with a larger ACG set.

Despite the profile similarity revealed by the clustering, there is no consensus set of genes present in all of these organisms (profiles are described in detail in supplementary table S1, Supplementary Material online). However, two functional categories of genes seem to be more specifically retained: core components of the intraflagellar transport machinery (IFT-A and/or IFT-B) with important roles in cilia assembly and genes involved in motility of the flagellum. In the extreme case of *Plasmodium*, even the intraflagellar machinery components are lost.

Some of the "Potentially Ciliated Species" in this study are also present in cluster δ close to ciliated species, which suggests the presence or absence of cilium based on their position in the clustering dendrogram. *Chlorella variabilis* is close to *Thalassiosira pseudonana* and possesses 73 ciliary genes including some components of intraflagellar transport complexes and motility associated genes, suggesting the existence of a gametic flagellated stage. The case of the two *Ostreococcus* species is less clear: they cluster in the same group as the ciliated species with reduced gene sets, but seem to be in an intermediate position between nonciliated and ciliated species, in term of ACG number (30 and 34) and composition.

Ultimately, the nonciliated species in subcluster δa possess, as expected, very few ACG (1–23). However, it is worth noting that an important part of these correspond to dynein heavy chains involved in flagellum motility, one of the aforementioned well retained functional categories, which could explain their inclusion in this cluster. Dynein heavy chains in these four species (Kollmar 2016) vary greatly from all eukaryotic ones and thus are detected as orthologs of both cilium and cytoplasmic human dynein heavy chain, explaining this unexpected profile.

Functional Categorization of Ciliary Genes Based on Evolutionary Diversity among Metazoa

Even though all Metazoa are ciliated organisms, the species-oriented clustering reveals significant diversity in the gene repertoires in their clades. Specifically, Ecdysozoa representatives cluster apart from other Metazoa and are distributed in clusters δ and γ. This repartition can be attributed to massive gene losses in this taxon, implying that analysis of ciliary gene distribution between metazoan species is quite informative. Using the profile of presence/absence of genes in Ecdysozoa lineages, it is possible to identify four evolutionary modules within the 274 ACG (see Methods): 91 genes conserved in all metazoan lineages (referred to as the "all-Metazoa" module), 73 genes lost in Nematodes but conserved in a majority of

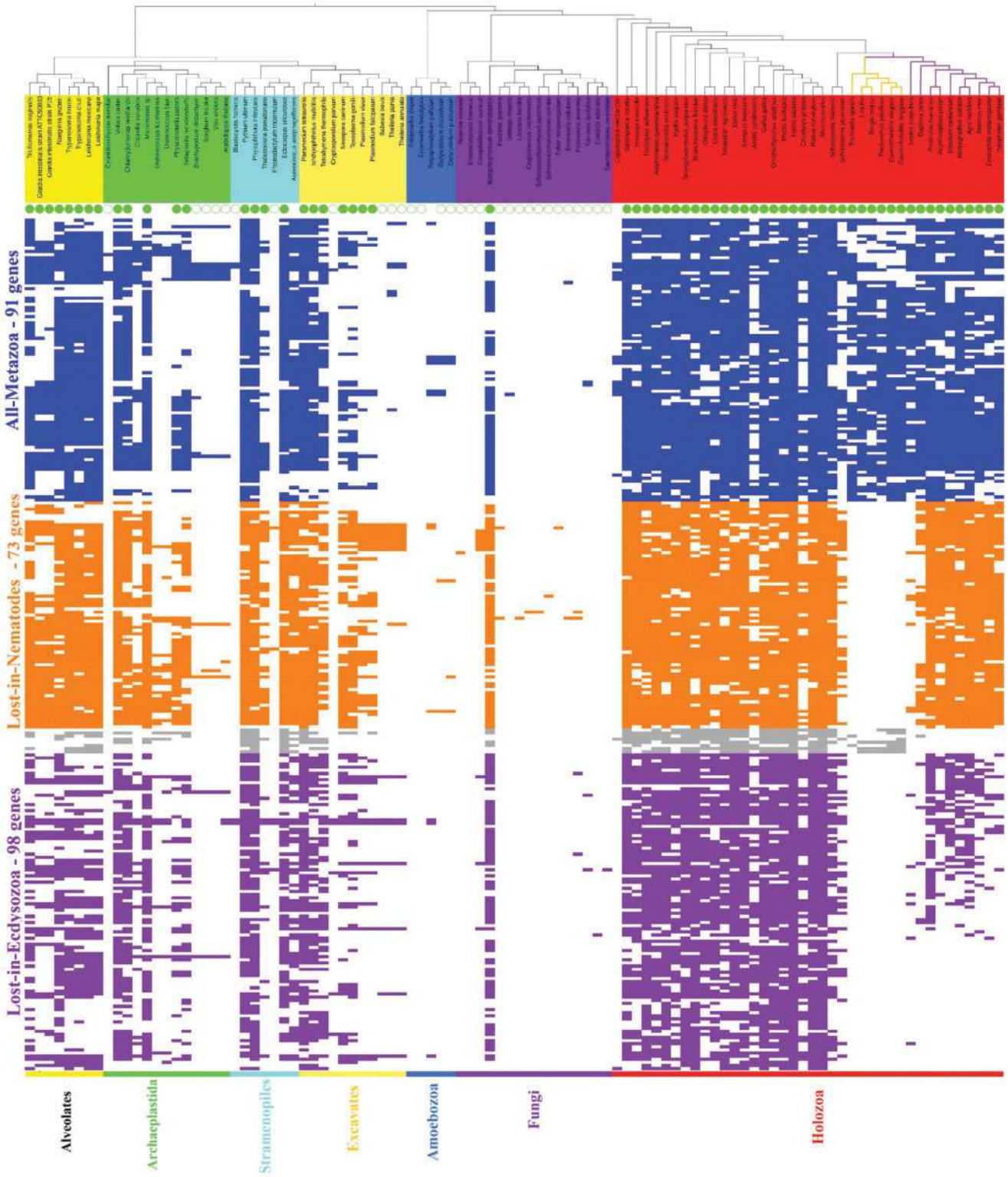


Fig. 6. Evolutionary modules based on divergence in metazoan phylogenetic profiles.Partition of phylogenetic profiles of the 274 ACG into four evolutionary modules. Species names are colored according to major eukaryotic clades (from left to right: Excavates, Plants, Stramenopiles, Alveolates, Amoebozoa, Fungi, and Holozoa). Ciliated species are annotated with a full green circle and nonciliated with an empty circle. Species for which existence of a ciliated stage is unclear have no circle. 274 binary profiles of presence (full colored rectangle) and absence (blank) of proteins colored according to their evolutionary modules: blue for proteins conserved in both nematodes and insects, orange for absence in all nematodes and presence in most insects, gray for presence in nematodes but absence in all insects (eight genes), purple for the rest (lost in most Ecdysozoa). This figure was generated using the iTol website (Letunic and Bork 2011).

insects ("lost-in-Nematodes" module), eight genes conserved in Nematodes but lost in insects ("lost-in-Insects") and finally 102 other genes lost in Nematodes and most other Ecdysozoa ("lost-in-Ecdysozoa" module; fig. 6). We performed a functional analysis of these modules to verify whether a potential association exists with particular ciliary functions.

The "all-Metazoa" module (blue in fig. 6) groups 91 ACG. Among these genes, 70 have a documented implication in cilia-related functions and/or ciliopathies. This set includes most components of highly studied complexes essential for cilium assembly, such as the intraflagellar transport complex or the transition zone. According to the guilt by association principle, it is probable than the remaining 21 genes with undocumented ciliary role belong to a functional pathway critical to cilia function. Intraflagellar transport is the mechanism transporting molecular cargo from the base to the tip of the cilium and vice-versa, and is critical for the molecular integrity of the cilium, as protein synthesis does not occur in the organelle. This function is mainly performed by three complexes, the IFT-A, IFT-B, and BBSome, whose disruption causes pleiotropic ciliopathies. Almost all components of these three complexes are found in the "all-Metazoa" module (5 out of 6, 14 out of 16, and 8 out of 9, respectively). In addition, the "all-Metazoa" module includes six (CC2D2A, TMEM67, AHI1, B9D1, B9D2, and TMEM231) of the 12 genes constituting the MKS complex (for a review, see [Garcia-Gonzalo and Reiter 2012]), an important component of the Transition Zone (TZ) that regulates trafficking from the cytosol to the cilium compartment. The MKS complex plays a critical role in cilium maintenance (Chih et al. 2012) and contains most of the genes implicated in two pleiotropic ciliopathies: Joubert Syndrome and Meckel Syndrome. Since these six genes are present in most ciliated species, notably the Ecdysozoa taxon, they probably correspond to components of LECA's ancestral Transition Zone complex and perform critical and conserved roles. These observations agree with the results of a more dedicated study of Transition Zone components across 52 eukaryotic species (Barker et al. 2014), in which five of these genes (with the exception of TMEM231) were identified as conserved "core components" of the TZ. Our results indicate that these five genes are part of a broader category of highly retained ciliary genes.

The "lost-in-Nematodes" module (orange in fig. 6) contains 73 ACG, including 59 genes with documented implication in cilia. Known genes in this set are linked to molecular complexes of motile cilia: dynein arms, radial spoke and central doublets of microtubules. These structures are specific to motile cilium and are necessary for proper motility. In fact, among the 274 ACG, 25 genes are annotated with the Gene Ontology term "cilium movement," among which 16 are part of the "lost-in-Nematodes" module (64%). Ciliary motile dysfunction is the cause of an important class of ciliopathies, known as the Primary Ciliary Dyskinesia. Accordingly, 22 of the 32 genes associated with these diseases in Orphanet (Orphanet website) belong to this module. This functional bias toward motile cilia for genes lost in Nematodes is not surprising, since the ability to construct motile cilia is lost in this clade (Ward et al. 1982). Interestingly, a proportion of genes lost in Nematodes are also lost in two Arthropod species that had unusual characteristics in our hierarchical clustering species dendrogram: the tick *Ixodes scapularis* (seven genes out of 73) and the water flea *Daphnia pulex* (35 genes out of 73). As genes of this module are present in insects, this distribution might explain our observations in the hierarchical dendrogram and suggests a partial loss of functional modules corresponding to motile cilium in these species, independent of the loss in nematodes.

The "lost-in-Insects" module (gray in fig. 6) is composed of eight ACG, including five poorly characterized ones, and it is thus difficult to attribute a specific role to this limited gene set. Nevertheless, it should be noted that two genes are part of the TZ: RPGRIP1L and NPHP4. These genes are part of the NPHP complex, an important component of the TZ, and their products have been reported to interact closely together and with the NPHP1 gene product (Sang et al. 2011). The NPHP1 gene is metazoan-specific and also lost in insects, suggesting the loss of the entire complex. Though loss of function of these genes has been linked to apical organization defects in Vertebrates, the precise role of the NPHP complex is not clearly established and the functional impact of the loss of the NPHP complex in insect cilia needs to be investigated.

Finally, gene profiles in the "lost-in-Ecdysozoa" module (violet in fig. 6) are heterogeneous and there is a continuous spectrum ranging from presence in some insects to absence in all of them, with no distinguishable intermediate stage. The loss is especially notable in the dipterans *Drosophila melanogaster* (91 genes lost out of the 102 genes) and *Anopheles gambiae* (93 out of 102). Characterization of this module is complicated by the number of genes of unknown function: roughly half of the 102 genes (56) have a documented ciliary function, of which 25 are well-validated ciliary genes (members of the positive reference set).

GO term analysis of the "lost-in-Ecdysozoa" module shows a significant enrichment in terms related to "microtubule cytoskeleton" and "centrosome" (6.67-fold enrichment, *P* value: 5.33^{*-17} and 10.15-fold enrichment, *P* value: $1.41 *10^{-14}$, respectively). Additionally, some genes are present in other previously mentioned complexes, notably in the IFT and BBSome complexes (LZTFL1 for BBSome, IFT27 and IFT25/HSPB11 for IFT-B). It has been shown that the products of these genes participate in these two complexes in human but are dispensable for complex stability, cilium assembly and intraflagellar transport (Bhogaraju et al. 2013). Interestingly, these three genes have a functional implication in the Sonic Hedgehog pathway that is mediated by primary cilia in Vertebrates. This pathway has been either lost or is mostly independent of cilia in the Ecdysozoa model organisms *Caenorhabditis elegans* and *D. melanogaster* (Ingham et al. 2011). Thus, while functional annotation of this module is less well defined than the first two, evidence points to its implication in the centriole and specific cilia-associated mechanisms.

Experimental Determination of Ciliary Localization for a Subset of Newly Identified Ciliary Genes

As previously mentioned, nearly half of the genes belonging to the "lost-in-Ecdysozoa" module are poorly characterized

(46 out of 102). Among the 274 ACG, they constitute an important proportion of the genes with no documented ciliary role (53%, 46 out of 87). The proportion of validated ciliary genes in this module is less important than in the others. This could be linked to a general bias from the scientific community, who tend to focus on genes conserved during evolution and present in *D. melanogaster* and *C. elegans*, two major animal models. However, as gene profiles in that module are less consensual, the possibility that the uncharacterized genes in this set are erroneous predictions might be a concern. To address this possibility and provide experimental support for our predictions, we tested the cellular localization in established *in vitro* models of ciliogenesis for the proteins corresponding to a subset of poorly documented genes belonging to the "lost-in-Ecdysozoa" module.

The genes were selected based on the commercial availability of reliable antibodies (supplementary table S5, Supplementary Material online). Ciliary localization of 17 corresponding proteins was tested by immunocytochemistry in human kidney proximal tubule epithelial cells (HK2) and human telomerase reverse-transcriptase immortalized retinal pigmented epithelial cells (hTERT-RPE1). We also tested antibodies that were predicted to cross react with the mouse protein in mouse inner medullary collecting duct (mIMCD3) cells. All these cell lines are established models of ciliogenesis (Rambhatla et al. 2002; Mai et al. 2005; van Rooijen et al. 2008).

Of these 17 antibodies, four were ciliary localized in HK2 cells: Leucine Rich Repeat Containing 34 (LRRC34), Leucine Rich Repeat Containing 23 (LRRC23), Testis Expressed 9 (TEX9), and WD40 Repeat 27 (WDR27; fig. 7*A*–*D*). All proteins were expressed along the entire length of the cilia in addition to some nuclear or cytoplasmic staining. The ciliary expression of LRRC34 was replicated in hTERT-RPE1 cells, (supplementary fig. S2, Supplementary Material online), where the protein also showed expression consistent with the cytoskeleton. Although the LRRC23 and TEX9 proteins could be detected in hTERT-RPE1 cells, protein expression did not colocalize with cilia (data not shown). WDR27 could not be detected in hTERT-RPE1 cells (data not shown), suggesting potential cell-specific functions of these proteins. Finally, upon testing a subset of the antibodies (those cross-reactive with the mouse antigen) in mIMCD3 cells, the Basic Immunoglobulin-like variable motif containing protein (Bivm) showed expression at the base of the cilium as shown in figure 7*E*, consistent with centriolar expression. However, this expression pattern was not seen in HK2 or hTERT-RPE1 cells, again suggesting either cell-specific or species-specific functions of this protein. None of the other tested proteins showed ciliary or centriolar localization in any of the cell lines. Thus, among the 17 tested genes belonging to the "lost-in-Ecdysozoa" module, our results confirm ciliary expression for five genes and the potential involvement in the centriole/cilium for genes predicted in the "lost-in-Ecdysozoa" module. It should be noted that absence of ciliary localization of the other tested genes does not rule out a potential ciliary role: they could either contribute to ciliary function in different ways or be specific to other cell types.

The exact function of these five genes could not be inferred on the basis of published information, given the sparse literature concerning them. However, both Leucine Rich Repeats (present in LLRC23 and LLRC34 with two and nine repeats, respectively) and WD40 Repeats (repeated 10-fold in WDR27) are structural elements involved in protein–protein interactions, suggesting a role in multiprotein complexes for LLRC23, LLRC34, and WDR27. It is worth noting that these repeats are frequent in ciliary proteins, notably WD40 Repeats are well-represented elements in IFT complexes (Cole 2003).

Discussion

Reference Data Sets for Future Analyses

Here, we performed phylogenetic profiling of all human genes in 100 eukaryotic species representative of major lineages of the tree of life using a robust and balanced method of orthology relationship inference coupled with three methods for phylogenetic profile analysis. We used these data to predict ciliary genes on the basis of their conservation in ciliated species and to characterize them in regard to their evolutionary history. This study focused on the cilium, an organelle with an unusual evolutionary history and involved in an emerging class of genetic diseases: the ciliopathies. However, the general principles and challenges concerning the phylogenetic profile approach, are likely to hold true regardless of the studied process. Thus, we hypothesize that dedicated studies based on our phylogenetic profiles could be performed for studying other cellular components and other genetic diseases.

To our knowledge, this is the first time that three independent methods have been compared and combined to predict gene sets with similar distributions over a wide range of eukaryotic species and to infer functional associations from phylogenetic profiling. The need for objective assessment criteria led us to develop positive and negative gene sets corresponding to ciliary processes. A well-defined positive set of known ciliary genes was already available under the form of a *Ciliary Gold Standard* provided by the *Syscilia* consortium and we updated this set using annotations derived from experiments for a total set of 377 genes. In contrast, a negative set of ciliary genes does not currently exist and its construction is an ongoing challenge for the community. As a general principle, construction of a negative set relative to biological processes is much more difficult than a positive one under the open-world assumption: we cannot dismiss the implication of a gene in a process if the only argument is that it was never observed before. Construction of a negative set linked to an organelle has been performed previously for mitochondria by selecting genes with GO annotations of other cellular compartments (Pagliarini et al. 2008). In the case of the cilium, using cellular localization criteria is not ideal due to its status as a nonenclosed organelle, with close links to plasma membranes, the cytoskeleton and cellular trafficking organelles (ER, Golgi). Instead we chose to discriminate them by functional processes, by selecting genes linked to Reactome pathways with no link to the cilium. Thus, the obtained set of 971 negative genes is fairly conservative and is linked to a wide

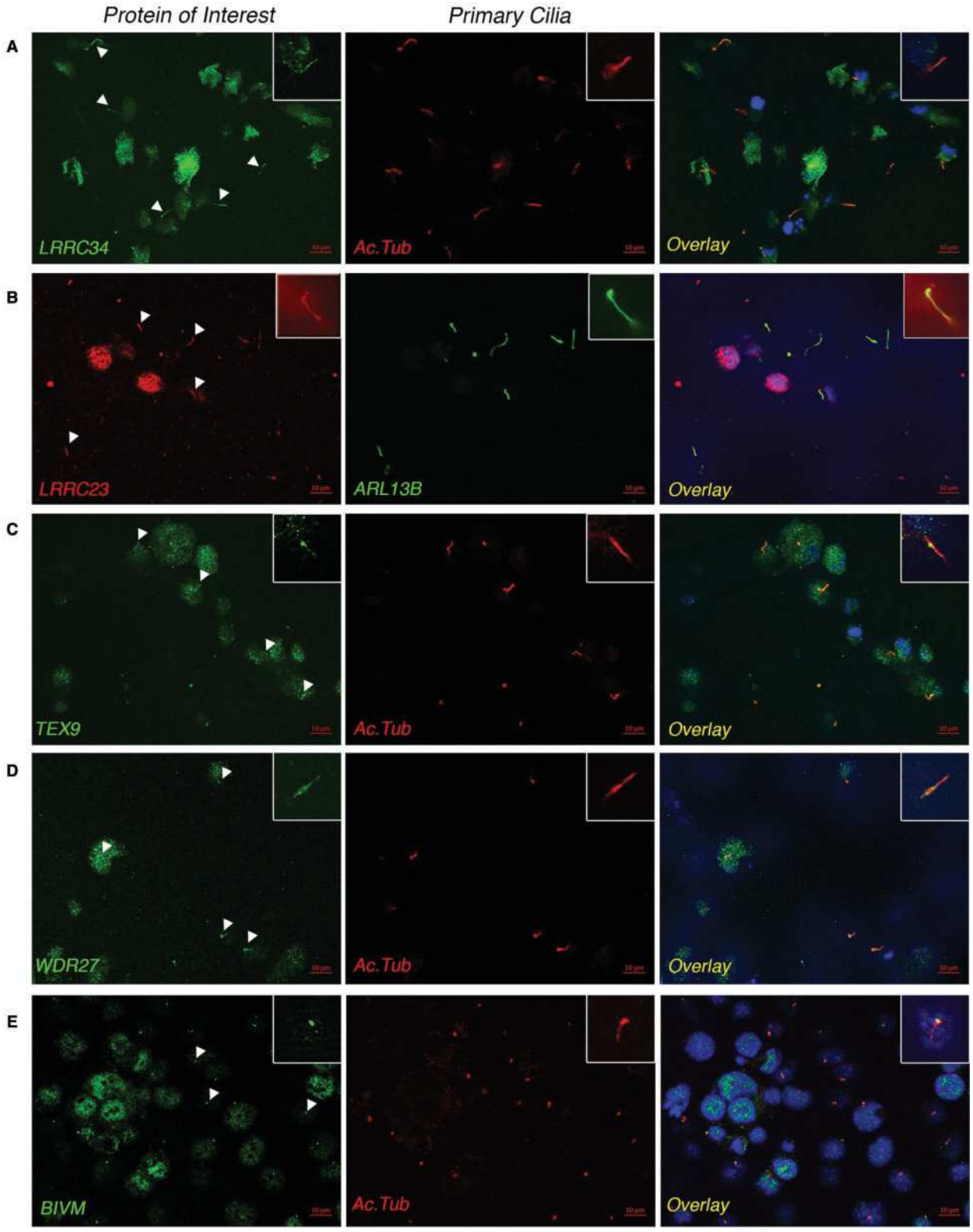


FIG. 7. Ciliary localization of five predicted ciliary proteins. Immunofluorescence of predicted ciliary proteins in ciliated mammalian cells. The cilia are labeled with antibodies targeting the ciliary markers ARL13B or acetylated tubulin (Ac Tub). (*A*–*D*) LRRC34, LRRC23, TEX9, and WDR27 colocalize with cilia in HK2 cells (arrowheads). (*E*) The mouse ortholog of BIVM localizes to the base of cilia (basal body and centriole, arrowheads) in mIMCD3 cells. Insets show a magnified view of a single cilium taken from an independent field of view.

variety of cellular processes. Obviously, this set is based on a snapshot of current knowledge and should be both increased and refined as our understanding of cilia advances.

Multi-Level Phylogenetic Profiling to Identify and Characterize Ciliary Genes

In the light of these two sets, it is possible to study the complexity and diversity of the evolutionary histories and biological processes related to cilia, by comparing the results obtained by three independent methods. Despite a significant overlap between the three methods (157 genes predicted by all three), there is significant variation between the predicted sets of ciliary genes (fig. 3), especially between results of the hierarchical clustering and the empirical method. Some genes are predicted exclusively by one of these two methods, including true positive genes that correspond to ancestral genes whose profiles differ slightly from a "canonical ciliary profile." This divergence could be caused by annotation errors in some genomes (technical noise), but also reveals the diversity of cilium protein contents in the eukaryotic domain underpinned by a variety of evolutionary histories (biological noise). Confronted with this heterogeneity, each method can be used to detect a particular set of genes that could not be simply described by a canonical profile. However, this is balanced by the risk of falsely predicting genes as belonging to a process, illustrated by the results of the clustering method that also produces two false positives. These issues outline the need for a comparison of multiple and independent prediction methods when analyzing phylogenetic profiling data, which are heterogeneous by nature. Combining the results of methods based on this comparison is thus an interesting way to broaden and cross-validate predictions. Here, by selecting the 274 genes predicted by at least two methods we were able to obtain a reasonable proportion of true positives found by at least one of the methods (116 out of 131) while completely avoiding false positives.

It is worth noting that we focused on ancestral and cilia-specific genes (genes conserved in a wide range of ciliated species) that are likely to be important for the functionality of the organelle since the LECA. This definition does not hold for all human genes involved in cilia and therefore does not cover the totality of the positive ciliary genes. In fact, of the 377 genes in our positive set, roughly 184 genes could be identified by comparative genomics. The remaining positive ciliary genes exhibit different evolutionary histories, which is again a tribute to the complexity of the cilium. Upon inspection of their phylogenetic profiles, it appears that these genes could be broadly classified into two categories (supplementary table S5, Supplementary Material online).

The first category corresponds to "universal" genes, that are present in most eukaryotes but are involved in other functions in addition to the ciliary-related ones and as such are also conserved in nonciliated species. We can cite for example the classical components of microtubules: subunits of α and β tubulins, nucleoporins and genes in the intracellular trafficking pathways (e.g., exocyst complex components). These "universal" genes do not exhibit cilia-related phylogenetic profiles, as orthologs of these genes are also present in nonciliated species. However, one can expect that, in nonciliated species, these genes might exhibit some imprints related to the loss of the process. In the future, it would be useful to complement an ortholog presence–absence profile with more precise evolutionary information, such as differential evolutionary rates or presence–absence of a particular protein domain, to allow to distinguish ciliary "universal" genes from completely unrelated ones.

The second category of ciliary genes without a ciliary phylogenetic profile corresponds to genes that emerged later during evolution in the common ancestor of Opisthokonts, Metazoa or Vertebrates. This is notably the case for genes linked to taxa-specific functions of cilia, like the hedgehog developmental pathways in Vertebrates or chaperones of the BBSome complex (BBS10 and BBS12) that emerged in metazoans. These genes have emerged too recently and are too related to clade-specific innovations to be characterized only by a broad evolutionary point-of-view. This highlights the importance of integrating data from different sources and approaches for identifying and studying ciliary genes, both derived from *in silico* studies and from leading-edge high-throughput studies such as the recent proteomic landscape by Boldt and colleagues (Boldt et al. 2016).

Nevertheless, analysis of phylogenetic profiles can leverage the inherent diversity of evolutionary histories to functionally categorize genes even at the level of more recent evolutionary events. In the case of our study, this fine-grained analysis allowed to identify a surprising pattern of gene losses in ecdysozoan species. These losses are especially striking as they concern genes retained since the LECA that might be critical for cilia functioning. It can also be used to classify genes into functional and evolutionary modules. Two of these modules have strong implications, both from a functional and from a disease-centric point of view. The first module, linked to pleiotropic ciliopathy classes, includes 91 genes that are well-conserved in metazoan taxa and enriched in central components of the most iconic cilium complexes (BBSome, IFT complexes, and TZ complexes). The second module, composed of 73 genes lost in nematodes, has a strong functional link to motile cilium and Primary Cilia Dyskinesia. This highlights the importance of investigating and exploiting the specificities of the profiles corresponding to a general process in order to identify functional modules.

However, as emphasized by the "lost-in-Ecdysozoa" gene set, analysis of such evolutionary modules might not be that straightforward, especially when numerous genes are present with few functional annotations. It should be noted that genes belonging to the "lost-in-Ecdysozoa" module exhibit heterogeneous distributions, notably in insect species. A more extensive taxonomic sampling of insect genomes and more generally of Ecdysozoa with better representation of their diverse taxonomic subdivisions, might allow a better distinction between gene losses and genes missed during the annotation process. More generally, the precision of phylogenetic profiling analyses will improve with the ever-increasing availability and diversity of sequenced genomes

(Škunca and Dessimoz 2015), provided that they are complete and sufficiently well annotated.

Evolution of Ciliary Processes Across Eukaryotic Species

As illustrated in this study, phylogenetic profiling is not only a way to characterize genes associated with a process by taking advantage of diverging evolutionary histories, but also to better understand the diversity of ciliary processes in various eukaryotic species. In this context, we performed species-oriented phylogenetic profile analyses to cluster ciliated species sharing similar gene repertories and, potentially, similar ciliary functions.

These analyses allowed us to identify a category of taxonomically unrelated species that possess a low number of ACG. They share the particularity of developing a flagellum uniquely during their gamete life stage and preferentially retained, although to a different degree, genes corresponding to the core machinery of intraflagellar transport (i.e., IFTA-A and IFT-B) and the motility-associated complexes (radial spoke and dynein arm). Thus, our study reveals that IFT genes are one of the most conserved cilia-associated gene sets in ciliated species. Overall, the diminution of the ciliary gene repertory in gamete-only flagellated species might correspond to reduced evolutionary constraints on the cilium due to the limitation of its role to the male gamete. In this context, intraflagellar transport and motility-associated genes seem to correspond to the minimal gene set necessary for the construction of a motile flagellum. In extreme cases, even these essential components can be partially degraded: *Toxoplasma gondii* and *Neospora caninum* have lost several components of IFT-B, whereas *Thalassiosira pseudonana* has lost the IFT-A complex and *Plasmodium* has lost the two complexes. These partial losses were previously recorded in a comparative genomic studies focused on intraflagellar-transport complexes, suggesting that they may be one of the last steps of complete ciliary loss (van Dam et al. 2013).

The species-oriented approach also allowed us to infer the existence of ciliated life stages in two species without observed cilia, namely *A. anophagefferens* and *C. variabilis* that conserve a significant number of ciliary genes. The presence of ciliary gene orthologs and the hypothesis that a ciliated life stage may exist were already mentioned during genome annotation of *C. variabilis* (Blanc et al. 2010) and other previous studies pointed out individual ciliary genes in the *A. anophagefferens* genome (Woodland and Fry 2008). Nevertheless, to our knowledge, this is the first time it was identified in a large-scale comparative genomics study. Moreover, comparison of the gene repertories with other ciliated species allows functional inferences. Indeed, the gene catalog of *A. anophagefferens* is coherent with the existence of a fully-fledged flagellum in a vegetative ciliated stage, whereas *C. variabilis* has a more reduced ciliary gene repertory akin to the repertories observed in species exhibiting only a gamete flagellum. Even though sexual reproduction was never observed in *C. variabilis*, it was noted during its genome annotation (Blanc et al. 2010) that it retains meiosis specific genes and will probably have an as yet unobserved haploid flagellated stage.

In Metazoa, our species-oriented phylogenetic profile analysis revealed massive losses of ACG in Ecdysozoa (Nematodes and Arthropods).This observation is in line with a general evolutionary trend, not specific to ciliary genes, as Ecdysozoan have suffered more extensive gene losses than other metazoan species (for a review, see Albalat and Cañestro 2016). In the case of ciliary genes, these losses seem to correlate with particular functions of cilium The 73 "Lost-in-Nematodes" genes are enriched in the Gene Ontology term "cilium movement," correlating with the well-known absence of sperm flagella and motile cilium in nematodes. Some genes of this module are also absent in the Arthropods *Daphnia pulex* and *Ixodes scapularis* that also exhibit aflagellate sperm (Morrow 2004). Surprisingly, given its aflagellate sperm, *Daphnia pulex* still conserves a number of genes involved in cilium movement and notably, orthologs of the dynein-arm component, a motile specific complex. To our knowledge, no direct observation of this complex has been made in *D. pulex*, but sensory cilia with dynein arms have been reported in other crustaceans (Geiselbrecht and Melzer 2014) and we hypothesize that the retained genes in *Daphnia* are used in similar structures.

In addition, 102 genes are lost in all or most ecdysozoan species. They seem to be functionally linked to microtubules and centrosome function, but many of them are yet undocumented, which can be partially explained by their absence in *D. melanogaster* and *C. elegans*, two of the classical models used for ciliogenesis studies. Sequence divergence between Ecdysozoa and other Metazoa has been previously noted for the centriole-associated protein PIX (Woodland and Fry 2008), but, according to the present study, the divergence is more general than previously suspected. As the author noted, the discrepancy might be associated with a centriole simplification in Ecdysozoa. Correspondingly, both centriole and flagellum axonemes are subject to an important diversity in insects and deviate from the canonical structure found in most ciliated eukaryotes (Mencarelli et al. 2008; Ross and Normark 2015). Regardless, the exact functions of most genes associated with this module have not been investigated, but they constitute promising targets for completing our current understanding of the cilium.

New Candidates and Confirmed Ciliary Genes

In this study, we identified 87 poorly characterized genes with ciliary phylogenetic profiles that constitute new targets. This includes 21 genes conserved in all metazoan taxa, a category rich in genes responsible for ciliary assembly and involved in pleiotropic ciliopathies, as well as 15 genes lost in nematode species, a category rich in motile genes linked to Primary Cilia Dyskinesia, the motility-associated ciliopathy. The link between these unknown genes and functional or disease-related modules could be used to orientate their functional investigation in the context of the cilium and help to prioritize them if rare variants are observed in still unresolved cases of ciliopathies. The remaining 46 poorly annotated genes are part of the "Lost in Ecdysozoa" modules. Ciliary localization of 17 of the corresponding proteins has been experimentally investigated and five proteins were physically present in the

cilium of ciliated mammalian cell lines. It is worth noting that this localization was observed in a cell type-specific context and in one case (BIVM) in a species-specific context. Obviously, these experiments do not prove that the remaining 12 genes do not play a role in cilium-related processes as they could be located in the cilium in other biological contexts (other cell types, specific development stages, etc.) or have an indirect impact on ciliary processes from another location, as is the case for the cytoplasm-located protein LZTFL1, whose defect is at the origin of Bardet–Biedl Syndrome (Seo et al. 2011). Nevertheless, whether the five new genes with confirmed ciliary location are at the basis of a specific class of ciliopathies is still to be determined.

Materials and Methods

Construction of Phylogenetic Profiles

Using OrthoInspector 2.0 (Linard et al. 2015), we predicted orthologs in 100 eukaryotic species of the 20,193 reviewed human protein-coding genes in the SwissProt database (February 2015; UniProt Consortium 2015). The chosen species were restricted to well annotated genomes and chosen to represent all major lineages of the eukaryote domain. Orthology predictions were used to construct a binary presence/absence matrix $X_{p,s}$, where an entry (g,s) is equal to 1 if at least one ortholog of human gene g was found in species s and 0 otherwise. Each matrix row defines a gene phylogenetic profile.

Definition of Validation Sets

To compare the accuracy of different ciliary gene prediction methods, we constructed positive and negative sets of ciliary genes. The positive set integrates the 302 genes from the *Ciliary Gold Standard* (van Dam et al. 2013) and an additional list of 75 genes carefully selected according to their Gene Ontology (GO) annotations (Gene Ontology Consortium 2015; November 2015). The additional ciliary genes were retained if they were annotated with GO terms related to cilia ("cilium," "cilium movement," "cilium organization," and "cilium morphogenesis") or their children, with experimental evidence codes (IMP, IGI, IPI, IDA, IEP, and EXP).

For the negative set, we determined a set of genes unlikely to be implicated in cilia, but representative of diverse functions by gathering genes from pathways with a minimum overlap with ciliary genes.

First, a list of genes directly or indirectly linked to cilia was created by adding to our positive gene set all genes interacting with a high confidence level (>0.7) with at least two ciliary genes in the version ten of the STRING database (Szklarczyk et al. 2015). From the 674 canonical pathways of the Reactome database (Croft et al. 2014), 68 pathways were retained containing no genes (for small pathways of <50 genes) or not more than a single gene (for larger pathway) from the extended ciliary gene set described above. Genes belonging to the 68 pathways unrelated to cilia constitute our negative set of 971 genes.

Analyses of Phylogenetic Profiles and Prediction of Ciliary Genes by Three Independent Methods

Knowledge-Guided Score

Knowledge-guided prediction of ciliary genes was performed under the hypothesis that these genes were differentially distributed among ciliated and nonciliated species within the following eukaryotic lineages: Stramenopiles, Archaeplastida, Alveolata, Excavata, Amoebozoa, and Fungi.

Accordingly, we assigned a score to all genes using the following:

$$\text{Score} = \sum_{\text{lineages}} \left(C_{\text{lineage}} - 2N_{\text{lineage}}\right)$$

Where:

$$C_{\text{lineage}} = \begin{cases} 1, & \textit{if the protein is present in at least 25\% of the ciliated species in the considered lineage} \\ 0, & \textit{otherwise} \end{cases}$$

$$N_{\text{lineage}} = \begin{cases} 1, & \textit{if the protein is present in at least 10\% of the non–ciliated species in the considered lineage} \\ 0, & \textit{otherwise} \end{cases}$$

The thresholds of 25% and 10% were chosen to establish the presence or absence of a gene on the basis of observations in several species, that is, to avoid noise due to prediction or annotation errors in some genomes and to accommodate variations among species. To avoid false positives, the threshold was more conservative for absence in nonciliated species.

Genes were defined as ciliary if they the score was $>=2$.

Hierarchical Clustering of Genes

The pairwise distance between two phylogenetic profiles x, y was estimated by the complement of the Pearson correlation coefficient $dr = 1 - corr(x,y)$ using the R package *amap*. Based on this pairwise distance, hierarchical clustering of all 20,193 phylogenetic profiles was performed using the Ward algorithm (Ward 1963), as implemented in the R function *hclust*. Gene clusters were defined as branches from the resulting dendrogram, using the dynamic tree-cutting algorithm (Langfelder et al. 2008).

CLIME Algorithm

The CLIME algorithm (Li et al. 2014) predicts genes sharing a given evolutionary history using a user provided gene training set, a binary species tree and a complete phylogenetic matrix. The training set is partitioned into Evolutionary Conserved Modules (ECM), according to the presence/absence profiles. Each module is then expanded by scanning the full matrix for

other genes with a similar profile. The predicted new genes constitute an extended ECM.

The species tree was extracted from the NCBI Taxonomy (Sayers et al. 2009; March 2015) and manually revised according to the literature to resolve cases of polytomy. For the ciliary training set, we used the *Ciliary Gold Standard*, a list of 302 expert-curated ciliary genes (van Dam et al. 2013).

Genes of extended ECM were selected as predicted ciliary genes when the corresponding ECM was flagged as "Informative" in the CLIME output.

GO Term Enrichment

All Gene Ontology term enrichments were realized using Panther (Mi et al. 2016) with the list of 20,193 human protein-coding genes as a background.

Species Profile Clustering

Clustering of phylogenetic profiles for the 274 ACG (Ancestral Ciliary Genes) was performed for 97 of the 100 eukaryotic species. *Anolis carolinensis*, *Cricetulus griseus*, and *Schistosoma japonicum* were excluded from the analysis due to errors in their genome annotation. Bootstrapped hierarchical clustering was realized in R, using the package *pvclust* (Suzuki and Shimodaira 2006) with binary distances, the Ward algorithm (Ward 1963) and 10,000 bootstrap replicates. Approximately Unbiased (AU) *P* values (Shimodaira 2004) were computed using multiscale bootstrapping. Clusters corresponding to branches with AU $\geq$ 0.9 were considered to be significant.

Definition of Evolutionary Modules

Predicted ciliary genes were divided into four evolutionary modules on the basis of their presence/absence profiles in Ecdysozoa. Genes present in at least one representative of both Nematodes and insects were assigned to the "all-Metazoa" module. Genes absent in all nematodes but well conserved in insects (present in six or more of the eight insects) were assigned to the "lost-in-Nematodes" module. Genes with the inverse distribution were grouped in a small "lost-in-Insects" module. The remaining genes were lost in most Ecdysozoa and grouped in the "lost-in-Ecdysozoa" module.

Cell Culture and Immunocytochemistry

Mouse inner medullary collecting duct (mIMCD3) cells (ATCC, CRL 212, USA), HK2 (human proximal tubule epithelial cell line) cells (ATCC, CRL-2190), and human telomerase reverse transcriptase immortalized-retinal pigmented epithelial cells (hTERT-RPE1, a kind gift from Dr. Séverine Bär, University of Strasbourg) were cultured in Dulbecco's modified Eagle medium (DMEM)-F12 (1:1) + GlutaMAX (Gibco, 31331, USA) with 10% fetal bovine serum (FBS, Gibco, 10500) and 1% Anti–Anti (Gibco, 15240-062) at 37 °C and 5% CO_2. To induce ciliogenesis, 2×10^5 HK2 cells were plated in each well of an 8-well Labtek chamber slide (Nunc, 177445, USA). Upon reaching confluence, the cells were cultured in media with 0.2% FBS for 48 h to induce ciliogenesis. To induce ciliogenesis in mIMCD3 cells, 2×10^5 cells were plated per well of a Labtek chamber slide and were cultured for 72 h postconfluence in 10% serum media. To induce ciliogenesis in hTERT-RPE1 cells, 1×10^5 cells were plated per well of an 8-well Labtek chamber slide and cultured in 0.2% FBS media for 72 h post-confluence.

For immunocytochemistry, HK2 and hTERT-RPE1 cells were fixed in 4% paraformaldehyde for 30 min and permeabilized in 0.2% Triton X for 10 min. mIMCD3 cells were fixed and permeabilized in ice-cold methanol. Cells were blocked in 5% bovine serum albumin (BSA) for 1 h at room temperature and then incubated with primary antibodies overnight at 4 °C at the concentrations listed in supplementary table S6, Supplementary Material online. Cells were then incubated with fluorescence-conjugated secondary antibodies (supplementary table S6, Supplementary Material online) diluted at 1:500 in 5% BSA for 1 h at room temperature. Finally, nuclei were stained with Hoechst dye (Life technologies, H3569, USA) and the slides were mounted in Vectashield (Vector Labs, H-1000, USA). Cells were visualized using the Axio Imager 2 (Carl Zeiss, Germany) and images were acquired and processed using the Zen microscope software.

Supplementary Material

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

Acknowledgments

This work was supported by the Agence Nationale de la Recherche (BIPBIP: ANR-10-BINF-03-02; ReNaBi-IFB: ANR-11-INBS-0013); the Fondation pour la Recherche Médicale (DBI20131228569); and Institute funds from the CNRS, the Université de Strasbourg and the Faculté de Médecine de Strasbourg. We thank Dr Séverine Bär (University of Strasbourg) for providing the hTERT-RPE1 cells, and the BICS and BISTRO bioinformatics platforms for informatics support.

References

Adl SM, Simpson AGB, Lane CE, Lukeš J, Bass D, Bowser SS, Brown MW, Burki F, Dunthorn M, Hampl V, et al. 2012. The revised classification of eukaryotes. *J. Eukaryot Microbiol*. 59:429–493.

Albalat R, Cañestro C. 2016. Evolution by gene loss. *Nat Rev Genet*. 17:379–391.

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

Avidor-Reiss T, Leroux MR. 2015. Shared and distinct mechanisms of compartmentalized and cytosolic ciliogenesis. *Curr Biol*. 25:R1143–R1150.

Avidor-Reiss T, Maer AM, Koundakjian E, Polyanovsky A, Keil T, Subramaniam S, Zuker CS. 2004. Decoding cilia function: defining specialized genes required for compartmentalized cilia biogenesis. *Cell* 117:527–539.

Badano JL, Mitsuma N, Beales PL, Katsanis N. 2006. The ciliopathies: an emerging class of human genetic disorders. *Annu Rev Genomics Hum Genet*. 7:125–148.

Barker AR, Renzaglia KS, Fry K, Dawe HR. 2014. Bioinformatic analysis of ciliary transition zone proteins reveals insights into the evolution of ciliopathy networks. *BMC Genomics* 15:531.

Berbari NF, O'Connor AK, Haycraft CJ, Yoder BK. 2009. The primary cilium as a complex signaling center. *Curr Biol*. 19:R526–R535.

Bhogaraju S, Engel BD, Lorentzen E. 2013. Intraflagellar transport complex structure and cargo interactions. *Cilia* 2:10.
Blanc G, Duncan G, Agarkova I, Borodovsky M, Gurnon J, Kuo A, Lindquist E, Lucas S, Pangilinan J, Polle J, et al. 2010. The *Chlorella variabilis* NC64A genome reveals adaptation to photosymbiosis, coevolution with viruses, and cryptic sex. *Plant Cell* 22:2943–2955.
Boldt K, van Reeuwijk J, Lu Q, Koutroumpas K, Nguyen T-MT, Texier Y, van Beersum SEC, Horn N, Willer JR, Mans DA, et al. 2016. An organelle-specific protein landscape identifies novel diseases and molecular mechanisms. *Nat Commun*. 7:11491.
Breunig JJ, Sarkisian MR, Arellano JI, Morozov YM, Ayoub AE, Sojitra S, Wang B, Flavell RA, Rakic P, Town T. 2008. Primary cilia regulate hippocampal neurogenesis by mediating sonic hedgehog signaling. *Proc Natl Acad Sci U S A*. 105:13127–13132.
Carvalho-Santos Z, Azimzadeh J, Pereira-Leal JB, Bettencourt-Dias M. 2011. Evolution: tracing the origins of centrioles, cilia, and flagella. *J Cell Biol*. 194:165–175.
Chih B, Liu P, Chinn Y, Chalouni C, Komuves LG, Hass PE, Sandoval W, Peterson AS. 2012. A ciliopathy complex at the transition zone protects the cilia as a privileged membrane domain. *Nat Cell Biol*. 14:61–72.
Cole DG. 2003. The intraflagellar transport machinery of *Chlamydomonas reinhardtii*. *Traffic Cph Den*. 4:435–442.
Croft D, Mundo AF, Haw R, Milacic M, Weiser J, Wu G, Caudy M, Garapati P, Gillespie M, Kamdar MR, et al. 2014. The Reactome pathway knowledgebase. *Nucleic Acids Res*. 42:D472–D477.
van Dam TJ, Wheway G, Slaats, GG, SYSCILIA Study Group, Huynen MA, Giles RH. 2013. The SYSCILIA gold standard (SCGSv1) of known ciliary components and its applications within a systems biology consortium. *Cilia* 2:7.
van Dam TJP, Townsend MJ, Turk M, Schlessinger A, Sali A, Field MC, Huynen MA. 2013. Evolution of modular intraflagellar transport from a coatomer-like progenitor. *Proc Natl Acad Sci U S A*. 110:6943–6948.
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.
Garcia-Gonzalo FR, Reiter JF. 2012. Scoring a backstage pass: mechanisms of ciliogenesis and ciliary access. *J Cell Biol*. 197:697–709.
Gautam P, Chaurasia A, Bhattacharya A, Grover R, Indian Genome Variation Consortium, Mukerji M, Natarajan VT. 2015. Population diversity and adaptive evolution in keratinization genes: impact of environment in shaping skin phenotypes. *Mol Biol Evol*. 32:555–573.
Geiselbrecht H, Melzer RR. 2014. Fine structure and ecdysis of mandibular sensilla associated with the lacinia mobilis in Neomysis integer (Leach, 1814) (Crustacea, Malacostraca, Peracarida). *Arthropod Struct Dev*. 43:221–230.
Gene Ontology Consortium. 2015. Gene Ontology Consortium: going forward. *Nucleic Acids Res*. 43:D1049–D1056.
Hodges ME, Wickstead B, Gull K, Langdale JA. 2011. Conservation of ciliary proteins in plants with no cilia. *BMC Plant Biol*. 11:185.
Ingham PW, Nakano Y, Seger C. 2011. Mechanisms and functions of Hedgehog signalling across the metazoa. *Nat Rev Genet*. 12:393–406.
Kathem SH, Mohieldin AM, Nauli SM. 2014. The roles of primary cilia in polycystic kidney disease. *AIMS Mol Sci*. 1:27–46.
Ke Y-N, Yang W-X. 2014. Primary cilium: an elaborate structure that blocks cell division?. *Gene* 547:175–185.
Kollmar M. 2016. Fine-tuning motile cilia and flagella: evolution of the dynein motor proteins from plants to humans at high resolution. *Mol Biol Evol*. 33:3249–3267.
Lancaster MA, Schroth J, Gleeson JG. 2011. Subcellular spatial regulation of canonical Wnt signalling at the primary cilium. *Nat Cell Biol*. 13:700–707.
Langfelder P, Zhang B, Horvath S. 2008. Defining clusters from a hierarchical cluster tree: the Dynamic Tree Cut package for R. *Bioinformatics* 24:719–720.
Lechtreck KF. 2015. IFT-cargo interactions and protein transport in cilia. *Trends Biochem Sci*. 40:765–778.
Letunic I, Bork P. 2011. Interactive Tree Of Life v2: online annotation and display of phylogenetic trees made easy. *Nucleic Acids Res*. 39:W475–W478.
Li JB, Gerdes JM, Haycraft CJ, Fan Y, Teslovich TM, May-Simera H, Li H, Blacque OE, Li L, Leitch CC, 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 SE, Gutman R, Liu JS, Mootha VK. 2014. Expansion of biological pathways based on evolutionary inference. *Cell* 158:213–225.
Linard B, Allot A, Schneider R, Morel C, Ripp R, Bigler M, Thompson JD, Poch O, Lecompte O. 2015. OrthoInspector 2.0: software and database updates. *Bioinformatics* 31:447–448.
Mai W, Chen D, Ding T, Kim I, Park S, Cho S, Chu JSF, Liang D, Wang N, Wu D, et al. 2005. Inhibition of Pkhd1 impairs tubulomorphogenesis of cultured IMCD cells. *Mol Biol Cell*. 16:4398–4409.
Mencarelli C, Lupetti P, Dallai R. 2008. New insights into the cell biology of insect axonemes. *Int Rev Cell Mol Biol*. 268:95–145.
Merchant SS, Prochnik SE, Vallon O, Harris EH, Karpowicz SJ, Witman GB, Terry A, Salamov A, Fritz-Laylin LK, Maréchal-Drouard L, et al. 2007. The Chlamydomonas genome reveals the evolution of key animal and plant functions. *Science* 318:245–250.
Mi H, Poudel S, Muruganujan A, Casagrande JT, Thomas PD. 2016. PANTHER version 10: expanded protein families and functions, and analysis tools. *Nucleic Acids Res*. 44:D336–D342.
Moran J, Mckean PG, Ginger ML. 2014. Eukaryotic flagella: variations in form, function, and composition during evolution. *Bioscience* 64:1103–1114.
Morrow EH. 2004. How the sperm lost its tail: the evolution of aflagellate sperm. *Biol Rev Camb Philos Soc*. 79:795–814.
Orphanet: Primary ciliary dyskinesia [Internet]. Orphanet [cited 2017 Jan]. Available from: http://www.orpha.net/consor/cgi-bin/Disease_Search.php?lng=EN&data_id=665&Disease_Disease_Search_diseaseGroup=dyskinesia-syndrome&Disease_Disease_Search_diseaseType=Pat&Disease(s)/group%20of%20diseases=Primary-ciliary-dyskinesia&title=Primary-ciliary-dyskinesia&search=Disease_Search_Simple
Pagliarini DJ, Calvo SE, Chang B, Sheth SA, Vafai SB, Ong S-E, Walford GA, Sugiana C, Boneh A, Chen WK, et al. 2008. A mitochondrial protein compendium elucidates complex I disease biology. *Cell* 134:112–123.
Parisi M, Glass I, 2013. Joubert Syndrome and Related Disorders. In: Pagon RA, Adam MP, Ardinger HH, Wallace SE, Amemiya A, Bean LJ, Bird TD, Fong C-T, Mefford HC, Smith RJ, et al. editors. GeneReviews(®). Seattle (WA): University of Washington, Seattle. Available from: http://www.ncbi.nlm.nih.gov/books/NBK1325/
Pazour GJ, Dickert BL, Vucica Y, Seeley ES, Rosenbaum JL, Witman GB, Cole DG. 2000. Chlamydomonas IFT88 and its mouse homologue, polycystic kidney disease gene tg737, are required for assembly of cilia and flagella. *J Cell Biol*. 151:709–718.
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.
Praveen K, Davis EE, Katsanis N. 2015. Unique among ciliopathies: primary ciliary dyskinesia, a motile cilia disorder. *F1000prime Rep*. 7:36.
Rambhatla L, Chiu C-P, Glickman RD, Rowe-Rendleman C. 2002. *In vitro* differentiation capacity of telomerase immortalized human RPE cells. *Invest Ophthalmol Vis Sci*. 43:1622–1630.
Reiter JF, Blacque OE, Leroux MR. 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.
van Rooijen E, Giles RH, Voest EE, van Rooijen C, Schulte-Merker S, van Eeden FJ. 2008. LRRC50, a conserved ciliary protein implicated in polycystic kidney disease. *J Am Soc Nephrol*. 19:1128–1138.
Ross L, Normark BB. 2015. Evolutionary problems in centrosome and centriole biology. *J Evol Biol*. 28:995–1004.
Sang L, Miller JJ, Corbit KC, Giles RH, Brauer MJ, Otto EA, Baye LM, Wen X, Scales SJ, Kwong M, et al. 2011. Mapping the Nephronophthisis-Joubert-Meckel-Gruber protein network reveals ciliopathy disease genes and pathways. *Cell* 145:513–528.

Sayers EW, Barrett T, Benson DA, Bryant SH, Canese K, Chetvernin V, Church DM, DiCuccio M, Edgar R, Federhen S, et al. 2009. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res*. 37:D5–15.

Seo S, Zhang Q, Bugge K, Breslow DK, Searby CC, Nachury MV, Sheffield VC. 2011. A novel protein LZTFL1 regulates ciliary trafficking of the BBSome and Smoothened. *PLoS Genet*. 7:e1002358.

Shimodaira H. 2004. Approximately unbiased tests of regions using multistep-multiscale bootstrap resampling. *Ann Stat*. 32:2616–2641.

Škunca N, Dessimoz C. 2015. Phylogenetic profiling: how much input data is enough? *PloS One* 10:e0114701.

Suzuki R, Shimodaira H. 2006. Pvclust: an R package for assessing the uncertainty in hierarchical clustering. *Bioinformatics* 22:1540–1542.

Szklarczyk D, Franceschini A, Wyder S, Forslund K, Heller D, Huerta-Cepas J, Simonovic M, Roth A, Santos A, Tsafou KP, et al. 2015. STRING v10: protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Res*. 43:D447–D452.

Tabach Y, Golan T, Hernández-Hernández A, Messer AR, Fukuda T, Kouznetsova A, Liu J-G, Lilienthal I, Levy C, Ruvkun G. 2013. Human disease locus discovery and mapping to molecular pathways through phylogenetic profiling. *Mol Syst Biol*. 9:692.

UniProt Consortium. 2015. UniProt: a hub for protein information. *Nucleic Acids Res*. 43:D204–D212.

Ward JH Jr. 1963. Hierarchical grouping to optimize an objective function. *J Am Stat Assoc*. 58:236–244.

Ward S, Roberts TM, Nelson GA, Argon Y. 1982. The development and motility of *Caenorhabditis elegans* Spermatozoa. *J Nematol*. 14:259–266.

Wheway G, Parry DA, Johnson CA. 2014. The role of primary cilia in the development and disease of the retina. *Organogenesis* 10:69–85.

Woodland HR, Fry AM. 2008. Pix proteins and the evolution of centrioles. *PloS One* 3:e3778.

6.3 Discussion

6.3.1 Le cil, un domaine en constante évolution

L'analyse sur plusieurs niveaux des profils phylogénétiques comme marqueurs évolutifs des gènes ciliaires nous a permis de mieux comprendre l'évolution du cil à travers les Eucaryotes et d'identifier des gènes potentiellement associés au cil. L'étude du cil, à travers plusieurs prismes, est un domaine très actif et la caractérisation des gènes liés aux cils a évolué depuis notre analyse détaillée, que ce soit par le biais de nouvelles données expérimentales ou suite à des efforts de curation des bases de données (Duek et al., 2018 pour un exemple récent). Ainsi, parmi les 87 gènes peu caractérisés que nous avions identifiés comme potentiellement ciliaires, 16 gènes ont depuis été confirmés par d'autres sources (voir Tableau 6-1). Avec les 5 gènes validés expérimentalement lors de notre étude, cela porte le nombre de gènes ciliaires validés parmi nos candidats à 21. Parmi eux, deux gènes ont également été identifiés comme impliqués dans des ciliopathies : ARMC9 lié à la maladie de Joubert et CFAP300 (alors connu sous l'identifiant c11orf70), impliqué dans des PCD.

Tableau 6-1 **Gènes candidats dont le rôle dans le cil a été confirmé.**

Gène	Module évolutif	Source	Date
ARMC9	"Essentiels"	(Van De Weghe et al., 2017)	Juillet 2017
RSPH6A	"Motiles"	(Abbasi et al., 2018)	Octobre 2018
TCTE1	"Motiles"	(Castaneda et al., 2017)	Juillet 2017
C4orf22	"Motiles"	Curation	Août 2018
WDR66	"Motiles"	Curation	Avril 2017
C6orf165	"Motiles"	Curation	Août 2016
WDR92	"Motiles"	(Patel-King et King, 2016)	Avril 2016
IQCA1	"Motiles"	Curation	Avril 2018
LRRC34	"Absence Ecdysozoaires"	(Nevers et al., 2017)	Août 2017
C11orf70	" Absence Ecdysozoaires"	(Fassad et al., 2018; Höben et al., 2018)	Mai 2018
TEX9	" Absence Ecdysozoaires"	(Nevers et al., 2017)	Août 2017
LRRC23	" Absence Ecdysozoaires"	(Nevers et al., 2017)	Août 2017
CCDC108	" Absence Ecdysozoaires"	Curation	Septembre 2016
TTC18	" Absence Ecdysozoaires"	Curation	Mars 2015
WDR27	" Absence Ecdysozoaires"	(Nevers et al., 2017)	Août 2017
CDKL5	" Absence Ecdysozoaires"	(Canning et al., 2018)	Janvier 2018
WDR90	" Absence Ecdysozoaires"	(Hamel et al., 2017)	Août 2017
ARL13A	" Absence Ecdysozoaires"	(Song et Perkins, 2018)	Septembre 2018
MAATS1	" Absence Ecdysozoaires"	Curation	Avril 2017
CCDC42B	" Absence Ecdysozoaires"	Curation	Septembre 2016
BIVM	" Absence Ecdysozoaires"	(Nevers et al., 2017)	Août 2017

Ces découvertes récentes illustrent bien l'importante dynamique de l'étude du cil et amènent une confirmation supplémentaire de la puissance de la génomique comparative pour identifier de nouveaux gènes ciliaires et de potentielles cibles de ciliopathies. Les 66 gènes encore non confirmés restent plus que jamais des candidats potentiels. En ce sens, et bien qu'une implication dans le cil ne soit pas encore démontrée, deux gènes de cette liste ont récemment été associés à des symptômes associés aux ciliopathies : le gène KATNB1 est la cause d'asymétrie droite-gauche et d'insuffisance cardiaque chez la souris (Furtado et al., 2017) et le gène DNAJC7 est associé à l'obésité et au diabète (Cherian et al., 2018). Ces associations phénotypiques, couplées à leur signature évolutive, les rendent donc particulièrement prometteurs en tant que gènes responsables de ciliopathies.

Avec les progrès réalisés dans la recherche sur le cil, la représentation fonctionnelle a également évolué considérablement. En 2017, les consortia *Syscilia* et *Gene Ontology* se sont associés pour améliorer les annotations ayant trait au cil (Roncaglia et al., 2017). Ces travaux ont abouti à la création et la modification de 127 termes GO pour permettre une description affinée de l'organelle ainsi qu'à l'annotation d'un plus grand nombre de gènes ciliaires. On peut en voir l'impact en réitérant les analyses GO que nous avions réalisées au cours de notre étude, ainsi le *cluster* ciliaire démontre un enrichissement encore plus marqué pour le terme *cilium* (enrichi 15,53 fois avec une valeur P de 7.28^{e}-124, contre 14,96 et $1,84^{e}$-98). Le terme *cilium morphogenesis* (enrichi 25,98 fois avec une valeur P de $2,85^{e}$-91) n'existe plus et les termes de l'ontologie *biological process* les plus significativement enrichis sont maintenant les deux termes le remplaçant : *cilium organization* (enrichis 20,55 fois avec une valeur P de $4,19^{e}$-107) et *cilium assembly* (enrichis 21,14 fois avec une valeur P de $3,36^{e}$-106).

Ces annotations améliorées, décrivant plus efficacement les différents compartiments du cil, permettent des analyses d'enrichissement à un niveau plus fin que ce que nous avions pu faire précédemment. J'ai donc analysé l'enrichissement des trois grands modules évolutifs par rapport, non pas à tout le protéome, mais aux 274 gènes ciliaires, afin de détecter une spécialisation fonctionnelle au sein des gènes ciliaires. Ce type d'analyse repose justement sur une bonne définition fonctionnelle des gènes associés, pour permettre d'isoler un enrichissement fonctionnel significatif, malgré la taille réduite de l'ensemble d'analyse. Ainsi, la liste de 91 gènes du module « essentiel » est enrichie significativement pour une quinzaine de termes *biological process* dont *intraciliary transport involved in cilium assembly* (enrichi 11 fois, P valeur de $2,55*10^{e}$-4) et 8 termes *cellular component* incluant entre autres *intracilliary transport particle* (enrichi 8,54 fois, P valeur de $9,25^{e}$-4) et *ciliary transition zone* (enrichi 10,32 fois, P valeur de $1,11^{e}$-03). Un tel enrichissement correspond bien à nos observations qui montraient que ce module comprenait les complexes essentiels à l'assemblage du cil. Le module « motiles » montre quant à lui un enrichissement significatif pour 4 termes *biological process*, le plus spécifique étant *axonemal dynein complex* (3,2 fois, $1,08^{e}$-4), lié aux complexes essentiels de la motilité, ce qui corrobore là encore nos observations fonctionnelles. Pour finir, le troisième module (« absence chez les Ecdysozoaire ») n'est significativement enrichi dans aucun terme, ce qui peut s'expliquer par le nombre encore important de gènes de fonction inconnue dans cette catégorie (32/98).

L'étude du cil et des ciliopathies continuera à progresser dans les années futures, notamment par le biais des analyses omiques. Dans ce contexte, il sera également important de faire évoluer les prédictions basées sur l'évolution avec de nouvelles données de génomique comparative.

6.3.2 L'étude du cil dans OrthoInspector 3.0

Les 100 espèces que nous avions choisies pour l'analyse du cil ont été sélectionnées en fonction de leur diversité et de la représentation des clades non ciliés. Cette problématique de représentation de la diversité des espèces disponibles, en évitant les biais vers un clade donné, a bien sûr inspiré le processus de sélection d'organismes modèles d'OrthoInspector 3.0. Les nouvelles espèces inclues dans OrthoInspector 3.0 ouvrent également de nouvelles perspectives pour l'étude du cil. Parmi celles-ci, on compte *Perkinsella sp,* un représentant non-cilié des Excavés, précédemment représentés uniquement par des espèces ciliées, des champignons ciliés (*Spizellomyces punctatus*, *Gonapodya prolifera*, *Allomyces macrogynus, Rozella allomycis*) répartis dans plusieurs clades en addition de *Batrachochytrium dendrobatidis,* déjà présent dans notre premier ensemble d'analyse, un représentant cilié du clade des Cryptophytes (*Guillardia theta*) ou encore deux représentants ciliés des Haptophycae (*Chrysochromulina sp* et *Emiliana huxleyi*). Pour l'analyse de nos différents modules ciliaires, les espèces d'OrthoInspector 3.0 comptent également un nombre plus important d'Ecdysozoaires, avec notamment un représentant des Myriapodes.

L'une des méthodes utilisées lors de nos travaux sur le cil reposait sur l'obtention d'un score en fonction de la présence et de l'absence chez les espèces ciliées et non ciliées de chaque clade. Ce score basé sur les clades est plus informatif qu'un simple score basé sur le nombre d'espèces ciliées et non ciliées, qui ne prend pas en compte la diversité des espèces. Cette expérience nous a guidés dans le développement de la recherche par profil dans OrthoInspector 3.0 et le choix de placer des contraintes au niveau du clade. Pour autant, la recherche par profil est basée sur des critères stricts, en contraste avec la recherche basée sur des scores qui permet d'accepter que des protéines soient présentes chez certaines espèces pour lequel le phénotype est absent ou soient présentes dans seulement une partie des clades où il est observé. Ces deux approches diffèrent donc principalement par des critères de spécificité et de sensibilité. Le choix du système de contraintes pour OrthoInspector s'est fait naturellement car il permet de se concentrer spécifiquement sur un profil donné et qu'il ne nécessite pas la définition d'un score, qui peut dépendre du sujet d'étude. Comme notre expérience sur le cil le montre, il est toutefois souvent judicieux de faire preuve de souplesse pour l'étude des systèmes biologiques et la mise en place d'un système moins strict est un développement à envisager à l'avenir pour notre outil de recherche.

Pour conclure, outre l'identification de nouveaux gènes ciliaires et la caractérisation de modules évolutifs, le point qui ressort de notre étude du cil est la valeur ajoutée de notre approche consensus combinant les prédictions de trois méthodes indépendantes et l'intérêt de l'analyse des profils à différents niveaux taxonomiques, de l'ensemble des Eucaryotes à certains clades des Métazoaires. Les méthodes choisies sont fortement dépendantes de l'objet biologique, on ne retrouve pas, par exemple, d'autre système biologique eucaryote autre que le cil avec une

signature si visible par *clustering* hiérarchique. Si les outils d'OrthoInspector 3.0 sont, par dessein, des généralisations des méthodes utilisées lors du projet ciliaire, ils reposent sur les mêmes principes : la combinaison d'analyses guidées par la connaissance, l'utilisation de mesures de distance automatiques pour ordonner les profils et l'analyse des profils ainsi identifiés pour en tirer des associations plus fines.

7 Matériel et méthodes

7.1 Ressources bioinformatiques

Les bases de données et le portail d'OrthoInspector 3.0 utilisent des données issues de plusieurs ressources publiques. Cela comprend les protéomes qui ont servi pour l'inférence des relations, ainsi que toutes les données additionnelles utilisées dans le cadre de la contextualisation. Je décris ici rapidement chaque ressource, la façon dont nous accédons aux données et, le cas échéant, les traitements appliqués.

7.1.1 Protéomes UniProt

UniProt (*Universal Protein Resource*) est une ressource dédiée aux protéines, à leurs séquences, ainsi qu'à leurs annotations, issues de diverses sources. UniProtKB (*UniProt Knowledgebase*) est composée des bases de données SwissProt dont les séquences font l'objet de curation et d'annotations manuelles et TrEMBL (*Translated EMBL Nucleotide Sequence Data Library*) dont les séquences sont annotées automatiquement. Les protéomes UniProt regroupent, pour chaque organisme, l'ensemble de ses protéines. Ils comprennent généralement les séquences issues de la traduction du génome (incluant le cas échéant le génome des plasmides et des organelles). Les protéomes de référence sont un sous-ensemble de ces protéomes (non-redondants) sélectionnés de façon à couvrir l'ensemble du vivant, et les espèces d'intérêt pour la recherche biomédicale ou biotechnologique. Ce sont ces protéomes de référence qui ont été utilisés comme base pour OrthoInspector.

Nous avons téléchargé les protéomes de référence pour les Eucaryotes, les Archées et les Bactéries, au format FASTA via le serveur ftp d'UniProt. Tous les téléchargements ont eu lieu au cours du mois de novembre 2016 (version 2016_10 d'UniProt). Ce sont ces protéomes qui ont servi aux inférences d'orthologie des quatre bases d'OrthoInspector 3.0.

Comme spécifié dans la partie Contributions, ces protéomes ont été sujets à une procédure de contrôle qualité en deux étapes : la visualisation des distributions de longueurs des séquences protéiques et la détermination d'indicateurs de qualité. Ces procédures ont été réalisées à l'aide de deux scripts Python dédiés. Le premier permet la génération de graphiques représentant la distribution des longueurs des séquences protéiques de tous les protéomes d'un répertoire donné (voir section 4.2.2.1 pour ces graphiques). Le second extrait plusieurs indicateurs caractérisant ces protéomes et les retranscrit dans un format CSV. Je décris brièvement ces indicateurs ci-après :

- *Taille du protéome* : le nombre de protéines composant un protéome donné. Ceci permet d'identifier notamment les protéomes incomplets. Cette indication est complétée par le nombre de protéines issues de SwissProt et de TrEMBL, information extraite de l'en-tête des séquences FASTA de chaque protéine (respectivement *sp* ou *tr* en début d'en-

tête). Une forte proportion de protéines issues de SwissProt identifie les protéomes ayant fait l'objet de curation et donc supposés plus fiables.

- *Annotations de fragments* : la proportion de protéines annotées comme *fragment* dans UniProt. On extrait cette information en recherchant la présence du mot 'fragment(s)', avec ou sans majuscules, dans la courte description de l'en-tête des séquences au format FASTA.
- *Petites protéines* : la proportion de protéines faisant strictement moins de 100 acides aminés. Cette information est calculée sur la base d'un décompte de caractères dans les séquences protéiques. Dans le fichier, cet indicateur est complété par la taille des protéines au premier quartile de la distribution.
- *Protéines ne commençant pas par une méthionine* : cet indicateur se base sur la vérification du premier acide aminé de chaque séquence.

Ces informations nous ont servi à sélectionner les protéomes de bonne qualité. Pour les Archées et les Bactéries, cela correspond aux protéomes ayant moins de 10% de protéines annotées comme fragments, moins de 20% de petites protéines et moins de 10% de protéines ne commençant pas par une méthionine. Pour les Eucaryotes, nous avons conservé les protéomes ayant moins de 20% de petites protéines et moins de 55% de protéines ne commençant par une méthionine. Nous avons également exclu les protéomes présentant une taille réduite, selon des critères détaillés dans la partie Contribution. Les protéomes ainsi sélectionnés ont été utilisés pour créer les trois bases de données spécifiques à chaque Domaine du Vivant.

La base de données inter-domaines repose sur un sous-ensemble du jeu de protéomes total dont le protocole de sélection repose sur la diversité taxonomique. Les règles utilisées pour cette sélection sont décrites en détail dans la section 4.3.1 des Contributions.

7.1.2 Taxonomy

La base de données Taxonomy (Federhen, 2012) du NCBI est dédiée à la classification de l'ensemble des espèces, d'une façon standardisée, sous la forme d'un arbre. Elle permet notamment d'associer toute séquence nucléique ou protéique à un organisme jusqu'au niveau de la souche, sans ambiguïté et de retrouver la lignée complète souche, espèce, genre etc. Pour cela, un identifiant unique, le *taxonomy identifier* ou *taxid*, est attribué à chaque nœud de l'arbre. UniProt maintient une version spécifique de cette base de données, compatible avec les séquences protéiques d'UniprotKB.

OrthoInspector utilise les données de taxonomie pour plusieurs de ses options de génomique comparative. Il exploite pour cela une installation PostgreSQL locale de la base Taxonomy construite à partir des données taxonomiques téléchargées d'UniProt, et tenue à jour au même rythme. Notre installation locale n'adopte pas le schéma classique d'une base Taxonomy mais une implémentation en *nested set* qui permet des requêtes optimisées pour retrouver les lignages complets des espèces.

Afin de faciliter certaines requêtes, la lignée taxonomique de chaque espèce est directement intégrée dans la table *bank* des bases de données OrthoInspector. Cette table est synchronisée tous les mois avec l'instance locale de la base Taxonomy, pour les quatre bases de données d'OrthoInspector 3.0.

7.1.3 Gene Ontology

Gene Ontology permet de décrire, dans un vocabulaire standardisé, la fonction des gènes selon trois aspects principaux : sa fonction moléculaire, le processus biologique dans lequel il intervient et le composant cellulaire où il réalise sa fonction. Les termes GO peuvent être plus ou moins spécifiques et sont représentés par des liens les uns aux autres qui permet notamment de rendre compte d'une structure hiérarchique : les termes généraux comme *organelle* sont les parents/ancêtres de termes plus spécifiques comme *cilium* ou *mitochondrion.* Les termes GO peuvent être issus de plusieurs sources, dont la fiabilité est exprimée par des *evidence codes*. A titre d'exemple, les fonctions validées expérimentalement prennent le code EXP quand celles tirées d'inférences automatisées prennent le code IEA.

Les termes Gene Ontology sont utilisés dans OrthoInspector pour contextualiser la fonction des protéines, à la fois au niveau de la page descriptive de chaque protéine et des outils de génomique comparative. Toutes les données de GO utilisées dans OrthoInspector sont extraites d'une instance locale de la base de données officielle GO construite à partir du dépôt MySQL *assocdb* disponible à http://archive.geneontology.org/latest-full/. L'ensemble des termes concernant une protéine sont extraits de la base, sans filtre sur la source de l'annotation, par le biais de leur numéro d'accès UniProt.

7.1.4 InterPro

InterPro est une ressource d'analyse fonctionnelle des séquences protéiques, et notamment d'inférence de leur architecture en domaines. Elle combine les inférences issues de plusieurs bases de données pour permettre une caractérisation complète des séquences.

La page de protéine d'OrthoInspector exploite les informations de domaines issues d'InterPro en tant qu'outil de contextualisation fonctionnelle. Ces données sont récupérées dynamiquement, à la volée, par le biais d'un *webservice* de l'EBI (https://www.ebi.ac.uk/interpro/protein/{access}?export=tsv), en utilisant le numéro d'accès UniProt comme clé.

7.1.5 Enrichissement GO par Panther

Panther est une plateforme dédiée à la classification évolutive et fonctionnelle des gènes (décrite plus en détail à la section 3.3.3.2) visant à faciliter les analyses à haut-débit. La plateforme propose, entre autres, des tests d'enrichissement fonctionnel, supportant une totalité de 112 espèces, des trois Domaines de Vivant, sélectionnées pour leur intérêt pour la recherche.

OrthoInspector exploite les *webservices* de Panther (la documentation est disponible dans la section Help du site http://pantherdb.org) dédiés à ces tests de surreprésentation pour caractériser les listes de protéines obtenues par l'outil de recherche par profil phylogénétique. Dans un premier temps, la disponibilité de l'espèce dans Panther est vérifiée dynamiquement à chaque recherche par le bais d'un premier *webservice* (*search for supported organism*). En cas d'indisponibilité, le bouton dédié à l'enrichissement GO est désactivé. Dans le cas où il est accessible, l'enrichissement à la volée est effectué en utilisant le *webservice* d'enrichissement (*Over-representation test*), sous la forme de trois requêtes POST, une par catégorie de termes GO. Ces requêtes utilisent comme paramètres :

- *organism* -le nom de l'espèce requête.
- *geneList* - la liste des numéros d'accès UniProt issus de la recherche.
- *enrichmentType*- Successivement *fullGO_function*, *fullGO_process*, *fullGO_component*

Les résultats obtenus ainsi sont ensuite mis en forme dans un tableau, ordonné par valeur P croissante, avec la possibilité d'accéder à toutes les protéines de la liste associées aux termes GO enrichis.

7.2 Protocole d'installation

Le protocole d'installation d'OrthoInspector 3.0, conçu pour gérer de grande quantité de données est décrit, dans le principe, dans la partie contribution. J'en donne ici les détails techniques.

7.2.1 European Grid Infrastructure

L'*European Grid Infrastructure* (EGI) est une fédération de centres de calcul dédiée à la recherche. Elle permet l'accès à des ressources de calculs dans le monde entier, totalisant plus de 850 000 cœurs. Nous avons utilisé cette infrastructure pour distribuer nos comparaisons de séquences tous-contre-tous. Les différentes tâches ont été distribuées sur jusqu'à 4 000 cœurs simultanément, grâce à un orchestrateur interne, *Grilladin* développé par Arnaud Kress. *Grilladin* permet de distribuer les calculs sur des ressources distantes et hétérogènes. Il active un mécanisme de détection d'erreur ou de non-exécution de tâches et permet la redistribution de celles-ci sur d'autres serveurs permettant ainsi l'exécution de la totalité des calculs.

7.2.2 Fragmentation des tâches de BLAST

Les comparaisons tous-contre-tous ont été réalisées en utilisant la suite programmatique en ligne de commande BLAST+. Pour chacune des quatre bases de données, une banque BLAST a été générée en utilisant comme données d'entrée les protéomes correspondants, concaténés en un fichier. La commande utilisée est la suivante :

```
makeblastdb -in {input} -dbtype prot
```

Pour les protéomes volumineux, cette commande créée automatiquement plusieurs banques, d'une taill maximale de 1 Go chacune. Les tâches de BLAST proprement dites ont été

effectuées parallèlement sur les différents nœuds de l'EGI, par fichier FASTA de 500 séquences et sous-section de la base de données en utilisant la commande blastp, avec comme options :

- dbsize : la taille totale des séquences dans les protéomes concaténés utilisés pour construire les banques. Cette option permet de rendre comparable les résultats issus des différentes sous-sections de banque BLAST.
- *word_size : 3*
- *evalue :* 1.0 e^{-5}
- *max_target_seq :* 5000
- *outfmt: 7 qseqid sseqid pident length mismatch gapopen qstart qend sstart send evalue bitscore staxids*. Ce paramètre nous permet d'obtenir les résultats du BLAST dans un format compatible avec OrthoInspector.

Les fichiers de résultats de chaque tâche BLAST ont ensuite été fusionnés automatiquement en un fichier unique par le biais d'un programme TCL afin de réunir tous les *hits* ayant trait à une même protéine requête pour les différentes sous-sections de la base de données. Les résultats dans ce fichier sont ordonnés en fonction du *bitscore* obtenu.

7.2.3 Création des bases OrthoInspector

L'installation des quatre bases de données PostgreSQL d'OrthoInspector 3.0 a été effectuée, étape par étape, en utilisant la version du programme accessible publiquement sur http://www.lbgi.fr/orthoinspectorv3/download_Package. Toutes les étapes ont été réalisées par lignes de commandes suivant la procédure détaillée sur http://www.lbgi.fr/orthoinspectorv3/tutorials, et l'option *dump only*, permettant le contrôle manuel de l'insertion des données ainsi calculées dans la base de données.

Pour les bases de données les plus volumineuses, certaines étapes ont nécessité l'exécution parallèle du programme sur plusieurs cœurs. Le Tableau 7-1 résume les cas où cela a été nécessaire, ainsi que la taille des sous-ensembles de données utilisés. Les fichiers issus des étapes parallélisées ont ensuite été fusionnés automatiquement. A chaque étape, les données récupérées ont été insérées manuellement dans la base de données puis les tables correspondantes ont été indexées.

Tableau 7-1 **Parallélisation des étapes d'OrthoInspector**. « Tout » signifie que l'étape a été réalisée en une fois, pour les bases de données les moins volumineuses.

	BLAST parsing	***Inparalog validations***	***Orthology calculations***
Archées	Tout	Tout	Tout
Bactéries	Paquets de 50 000 séquences	Paquets de 100 espèces	Paquets de 20 espèces
Eucaryotes	Paquets de 100 espèces	Paquets de 10 espèces	Paquets de 10 espèces
Inter-domaines	Tout	Paquets de 50 espèces	Paquets de 20 espèces

Afin de faciliter la recherche dans les bases de données, nous avons ajouté automatiquement, pour les quatre bases de données, une table supplémentaire au schéma classique d'OrthoInspector, dédiée aux identifiants des bases de données externes. Cette table comprend les numéros d'accès et identifiants UniProt et facilite les requêtes exécutées sur le portail.

7.3 Implémentation du site web

Le portail OrthoInspector 3.0, que ce soit pour l'accès aux relations d'orthologie ou pour les outils de génomique comparative, est décrit en détail dans la partie Contributions. Ici, je reviens sur les aspects techniques de l'implémentation.

7.3.1 Technologies

Le site web OrthoInspector 3.0 est développé, au niveau serveur, en PHP et repose sur le *framework* Silex. L'architecture du portail suit le modèle de conception MCV (Modèle-Vue-Contrôleur) permettant une organisation structurée du code et sa modularité. La Figure *7-1* schématise cette organisation.

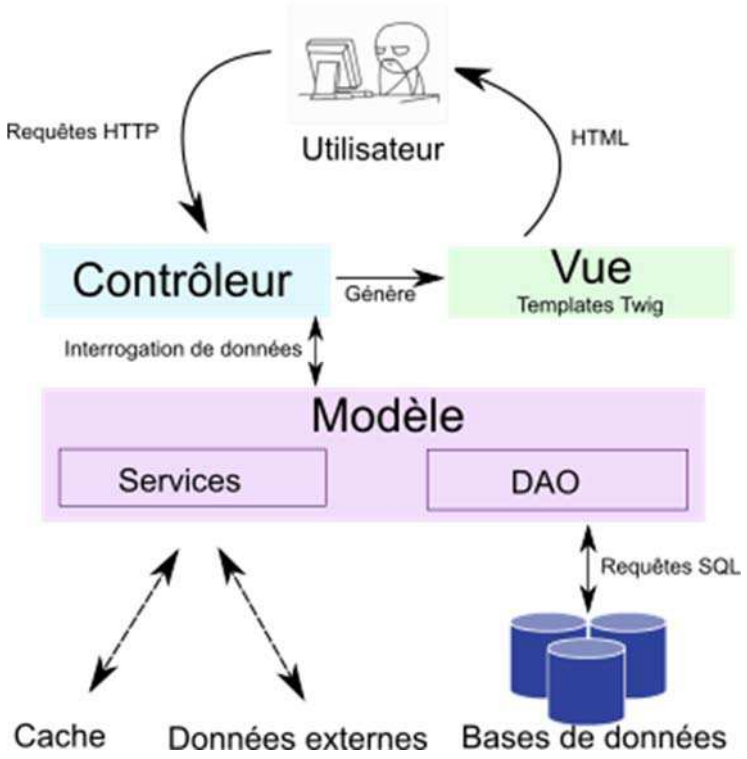


Figure 7-1 **Organisation technique du portail OrthoInspector**. Le code est séparé en trois modules. Le Contrôleur gère les requêtes de l'utilisateur, interroge les données *via* les sections de la partie Modèle, et utilise les données traitées à la partie Vue pour générer le code HTML visualisé par l'utilisateur.

La partie Contrôleur permet de gérer les requêtes de l'utilisateur, envoyées par le biais du protocole http et interagit avec les autres parties de l'application. La partie Modèle de l'application gère les interactions avec les données de l'application, elle se divise en une sous-

section Services, qui orchestre notamment le cache et l'interrogation de données externes (InterPro, Panther) et une sous-section DAO (*Data Access Object*) qui permet l'interaction avec toutes les bases de données SQL locales, par l'intermédiaire du service Doctrine. Finalement, la partie Vue de l'application gère le rendu des données à l'utilisateur. Cette dernière partie repose sur le moteur de *template* Twig, qui orchestre la génération des pages HTML.

Du coté client, le rendu des pages d'OrthoInspector repose principalement sur la très populaire librairie CSS Bootstrap 3.0 complétée par des feuilles de style personnalisées. Les éléments dynamiques des pages sont gérés par du code JavaScript, utilisant principalement la librairie JQuery. Le choix des technologies citées ici pour développer OrthoInspector 3.0 reflète principalement une volonté d'en garantir la pérennité et d'en faciliter la maintenance. Nous avons donc préféré utiliser uniquement des bibliothèques logicielles simples et stables.

Les développements du portail web d'OrthoInspector 3.0 sont gérés sur le serveur GitLab du laboratoire. Toutes les modifications sont suivies par le logiciel de *versioning* git et sont en premier lieu appliquées sur une version de travail du portail, distincte de celle accessible publiquement. Seules les modifications testées et validées sont intégrées à la version de production du site web.

7.3.2 Table d'orthologie et taxonomie

La table d'orthologie de la page de protéine constitue le principal accès aux données d'orthologie. Les données concernant les relations d'orthologie affichées proviennent de requêtes effectuées dans les différentes tables de la base de données OrthoInspector. Les relations d'orthologie sont réparties dans trois tables (*onetoone*, *onetomany*, *manytomany*) de la base, et sont représentées de façon dirigée (la relation part d'une protéine A vers une protéine B). Une combinaison de six requêtes interrogeant chaque table dans les deux sens, permet d'explorer toutes les relations d'orthologie d'une protéine donnée et de récupérer les identifiants, descriptions, espèces associées et séquences.

A ces données s'ajoutent des données taxonomiques, qui ne sont pas affichées en détail dans la table pour des raisons pratiques mais qui sont essentielles à son exploration. Ces données sont extraites de la base de données Taxonomy locale décrite précédemment. Ainsi, le lignage complet de chaque espèce (l'ensemble de ces clades ancestraux) est intégré dans une colonne « cachée » de la table d'orthologie. C'est cette colonne qui permet de filtrer les résultats affichés dans la table en fonction d'un clade d'intérêt.

L'ordre des résultats affichés dans la table d'orthologie dépend également de la taxonomie, et plus précisément de la distance taxonomique entre l'espèce requête et les espèces dans lesquelles sont détectés les orthologues. Cette distance taxonomique est mesurée à partir du lignage complet et correspond au nombre de clades qu'il faut remonter dans la taxonomie, à partir de l'espèce requête, pour retrouver l'ancêtre commun avec chaque espèce cible. Par exemple, la distance entre l'homme et le chimpanzé, tous deux des Homininae, serait de 2 : la position de ce clade dans le lignage de l'homme, selon le NCBI (***Homininae*** *; Homo ; Homo*

sapiens). Le rang taxonomique est également représenté dans une colonne « cachée » de la table d'orthologie, qui est utilisée par défaut pour trier les résultats, ce qui permet d'afficher les résultats de façon pertinente.

7.3.3 Recherche par profils

L'implémentation de la recherche par profil présente deux spécificités techniques principales : la sélection de contraintes sur l'arbre phylogénétique et la recherche de protéines proprement dite.

L'affichage de l'arbre taxonomique sur lequel sélectionner les contraintes utilise la bibliothèque libre de représentation d'arborescence JavaScript FancyTree, modifiée de façon à permettre trois états de sélection (sans contrainte, présence, absence) et à ce que les contraintes de présence et d'absence se répercutent dans le reste de l'arbre. Cette librairie exploite, pour la construction de l'arbre, des données le décrivant au format JSON. Les descriptions des arbres relatifs à chaque base de données sont générées automatiquement à partir de la liste des espèces présentes dans la base, en passant par la base de données Taxonomy locale. La construction automatique de ces arbres pouvant être longue, ces données sont conservées en cache au niveau du serveur pendant un mois, et sont recalculées à l'issue de ce délai pour tenir compte des mises à jour éventuelles de la taxonomie.

La recherche de protéines obéissant à un profil repose directement sur l'interrogation des bases de données d'orthologie par une requête SQL composite, pouvant prendre en compte plusieurs clades présents et absents. Ces requêtes sont composées d'un enchainement de blocs identiques dont la fonction est de retrouver les protéines de l'espèce requête ayant au moins un orthologue avec des protéines d'au moins une espèce dans une liste d'espèces cibles. Ces blocs de requêtes SQL sont enchainés les uns après les autres pour correspondre aux contraintes demandées, en utilisant les opérateurs SQL INTERSECT et EXCEPT, comme l'illustre la Figure 7-2. Cette structure modulaire permet de s'adapter à toutes les combinaisons de contraintes qu'il est possible de définir.

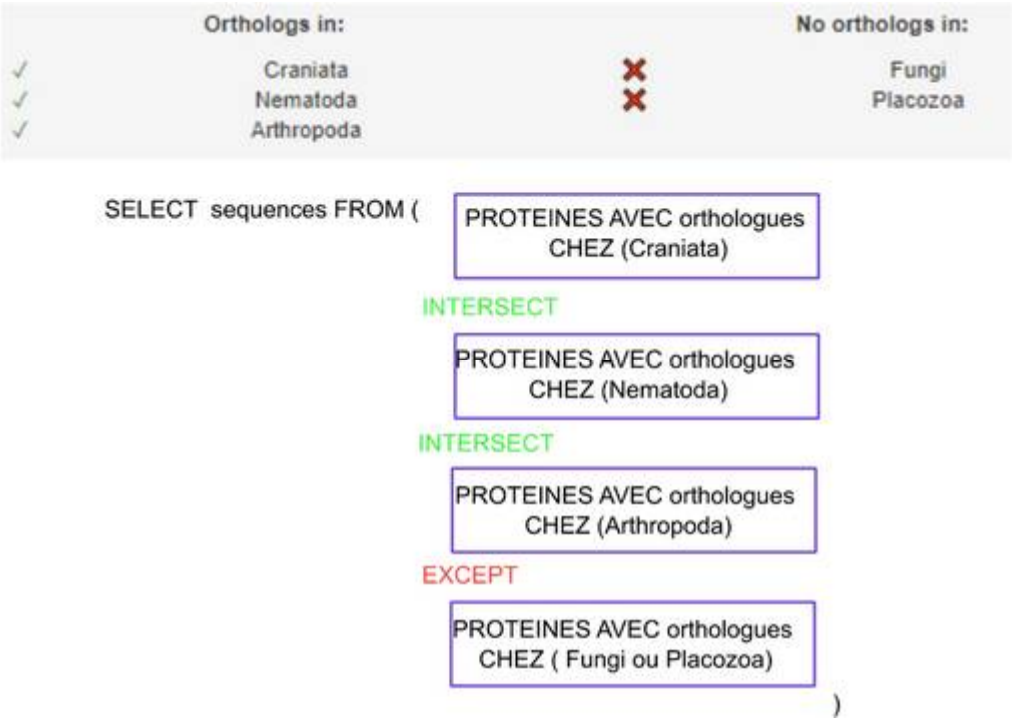


Figure 7-2 **Structure modulaire des recherches par profil**. Le bloc de base, en violet, permet de retrouver toutes les protéines ayant un orthologue dans au moins une espèce d'une liste donnée. On enchaîne autant de ces blocs que de contraintes de présence avec le mot clé INTERSECT, et on y rajoute un bloc commun pour toutes les contraintes d'absence avec le mot clé EXCEPT.

7.3.4 Distribution taxonomique

La visualisation synthétique des distributions taxonomiques est disponible à la fois sur la page de protéine et en tant que résultat des outils de génomique comparative. Cet outil de visualisation affiche la distribution des orthologues des protéines dans un certain nombre de clades, sélectionnés à l'avance. Techniquement, les clades choisis à chacun des trois niveaux (voir Contributions) sont définis dans des fichiers de paramètres dédiés, au format JSON, qui décrivent le nom de chaque clade et le ou les *taxid* des clades correspondants dans la taxonomie NCBI. On peut ainsi représenter des divisions taxonomiques reconnues dans la littérature mais non intégrées à la taxonomie du NCBI (par exemple, Excavate correspond à quatre *taxids* : 207245, 5719, 33682, 5752) ou regrouper plusieurs clades dans une catégorie *Other* si nécessaire. A partir de ces fichiers, la distribution est générée automatiquement, en utilisant les données taxonomiques des espèces dans lesquelles on retrouve un orthologue. Les divisions étant définies en termes taxonomiques, et non en fonction des espèces, cette implémentation est suffisamment flexible pour s'adapter aux évolutions des espèces répertoriées dans les bases de données OrthoInspector ou aux changements de classification.

Sur la page des protéines, les distributions sont construites directement à partir des données chargées de la table d'orthologie, ce qui permet d'éviter des requêtes supplémentaires dans les bases de données. Dans le cadre des outils de génomique comparative, il s'agit d'afficher la distribution de nombreuses protéines en peu de temps. L'affichage de ces distributions repose donc sur une base de données PostgreSQL stockant uniquement l'indicateur de présence dans

une espèce pour chaque protéine ainsi que les grandes catégories taxonomiques auxquelles les espèces appartiennent. Le bilan taxonomique de chaque protéine peut ainsi être obtenu par une requête SQL simple.

7.4 Profils phylogénétiques et similarité de profils

Les mesures de similarité entre profils permettent, à la fois dans MyGeneFriends et OrthoInspector de faire le lien entre des protéines ayant une histoire évolutive similaire. Dans les deux cas, leur calcul a nécessité la génération des profils phylogénétiques à partir des relations d'orthologie, ainsi que le calcul des distances proprement dites.

7.4.1 Génération des profils

Les profils phylogénétiques ont été générés à partir de bases de données d'orthologie par le biais du programme Phyligrane. Ce programme, développé en Python, permet de générer à partir d'une base de données OrthoInspector un fichier tabulé correspondant à une matrice phylogénétique, avec les protéines de l'espèce requête en lignes et les espèces cibles en colonnes. Brièvement, le programme prend comme argument :

- *orga_id* : l'identifiant OrthoInspector de l'espèce, ou son *taxid*.
- *output* : le nom du fichier de sortie.
- *orga_list* : les identifiants OrthoInspector des espèces cibles ou leur *taxid*. Par défaut, toutes les espèces de la base sont utilisées.
- *database* : un fichier décrivant la base de données OrthoInspector cible.

La matrice phylogénétique est construite protéine par protéine. Pour chacune d'entre elles, les orthologues sont récupérés dans la base de données par des requêtes SQL (dans les tables *onetoone*, *onetomany*, *manytomany*) et pour chacune des espèces cibles, la colonne correspondante est remplie par 1 si un orthologue existe chez cette espèce et 0 dans le cas contraire.

Le fichier généré peut comprendre des informations supplémentaires, telles que les espèces, la taxonomie, les descriptions fonctionnelles, pour faciliter l'inspection visuelle de la matrice phylogénétique ou décrire simplement les lignes et colonnes par leurs identifiants de protéines et *taxid* (option *minimal*), afin d'en faciliter la prise en charge par des programmes d'analyse automatiques.

Les profils utilisés pour définir les distances entre gènes dans MyGeneFriends, ont été construits en utilisant la base de données eucaryotes d'OrthoInspector v2 avec l'homme comme espèce requête et 100 eucaryotes en espèces cibles (les 100 espèces sont les même que celles utilisées dans l'article de caractérisation évolutive des gènes ciliaires (Nevers et al., 2017).).

Les profils utilisés pour définir les distances entres gènes dans OrthoInspector ont quant à eux été construits en se basant sur les bases de données intra-domaines et inter-domaines

d'OrthoInspector 3.0, en utilisant chacune des espèces des bases correspondantes comme espèce requête. Les espèces cibles utilisées par les bases de données inter-domaines, Eucaryote et Bactéries sont les espèces « modèles », définies dans la partie Contributions. Les espèces cibles utilisées pour les profils de la base de données Archées sont toutes les espèces de cette base de données.

7.4.2 Calculs de distances

Les mesures de distances ont été calculées en R, à partir des matrices extraites des fichiers CSV de distances. Les distances de Jaccard ont été calculées en utilisant la fonction dist.binary du package ade4 (Dray et Dufour, 2007), avec comme paramètre la matrice phylogénétique et l'option « meth=1 ».

En considérants deux profils phylogénétiques x et y, avec a le nombre d'espèces dans lesquelles une présence est observée pour x et y, b le nombre d'espèces dans lesquelles la présence est observée pour x seulement et c pour y seulement. La distance de Jaccard entre ces deux profils correspond à :

$$d_{Jaccard} = \sqrt{1 - \frac{a}{a+b+c}}$$

Les distances de corrélation ont été calculées en utilisant la fonction *dist* du package *amap* (Lucas, 2018), avec comme paramètre la matrice phylogénétique et l'option « method= 'correlation' »). Aucune corrélation ne pouvant être calculée pour les enregistrements sans variance, les lignes vides (protéines sans orthologue chez les espèces cibles) ont été supprimées au préalable de la matrice.

7.4.3 Stockage et accès aux distances

Les matrices de distances calculées de cette manière pour chaque profil phylogénétique ont été enregistrées sous formes de fichiers CSV. Les paires de protéines pour lesquelles la distance de Jaccard est inférieure à 0,5, ainsi que la valeur de distance exacte correspondante ont été extraites à l'aide d'un script Python et enregistrées dans une base de données PostgreSQL.

Les distances sont affichées sur le portail OrthoInspector *via* des requêtes SQL à cette base de données indexée. Ces requêtes récupèrent toutes les protéines similaires à une protéine cible, sous un second seuil de distances, fixé à 0,4. Ce second seuil étant un paramètre des requêtes, il peut techniquement être fixé de 0 à 0,5.

8 Conclusion et perspectives

L'explosion des masses de données de la dernière décennie a redéfini l'étude des liens génotype-phénotypes et favorisé l'avènement des méthodes intégratives. L'un des défis majeurs d'aujourd'hui est de savoir exploiter et combiner ces omiques pour décrire les différents aspects des systèmes biologiques. Mes travaux de thèse se sont inscrits dans ce contexte en s'intéressant aux données génomiques de l'ensemble du Vivant dans le cadre de la génomique comparative. On y retrouve deux grandes problématiques qui, à mon sens, sont emblématiques de l'exploitation des omiques en biologie. Comment peut-on extraire et synthétiser la connaissance de données nombreuses et hétérogènes ? Comment exploiter cette connaissance pour décrire les systèmes biologiques ?

Des marqueurs évolutifs pour synthétiser les données génomiques

Ma réponse à la première problématique a été la conception de marqueurs évolutifs sous la forme de profils phylogénétiques. Cette approche s'appuie sur un important effort de réduction des dimensions des données pour offrir une représentation résumée de l'histoire évolutive des gènes et du contenu en gènes des différents génomes. Définir des marqueurs évolutifs fiables et pertinents nécessite un travail de sélection des données, de développement de méthodes d'inférence et leur mise en œuvre à grande échelle pour exploiter ces données. Ce qui me paraît le plus important, avec le recul sur l'ensemble de mes travaux, n'est pas seulement d'extraire les informations des données, mais surtout de les représenter de façon pertinente.

Les marqueurs évolutifs intégrés à MyGeneFriends et OrthoInspector reposent sur un même concept, le profil phylogénétique et pourtant, ils le représentent de façon essentiellement différente. MyGeneFriends propose seulement de les catégoriser par grandes classes de profils, notamment en fonction de l'âge des gènes, alors que la distribution taxonomique d'OrthoInspector permet d'affiner l'information, en résumant les présences-absences clade par clade, plus ou moins finement selon les niveaux choisis par l'utilisateur. Ces différents choix de représentation obéissent à une logique de contextualisation : il n'est pas nécessaire de montrer toutes les informations, mais de montrer uniquement les informations pertinentes à l'analyse du sujet. C'est ce qu'exprime cette citation de Clay Shirky, à propos d'internet : « *It's not information overload. It's filter failure* ». En d'autres termes, la masse de données n'est pas un problème tant qu'on arrive à isoler les informations qui sont réellement utiles dans un contexte donné.

C'est la contextualisation qui décide également des informations mises à disposition sur les différents portails web. OrthoInspector intègre principalement des données fonctionnelles pour complémenter la génomique comparative, dans une logique de dialogue évolution-fonction permanent. MyGeneFriends intègre quant à lui des données de différents types, informant sur un gène à plusieurs niveaux. Un gène, ou ses produits, peuvent aujourd'hui être décrits selon bien des angles, par des données, *in silico*, expérimentales ou bibliographiques. Les choix de

contextualisation pertinents pour un sujet de recherche ne le seront pas nécessairement pour d'autres. Il n'existe donc pas de choix objectif absolu concernant les données à mettre en avant. MyGeneFriends répond à cette problématique à la façon des réseaux sociaux grâce à la mise en œuvre d'outils de personnalisation dans une approche participative et collaborative : c'est l'utilisateur, par ses actions, qui aide à mettre des éléments en valeur pour son propre sujet d'intérêt. Une autre façon d'y répondre est de passer par l'interopérabilité des bases de données et le web sémantique. Il s'agira pour chaque personne de choisir les données qui l'intéressent et de les croiser selon ses besoins, par une requête SPARQL, par exemple. Les développements futurs d'OrthoInspector le conduiront à emprunter également la voie du web sémantique pour renforcer sa logique de contextualisation.

La représentation des marqueurs évolutifs obéissant également à une logique de contexte, il est légitime de s'interroger sur leur devenir. Nous avons opté, dans OrthoInspector pour trois niveaux de granularité dans la représentation, afin de permettre des choix adaptés à diverses questions biologiques. Cette représentation par niveaux taxonomiques convient dans un large panel de situations et l'on pourrait poursuivre dans cette logique, en ajoutant des niveaux supplémentaires à ceux déjà existants (par exemple, la distribution au sein des Ascomycètes). Pour autant, on peut imaginer des cas où d'autres modalités de représentation seraient plus pertinentes, notamment en intégrant des catégories basées sur des clades ou des traits phénotypiques définis par l'utilisateur. Dans notre étude des ciliopathies, nous avons divisé les espèces non seulement par grands clades, mais aussi en espèces ciliées et non ciliées. Cette représentation permettant d'accéder directement aux corrélations génotype-phénotype est en parfaite adéquation avec le contexte de l'étude. La prise en compte implicite dans la représentation des besoins spécifiques à chaque question biologique passe là encore par l'utilisation des codes du Web 2.0 et par une approche participative.

Les marqueurs évolutifs pour la description de systèmes biologiques

La seconde problématique des omiques est la modalité d'utilisation des connaissances extraites des masses de données pour décrire les systèmes biologiques. Dans notre cas, il s'agissait d'exploiter les marqueurs évolutifs pour les analyses à grande échelle. A travers les outils développés au cours de cette thèse, j'ai apporté des éléments de réponse : on peut identifier les fonctions associées à des histoires évolutives, ou inversement, et surtout, on peut utiliser les similarités de profils entre gènes afin de retrouver les modules évolutifs et leur correspondance avec des modules fonctionnels. Ces approches sont complémentaires et permettent comme on l'a vu dans différents exemples, particulièrement pour les ciliopathies, d'étudier en détail des systèmes biologiques donnés.

Les résultats obtenus pour ces applications sont encourageants et permettent d'envisager une exploitation à encore plus grande échelle des marqueurs évolutifs. Mon objectif pour la suite est de caractériser non pas un système biologique ciblé, mais l'ensemble des gènes humains, ou d'autre espèce, à la lumière de leur histoire évolutive. Dans cette optique, le développement de représentations des relations évolutives sous forme de réseau ouvre des perspectives prometteuses : on peut envisager de représenter l'ensemble des gènes sous la forme de réseaux

dont les arrêtes seraient les distances évolutives, une sorte de représentation étendue de celle obtenue dans MyGeneFriends. Une telle représentation offre la possibilité d'identifier des modules évolutifs, c'est-à-dire des groupes de gènes avec une histoire similaire et d'associer chacun d'entre eux à une fonction donnée, en exploitant pour cela les techniques d'analyse mises au point, par exemple, pour l'étude des réseaux d'interaction. Une fois ces modules évolutifs mis en évidence, on peut envisager de les associer à des catégories fonctionnelles précises et ainsi, isoler les signatures évolutives de ces fonctions.

Les perspectives d'exploitation des marqueurs évolutifs sont encore plus vastes. Récemment, Wan et ses collaborateurs (Wan et al., 2015) ont caractérisé la trajectoire évolutive des protéines de 981 complexes métazoaires, en s'intéressant à l'âge des gènes les composant. Ils ont pu, de cette façon, discriminer les complexes d'origine ancestrale associés aux fonctions cellulaires centrales des innovations spécifiques aux métazoaires, liées notamment à la multicellularité. Une partie des complexes étant composée à la fois de protéines récentes et ancestrales, ces résultats mettent en évidence les modifications de ces complexes au cours de l'évolution. Dans cette logique, les profils phylogénétiques en tant que marqueurs évolutifs ont le potentiel de capter un spectre plus nuancé d'histoires évolutives pour caractériser les complexes macromoléculaires ou plus généralement, les systèmes biologiques. On peut envisager, avec une représentation adaptée des profils, de se diriger vers une carte évolutive des systèmes biologiques.

L'extension finale de ces travaux, cependant, ne se trouve pas au niveau de l'évolution seule, mais plus surement, dans le concept d'intégration des données. Si, comme on l'a vu, la coévolution de gènes corrèle de façon remarquable avec des modules fonctionnels, voire des pathologies, elle ne s'applique pas à l'intégralité des gènes concernés. Elle offre une vision précise, mais incomplète des relations dans les systèmes biologiques. Les données issues d'autres omiques, de co-expression (transcriptomique) ou d'interactions protéines-protéines (protéomiques) révèlent d'autres associations, parfois spécifiques d'un contexte donné. Représenter l'ensemble de ces données, par le biais d'un réseau multicouche par exemple, permettrait de décrire sous plusieurs angles les systèmes biologiques et d'intégrer « simplement » les notions issues de génomique comparative aux analyses à haut-débit. On passerait d'une carte évolutive à une carte multi-niveau des systèmes biologiques, que l'on pourrait finalement comparer aux modules fonctionnels ou pathologiques connus.

Du réseau d'espèces au réseau du Vivant ?

Les outils que j'ai développés et les pistes énoncées jusqu'à présent proposent d'identifier les relations au sein d'une espèce, soit des gènes entre eux, soit entre les gènes et un caractère phénotypique. Cette approche centrée sur les espèces ignore par principe les gènes perdus dans l'espèce considérée. Par exemple, nos travaux sur les gènes ciliaires humains ont négligé les gènes ciliaires perdus chez l'homme. Si l'on peut considérer que cela a peu d'importance pour la recherche de gènes associés aux ciliopathies humaines, cela peut avoir un impact sur la compréhension générale du système à travers le Vivant, et notamment chez les espèces modèles utilisées justement pour tester expérimentalement la fonction de tel ou tel gène humain.

Pourtant, la génomique comparative permet de s'émanciper des espèces modernes en s'intéressant aux évènements évolutifs ancestraux. Conceptuellement, les gènes descendant d'un ancêtre commun par spéciation sont tous orthologues entre eux et ont donc, le même profil phylogénétique. Il est donc possible de générer des profils phylogénétiques non pas, pour une espèce, mais pour l'ensemble des espèces en prenant comme base un orthogroupe et non plus un gène.

Un tel élargissement permettrait de changer d'échelle dans la conception des réseaux discutés plus haut. Les avantages d'un tel changement de perspective seraient un apport précieux dans l'analyse des modules fonctionnels que l'on pourrait caractériser au plus près de leur équivalent chez un ancêtre commun. D'une façon similaire à ce que KEGG propose sous la forme de PATHWAY *maps* (Aoki-Kinoshita et Kanehisa, 2007) pour les voies métaboliques, cette approche permettrait d'identifier les éléments de ces modules qui constituent des innovations ou des pertes chez les différentes espèces considérées. Pour reprendre l'exemple du cil, une telle représentation pourrait mettre plus facilement en avant les différences entre répertoires de gènes ciliaires de l'homme et des Nématodes dépourvus de cils motiles, dont le modèle *Caenorhabditis elegans* et ainsi, remettre en contexte les équivalences et les divergences entre les deux espèces. Finalement, s'intéresser aux profils phylogénétiques sans les ancrer aux espèces introduit la possibilité d'intégrer dans un même réseau des données omiques issues de différentes espèces, les nœuds d'un tel réseau n'étant plus liés aux gènes d'une seule espèce, mais à tous ses parents.

La définition de profils phylogénétiques indépendants d'une espèce donnée soulève différentes questions. Il s'agit dans un premier temps de passer des représentations paires à paires d'orthologie aux orthogroupes. Cette définition n'existe pas encore dans le cadre d'OrthoInspector et la construction de telles entités n'est pas triviale, notamment à cause d'inévitables erreurs d'inférence empêchant une simple définition par transitivité. Ces développements s'inscrivent cependant dans la droite ligne de mes travaux de thèse et pourront s'appuyer sur les définitions proposées par d'autres ressources d'orthologie, membres de *Quest For Orthologs*. Il est important de noter que, les orthogroupes réunissant tous les gènes ou protéines descendant d'un ancêtre commun, paralogues comme orthologues, ils doivent être définis pour un niveau taxonomique donné afin de prendre en compte les nuances introduites par les duplications à différents niveaux. Par extension, les profils phylogénétiques qui en résulteraient devront également être définis pour un niveau taxonomique donné.

Vers de nouveaux marqueurs évolutifs

Les pistes évoquées plus haut ouvrent de nouvelles perspectives d'exploitation en génomique comparative, tant pour des applications ciblées visant les relations génotype-phénotype dans un système biologique que pour l'étude globale des mécanismes évolutifs. Les marqueurs définis tout au long de mes travaux de thèse restent cependant inféodés aux relations d'orthologie entre protéines. Or, le profilage phylogénétique, tout comme la notion d'orthologie, s'applique à tous les objets biologiques qui descendent d'un ancêtre commun. Les relations d'orthologie au niveau du gène protéique ne prennent pas en compte la nature mosaïque des protéines, qui

peuvent être composées de plusieurs domaines. L'architecture en domaines de deux orthologues peut être différente, ce qui a, à son tour, un effet sur sa fonction. La question de la prise en compte du domaine dans les prédictions d'orthologie est au cœur des questionnements de la communauté et un thème central d'un des derniers *meetings* de *Quest For Orthologs* (Forslund et al., 2018). Si les domaines constituent l'unité fonctionnelle, voire l'unité évolutive, il apparait judicieux de les intégrer pleinement dans la représentation des profils phylogénétiques. Cette approche permettrait de rendre compte de l'évolution des protéines au-delà de la simple présence ou absence d'orthologue, en intégrant des pertes ou gains de domaines liés à des spécialisations fonctionnelles ou à des modifications d'interaction protéine-protéine. En définitive, intégrer le domaine dans la définition des marqueurs évolutifs ferait passer les analyses à un niveau de granularité plus fin dans l'extraction des données de génomique comparative. Cette progression vers un nouveau niveau de représentation encore plus précis se place dans la continuité directe de mes travaux.

Les gènes ou les domaines protéiques ne sont pas les seuls éléments en jeu dans les systèmes biologiques, les gènes non-codants ont également un rôle considérable. Je me suis concentré sur les protéines pour deux raisons principales : les annotations de gènes protéiques issues d'un génome sont relativement « complètes » pour la plupart des espèces considérées et les méthodes d'identification de gènes orthologues se basent le plus souvent sur les séquences protéiques. A l'inverse, à l'exception des ARN les plus conservés, à savoir les ARN ribosomiques et les ARN de transfert, les annotations de gènes non codants sont difficiles à réaliser à partir de la séquence génomique et reposent principalement sur des données de transcriptomique. Cette limitation réduit considérablement la puissance du profilage phylogénétique qui nécessite des répertoires de gènes complets pour être efficace. Un second défi de poids est bien sûr l'identification d'orthologues entre ces gènes, les séquences étant souvent moins conservées que celles des gènes protéiques. Si ces défis techniques constituent à l'heure actuelle des obstacles majeurs, le profilage phylogénétique demeure théoriquement applicable aux gènes non codants, voire à des éléments non géniques comme les éléments de régulation, ce qui laisse la porte ouverte aux avancées scientifiques du domaine dans les années à venir.

Références

1000 Genomes Project Consortium, Auton, A., Brooks, L.D., Durbin, R.M., Garrison, E.P., Kang, H.M., Korbel, J.O., Marchini, J.L., McCarthy, S., McVean, G.A., et al. (2015). A global reference for human genetic variation. Nature *526*, 68–74.

Abbasi, F., Miyata, H., Shimada, K., Morohoshi, A., Nozawa, K., Matsumura, T., Xu, Z., Pratiwi, P., and Ikawa, M. (2018). RSPH6A is required for sperm flagellum formation and male fertility in mice. J. Cell. Sci. *131*.

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*, 429–493.

Aguinaldo, A.M., Turbeville, J.M., Linford, L.S., Rivera, M.C., Garey, J.R., Raff, R.A., and Lake, J.A. (1997). Evidence for a clade of nematodes, arthropods and other moulting animals. Nature *387*, 489–493.

Albalat, R., and Cañestro, C. (2016). Evolution by gene loss. Nat. Rev. Genet. *17*, 379–391.

Alexander Pyron, R., and Wiens, J.J. (2011). A large-scale phylogeny of Amphibia including over 2800 species, and a revised classification of extant frogs, salamanders, and caecilians. Molecular Phylogenetics and Evolution *61*, 543–583.

Alexeyenko, A., Tamas, I., Liu, G., and Sonnhammer, E.L.L. (2006). Automatic clustering of orthologs and inparalogs shared by multiple proteomes. Bioinformatics *22*, e9-15.

Allot, A., Chennen, K., Nevers, Y., Poidevin, L., Kress, A., Ripp, R., Thompson, J.D., Poch, O., and Lecompte, O. (2017). MyGeneFriends: A Social Network Linking Genes, Genetic Diseases, and Researchers. J. Med. Internet Res. *19*, e212.

Altelaar, A.F.M., Munoz, J., and Heck, A.J.R. (2013). Next-generation proteomics: towards an integrative view of proteome dynamics. Nat. Rev. Genet. *14*, 35–48.

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*.

Altenhoff, A.M., Gil, M., Gonnet, G.H., and Dessimoz, C. (2013). Inferring hierarchical orthologous groups from orthologous gene pairs. PLoS ONE *8*, e53786.

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.

Altenhoff, A.M., Glover, N.M., Train, C.-M., Kaleb, K., Warwick Vesztrocy, A., Dylus, D., de Farias, T.M., Zile, K., Stevenson, C., Long, J., 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.

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.

Amar, D., Frades, I., Danek, A., Goldberg, T., Sharma, S.K., Hedley, P.E., Proux-Wera, E., Andreasson, E., Shamir, R., Tzfadia, O., 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*.

Amberger, J.S., and Hamosh, A. (2017). Searching Online Mendelian Inheritance in Man (OMIM): A Knowledgebase of Human Genes and Genetic Phenotypes. Curr Protoc Bioinformatics *58*, 1.2.1-1.2.12.

Aoki-Kinoshita, K.F., and Kanehisa, M. (2007). Gene annotation and pathway mapping in KEGG. Methods Mol. Biol. *396*, 71–91.

Arnaiz, O., Malinowska, A., Klotz, C., Sperling, L., Dadlez, M., Koll, F., and Cohen, J. (2009). Cildb: a knowledgebase for centrosomes and cilia. Database (Oxford) *2009*, bap022.

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.

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., et al. (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat. Genet. *25*, 25–29.

Aury, J.-M., Jaillon, O., Duret, L., Noel, B., Jubin, C., Porcel, B.M., Ségurens, B., Daubin, V., Anthouard, V., Aiach, N., et al. (2006). Global trends of whole-genome duplications revealed by the ciliate Paramecium tetraurelia. Nature *444*, 171–178.

Avery, O.T., MacLeod, C.M., and McCarty, M. (1944). STUDIES ON THE CHEMICAL NATURE OF THE SUBSTANCE INDUCING TRANSFORMATION OF PNEUMOCOCCAL TYPES. J Exp Med *79*, 137–158.

Avidor-Reiss, T., Maer, A.M., Koundakjian, E., Polyanovsky, A., Keil, T., Subramaniam, S., and Zuker, C.S. (2004). Decoding Cilia Function: Defining Specialized Genes Required for Compartmentalized Cilia Biogenesis. Cell *117*, 527–539.

Bairoch, A. (2000). The ENZYME database in 2000. Nucleic Acids Res. *28*, 304–305.

Barabási, A.-L., Gulbahce, N., and Loscalzo, J. (2011). Network Medicine: A Network-based Approach to Human Disease. Nat Rev Genet *12*, 56–68.

Barker, D., and Pagel, M. (2005). Predicting Functional Gene Links from Phylogenetic-Statistical Analyses of Whole Genomes. PLoS Comput Biol *1*.

Barker, D., Meade, A., and Pagel, M. (2007). Constrained models of evolution lead to improved prediction of functional linkage from correlated gain and loss of genes. Bioinformatics *23*, 14–20.

Beck, C., Knoop, H., and Steuer, R. (2018). Modules of co-occurrence in the cyanobacterial pan-genome reveal functional associations between groups of ortholog genes. PLoS Genetics *14*.

Bedez, F., Linard, B., Brochet, X., Ripp, R., Thompson, J.D., Moras, D., Lecompte, O., and Poch, O. (2013). Functional insights into the core-TFIIH from a comparative survey. Genomics *101*, 178–186.

Bennett, G.M., and Moran, N.A. (2013). Small, Smaller, Smallest: The Origins and Evolution of Ancient Dual Symbioses in a Phloem-Feeding Insect. Genome Biol Evol *5*, 1675–1688.

Bentley, D.R., Balasubramanian, S., Swerdlow, H.P., Smith, G.P., Milton, J., Brown, C.G., Hall, K.P., Evers, D.J., Barnes, C.L., Bignell, H.R., et al. (2008). Accurate whole human genome sequencing using reversible terminator chemistry. Nature *456*, 53–59.

Bininda-Emonds, O.R.P., Cardillo, M., Jones, K.E., MacPhee, R.D.E., Beck, R.M.D., Grenyer, R., Price, S.A., Vos, R.A., Gittleman, J.L., and Purvis, A. (2007). The delayed rise of present-day mammals. Nature *446*, 507–512.

Binns, D., Dimmer, E., Huntley, R., Barrell, D., O'Donovan, C., and Apweiler, R. (2009). QuickGO: a web-based tool for Gene Ontology searching. Bioinformatics *25*, 3045–3046.

Bitard-Feildel, T., Kemena, C., Greenwood, J.M., and Bornberg-Bauer, E. (2015). Domain similarity based orthology detection. BMC Bioinformatics *16*.

Blanchette, M., and Tompa, M. (2002). Discovery of Regulatory Elements by a Computational Method for Phylogenetic Footprinting. Genome Res *12*, 739–748.

Boffelli, D., McAuliffe, J., Ovcharenko, D., Lewis, K.D., Ovcharenko, I., Pachter, L., and Rubin, E.M. (2003). Phylogenetic Shadowing of Primate Sequences to Find Functional Regions of the Human Genome. Science *299*, 1391–1394.

Boldt, K., van Reeuwijk, J., Lu, Q., Koutroumpas, K., Nguyen, T.-M.T., Texier, Y., van Beersum, S.E.C., Horn, N., Willer, J.R., Mans, D.A., et al. (2016). An organelle-specific protein landscape identifies novel diseases and molecular mechanisms. Nat Commun *7*, 11491.

Brownlee, G.G., Sanger, F., and Barrell, B.G. (1967). Nucleotide sequence of 5S-ribosomal RNA from Escherichia coli. Nature *215*, 735–736.

Buchfink, B., Xie, C., and Huson, D.H. (2015). Fast and sensitive protein alignment using DIAMOND. Nat. Methods *12*, 59–60.

Buljan, M., Frankish, A., and Bateman, A. (2010). Quantifying the mechanisms of domain gain in animal proteins. Genome Biology *11*, R74.

Calvo, S., Jain, M., Xie, X., Sheth, S.A., Chang, B., Goldberger, O.A., Spinazzola, A., Zeviani, M., Carr, S.A., and Mootha, V.K. (2006). Systematic identification of human mitochondrial disease genes through integrative genomics. Nat. Genet. *38*, 576–582.

Calvo, S.E., Clauser, K.R., and Mootha, V.K. (2016). MitoCarta2.0: an updated inventory of mammalian mitochondrial proteins. Nucleic Acids Res. *44*, D1251-1257.

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.

Canning, P., Park, K., Gonçalves, J., Li, C., Howard, C.J., Sharpe, T.D., Holt, L.J., Pelletier, L., Bullock, A.N., and Leroux, M.R. (2018). CDKL Family Kinases Have Evolved Distinct Structural Features and Ciliary Function. Cell Rep *22*, 885–894.

Castaneda, J.M., Hua, R., Miyata, H., Oji, A., Guo, Y., Cheng, Y., Zhou, T., Guo, X., Cui, Y., Shen, B., et al. (2017). TCTE1 is a conserved component of the dynein regulatory complex and is required for motility and metabolism in mouse spermatozoa. Proc. Natl. Acad. Sci. U.S.A. *114*, E5370–E5378.

Castelle, C.J., Wrighton, K.C., Thomas, B.C., Hug, L.A., Brown, C.T., Wilkins, M.J., Frischkorn, K.R., Tringe, S.G., Singh, A., Markillie, L.M., et al. (2015). Genomic Expansion of Domain Archaea Highlights Roles for Organisms from New Phyla in Anaerobic Carbon Cycling. Current Biology *25*, 690–701.

Chen, X., and Zhang, J. (2012). The Ortholog Conjecture Is Untestable by the Current Gene Ontology but Is Supported by RNA Sequencing Data. PLOS Computational Biology *8*, e1002784.

Chen, F., Mackey, A.J., Stoeckert, C.J., and Roos, D.S. (2006). OrthoMCL-DB: querying a comprehensive multi-species collection of ortholog groups. Nucleic Acids Res. *34*, D363-368.

Cherian, P.T., Al-Khairi, I., Sriraman, D., Al-Enezi, A., Al-Sultan, D., AlOtaibi, M., Al-Enezi, S., Tuomilehto, J., Al-Mulla, F., Abubaker, J.A., et al. (2018). Increased Circulation and Adipose Tissue Levels of DNAJC27/RBJ in Obesity and Type 2-Diabetes. Front Endocrinol (Lausanne) *9*, 423.

Chiu, J.C., Lee, E.K., Egan, M.G., Sarkar, I.N., Coruzzi, G.M., and DeSalle, R. (2006). OrthologID: automation of genome-scale ortholog identification within a parsimony framework. Bioinformatics *22*, 699–707.

Clarke, J., Wu, H.-C., Jayasinghe, L., Patel, A., Reid, S., and Bayley, H. (2009). Continuous base identification for single-molecule nanopore DNA sequencing. Nat Nanotechnol *4*, 265–270.

Cobb, M. (2006). Heredity before genetics: a history. Nat. Rev. Genet. *7*, 953–958.

Cooper, G.M., Stone, E.A., Asimenos, G., Green, E.D., Batzoglou, S., and Sidow, A. (2005). Distribution and intensity of constraint in mammalian genomic sequence. Genome Res *15*, 901–913.

Corradi, N., and Slamovits, C.H. (2011). The intriguing nature of microsporidian genomes. Brief Funct Genomics *10*, 115–124.

Cosentino, S., and Iwasaki, W. (2018). SonicParanoid: fast, accurate, and easy orthology inference. Bioinformatics.

Crawford, N.G., Faircloth, B.C., McCormack, J.E., Brumfield, R.T., Winker, K., and Glenn, T.C. (2012). More than 1000 ultraconserved elements provide evidence that turtles are the sister group of archosaurs. Biol. Lett. *8*, 783–786.

Crick, F.H. (1958). On protein synthesis. Symp. Soc. Exp. Biol. *12*, 138–163.

Criscuolo, A., and Gribaldo, S. (2011). Large-Scale Phylogenomic Analyses Indicate a Deep Origin of Primary Plastids within Cyanobacteria. Mol Biol Evol *28*, 3019–3032.

Croft, D., Mundo, A.F., Haw, R., Milacic, M., Weiser, J., Wu, G., Caudy, M., Garapati, P., Gillespie, M., Kamdar, M.R., et al. (2014). The Reactome pathway knowledgebase. Nucleic Acids Res. *42*, D472-477.

Cunningham, F.X., Lafond, T.P., and Gantt, E. (2000). Evidence of a Role for LytB in the Nonmevalonate Pathway of Isoprenoid Biosynthesis. J Bacteriol *182*, 5841–5848.

Curtis, D.S., Phillips, A.R., Callister, S.J., Conlan, S., and McCue, L.A. (2013). SPOCS: software for predicting and visualizing orthology/paralogy relationships among genomes. Bioinformatics *29*, 2641–2642.

van Dam, T.J., Wheway, G., Slaats, G.G., SYSCILIA Study Group, Huynen, M.A., and Giles, R.H. (2013). The SYSCILIA gold standard (SCGSv1) of known ciliary components and its applications within a systems biology consortium. Cilia *2*, 7.

Darby, C.A., Stolzer, M., Ropp, P.J., Barker, D., and Durand, D. (2017). Xenolog classification. Bioinformatics *33*, 640–649.

Dehal, P., and Boore, J.L. (2005). Two rounds of whole genome duplication in the ancestral vertebrate. PLoS Biol. *3*, e314.

Dehal, P.S., and Boore, J.L. (2006). A phylogenomic gene cluster resource: the Phylogenetically Inferred Groups (PhIGs) database. BMC Bioinformatics *7*, 201.

DeLuca, T.F., Cui, J., Jung, J.-Y., St. Gabriel, K.C., and Wall, D.P. (2012). Roundup 2.0: enabling comparative genomics for over 1800 genomes. Bioinformatics *28*, 715–716.

Dessimoz, C., Gabaldón, T., Roos, D.S., Sonnhammer, E.L.L., Herrero, J., Consortium, the Q. for O., Altenhoff, A., Apweiler, R., Ashburner, M., Blake, J., et al. (2012). Toward community standards in the quest for orthologs. Bioinformatics *28*, 900.

Dey, G., and Meyer, T. (2015). Phylogenetic Profiling for Probing the Modular Architecture of the Human Genome. Cell Syst *1*, 106–115.

Drăgan, M.-A., Moghul, I., Priyam, A., Bustos, C., and Wurm, Y. (2016). GeneValidator: identify problems with protein-coding gene predictions. Bioinformatics *32*, 1559–1561.

Dray, S., and Dufour, A.-B. (2007). The ade4 Package: Implementing the Duality Diagram for Ecologists. Journal of Statistical Software, Articles *22*, 1–20.

Duek, P., Gateau, A., Bairoch, A., and Lane, L. (2018). Exploring the Uncharacterized Human Proteome Using neXtProt. J. Proteome Res.

Dufayard, J.-F., Duret, L., Penel, S., Gouy, M., Rechenmann, F., and Perrière, G. (2005). Tree pattern matching in phylogenetic trees: automatic search for orthologs or paralogs in homologous gene sequence databases. Bioinformatics *21*, 2596–2603.

Dunne, M.P., and Kelly, S. (2017). OrthoFiller: utilising data from multiple species to improve the completeness of genome annotations. BMC Genomics *18*, 390.

Dutkowski, J., Kramer, M., Surma, M.A., Balakrishnan, R., Cherry, J.M., Krogan, N.J., and Ideker, T. (2013). A gene ontology inferred from molecular networks. Nat. Biotechnol. *31*, 38–45.

Ebersberger, I., Strauss, S., and von Haeseler, A. (2009). HaMStR: profile hidden markov model based search for orthologs in ESTs. BMC Evol. Biol. *9*, 157.

Edgar, R.C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. *32*, 1792–1797.

Eichler, E.E., and Sankoff, D. (2003). Structural dynamics of eukaryotic chromosome evolution. Science *301*, 793–797.

Eid, J., Fehr, A., Gray, J., Luong, K., Lyle, J., Otto, G., Peluso, P., Rank, D., Baybayan, P., Bettman, B., et al. (2009). Real-time DNA sequencing from single polymerase molecules. Science *323*, 133–138.

Ekseth, O.K., Kuiper, M., and Mironov, V. (2014). orthAgogue: an agile tool for the rapid prediction of orthology relations. Bioinformatics *30*, 734–736.

Emms, D.M., and Kelly, S. (2015). OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. Genome Biol *16*.

Enault, F., Suhre, K., Abergel, C., Poirot, O., and Claverie, J.-M. (2003). Annotation of bacterial genomes using improved phylogenomic profiles. Bioinformatics *19 Suppl 1*, i105-107.

Enright, A.J., Iliopoulos, I., Kyrpides, N.C., and Ouzounis, C.A. (1999). Protein interaction maps for complete genomes based on gene fusion events. Nature *402*, 86–90.

Eyre, T.A., Wright, M.W., Lush, M.J., and Bruford, E.A. (2007). HCOP: a searchable database of human orthology predictions. Brief. Bioinformatics *8*, 2–5.

Fassad, M.R., Shoemark, A., le Borgne, P., Koll, F., Patel, M., Dixon, M., Hayward, J., Richardson, C., Frost, E., Jenkins, L., et al. (2018). C11orf70 Mutations Disrupting the Intraflagellar Transport-Dependent Assembly of Multiple Axonemal Dyneins Cause Primary Ciliary Dyskinesia. Am. J. Hum. Genet. *102*, 956–972.

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

Fernández-Breis, J.T., Chiba, H., Legaz-García, M.D.C., and Uchiyama, I. (2016). The Orthology Ontology: development and applications. J Biomed Semantics *7*, 34.

Fiers, W., Contreras, R., Duerinck, F., Haegeman, G., Iserentant, D., Merregaert, J., Min Jou, W., Molemans, F., Raeymaekers, A., Van den Berghe, A., et al. (1976). Complete nucleotide sequence of bacteriophage MS2 RNA: primary and secondary structure of the replicase gene. Nature *260*, 500–507.

Finn, R.D., Bateman, A., Clements, J., Coggill, P., Eberhardt, R.Y., Eddy, S.R., Heger, A., Hetherington, K., Holm, L., Mistry, J., et al. (2014). Pfam: the protein families database. Nucleic Acids Res. *42*, D222-230.

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.

Fitch, W.M. (1970). Distinguishing homologous from analogous proteins. Syst. Zool. *19*, 99–113.

Fleischmann, R.D., Adams, M.D., White, O., Clayton, R.A., Kirkness, E.F., Kerlavage, A.R., Bult, C.J., Tomb, J.F., Dougherty, B.A., and Merrick, J.M. (1995). Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. Science *269*, 496–512.

Force, A., Lynch, M., Pickett, F.B., Amores, A., Yan, Y.L., and Postlethwait, J. (1999). Preservation of duplicate genes by complementary, degenerative mutations. Genetics *151*, 1531–1545.

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

Forslund, K., Pereira, C., Capella-Gutierrez, S., da Silva, A.S., Altenhoff, A., Huerta-Cepas, J., Muffato, M., Patricio, M., Vandepoele, K., Ebersberger, I., et al. (2018). Gearing up to handle the mosaic nature of life in the quest for orthologs. Bioinformatics *34*, 323–329.

Fouts, D.E., Brinkac, L., Beck, E., Inman, J., and Sutton, G. (2012). PanOCT: automated clustering of orthologs using conserved gene neighborhood for pan-genomic analysis of bacterial strains and closely related species. Nucleic Acids Res. *40*, e172.

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

Frost, D.R., Grant, T., Faivovich, J., Bain, R.H., Haas, A., Haddad, C.F.B., Sa, D., O, R., Channing, A., Wilkinson, M., et al. (2006). The amphibian tree of life. Bulletin of the AMNH ; no. 297.

Fulton, D.L., Li, Y.Y., Laird, M.R., Horsman, B.G.S., Roche, F.M., and Brinkman, F.S.L. (2006). Improving the specificity of high-throughput ortholog prediction. BMC Bioinformatics *7*, 270.

Furtado, M.B., Merriner, D.J., Berger, S., Rhodes, D., Jamsai, D., and O'Bryan, M.K. (2017). Mutations in the Katnb1 gene cause left-right asymmetry and heart defects. Dev. Dyn. *246*, 1027–1035.

Gabaldón, T., and Huynen, M.A. (2004). Prediction of protein function and pathways in the genome era. Cell. Mol. Life Sci. *61*, 930–944.

Gabaldón, T., Dessimoz, C., Huxley-Jones, J., Vilella, A.J., Sonnhammer, E.L., and Lewis, S. (2009). Joining forces in the quest for orthologs. Genome Biology *10*, 403.

Galperin, M.Y., and Koonin, E.V. (2000). Who's your neighbor? New computational approaches for functional genomics. Nat. Biotechnol. *18*, 609–613.

Garcia-Garcia, J., Schleker, S., Klein-Seetharaman, J., and Oliva, B. (2012). BIPS: BIANA Interolog Prediction Server. A tool for protein–protein interaction inference. Nucleic Acids Res *40*, W147–W151.

Gaudet, P., Livstone, M.S., Lewis, S.E., and Thomas, P.D. (2011). Phylogenetic-based propagation of functional annotations within the Gene Ontology consortium. Brief. Bioinformatics *12*, 449–462.

Gaudet, P., Michel, P.-A., Zahn-Zabal, M., Cusin, I., Duek, P.D., Evalet, O., Gateau, A., Gleizes, A., Pereira, M., Teixeira, D., et al. (2015). The neXtProt knowledgebase on human proteins: current status. Nucleic Acids Res. *43*, D764-770.

Gawad, C., Koh, W., and Quake, S.R. (2016). Single-cell genome sequencing: current state of the science. Nature Reviews Genetics *17*, 175–188.

Gene Ontology Consortium (2015). Gene Ontology Consortium: going forward. Nucleic Acids Res. *43*, D1049-1056.

Gertz, E.M., Yu, Y.-K., Agarwala, R., Schäffer, A.A., and Altschul, S.F. (2006). Composition-based statistics and translated nucleotide searches: improving the TBLASTN module of BLAST. BMC Biol. *4*, 41.

Gilbert, J.A., Blaser, M.J., Caporaso, J.G., Jansson, J.K., Lynch, S.V., and Knight, R. (2018). Current understanding of the human microbiome. Nature Medicine *24*, 392–400.

Gilks, W.R., Audit, B., de Angelis, D., Tsoka, S., and Ouzounis, C.A. (2005). Percolation of annotation errors through hierarchically structured protein sequence databases. Math Biosci *193*, 223–234.

Glazko, G.V., and Mushegian, A.R. (2004). Detection of evolutionarily stable fragments of cellular pathways by hierarchical clustering of phyletic patterns. Genome Biol *5*, R32.

Gligorijević, V., and Pržulj, N. (2015). Methods for biological data integration: perspectives and challenges. J R Soc Interface *12*.

Glover, N.M., Redestig, H., and Dessimoz, C. (2016). Homoeologs: What Are They and How Do We Infer Them? Trends Plant Sci *21*, 609–621.

Goffeau, A., Barrell, B.G., Bussey, H., Davis, R.W., Dujon, B., Feldmann, H., Galibert, F., Hoheisel, J.D., Jacq, C., Johnston, M., et al. (1996). Life with 6000 genes. Science *274*, 546, 563–567.

Gray, G.S., and Fitch, W.M. (1983). Evolution of antibiotic resistance genes: the DNA sequence of a kanamycin resistance gene from Staphylococcus aureus. Mol. Biol. Evol. *1*, 57–66.

Gray, M.W., Burger, G., and Lang, B.F. (1999). Mitochondrial Evolution. Science *283*, 1476–1481.

Gros, F., Hiatt, H., Gilbert, W., Kurland, C.G., Risebrough, R.W., and Watson, J.D. (1961). Unstable ribonucleic acid revealed by pulse labelling of Escherichia coli. Nature *190*, 581–585.

Grossetête, S., Labedan, B., and Lespinet, O. (2010). FUNGIpath: a tool to assess fungal metabolic pathways predicted by orthology. BMC Genomics *11*, 81.

Grubb, S.C., Bult, C.J., and Bogue, M.A. (2014). Mouse Phenome Database. Nucleic Acids Res *42*, D825–D834.

Haag, K.L., James, T.Y., Pombert, J.-F., Larsson, R., Schaer, T.M.M., Refardt, D., and Ebert, D. (2014). Evolution of a morphological novelty occurred before genome compaction in a lineage of extreme parasites. PNAS *111*, 15480–15485.

Hamel, V., Steib, E., Hamelin, R., Armand, F., Borgers, S., Flückiger, I., Busso, C., Olieric, N., Sorzano, C.O.S., Steinmetz, M.O., et al. (2017). Identification of Chlamydomonas Central Core Centriolar Proteins Reveals a Role for Human WDR90 in Ciliogenesis. Curr. Biol. *27*, 2486-2498.e6.

Hedges, S.B., and Poling, L.L. (1999). A molecular phylogeny of reptiles. Science *283*, 998–1001.

van der Heijden, R.T., Snel, B., van Noort, V., and Huynen, M.A. (2007). Orthology prediction at scalable resolution by phylogenetic tree analysis. BMC Bioinformatics *8*, 83.

Hennig, W. (1950). Grundzüge einer Theorie der Phylogenetischen Systematik (Dt. Zentralverl.).

Henricson, A., Forslund, K., and Sonnhammer, E.L.L. (2010). Orthology confers intron position conservation. BMC Genomics *11*, 412.

Herrero, J., Muffato, M., Beal, K., Fitzgerald, S., Gordon, L., Pignatelli, M., Vilella, A.J., Searle, S.M.J., Amode, R., Brent, S., et al. (2016). Ensembl comparative genomics resources. Database (Oxford) *2016*.

Hershey, A.D., and Chase, M. (1952). Independent functions of viral protein and nucleic acid in growth of bacteriophage. J. Gen. Physiol. *36*, 39–56.

Hinchliff, C.E., Smith, S.A., Allman, J.F., Burleigh, J.G., Chaudhary, R., Coghill, L.M., Crandall, K.A., Deng, J., Drew, B.T., Gazis, R., et al. (2015). Synthesis of phylogeny and taxonomy into a comprehensive tree of life. PNAS *112*, 12764–12769.

Ho, M.-R., Chen, C., and Lin, W. (2010). Gene-oriented ortholog database: a functional comparison platform for orthologous loci. Database (Oxford) *2010*, baq002.

Höben, I.M., Hjeij, R., Olbrich, H., Dougherty, G.W., Nöthe-Menchen, T., Aprea, I., Frank, D., Pennekamp, P., Dworniczak, B., Wallmeier, J., et al. (2018). Mutations in C11orf70 Cause Primary Ciliary Dyskinesia with Randomization of Left/Right Body Asymmetry Due to Defects of Outer and Inner Dynein Arms. Am. J. Hum. Genet. *102*, 973–984.

Hodges, M.E., Wickstead, B., Gull, K., and Langdale, J.A. (2011). Conservation of ciliary proteins in plants with no cilia. BMC Plant Biol. *11*, 185.

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

Holley, R.W., Apgar, J., Everett, G.A., Madison, J.T., Marquisee, M., Merrill, S.H., Penswick, J.R., and Zamir, A. (1965). STRUCTURE OF A RIBONUCLEIC ACID. Science *147*, 1462–1465.

Horiike, T., Minai, R., Miyata, D., Nakamura, Y., and Tateno, Y. (2016). Ortholog-Finder: A Tool for Constructing an Ortholog Data Set. Genome Biol Evol *8*, 446–457.

Houle, D., Govindaraju, D.R., and Omholt, S. (2010). Phenomics: the next challenge. Nature Reviews Genetics *11*, 855–866.

Hu, Y., Flockhart, I., Vinayagam, A., Bergwitz, C., Berger, B., Perrimon, N., and Mohr, S.E. (2011). An integrative approach to ortholog prediction for disease-focused and other functional studies. BMC Bioinformatics *12*, 357.

Huang, N., Xia, Y., Zhang, D., Wang, S., Bao, Y., He, R., Teng, J., and Chen, J. (2017). Hierarchical assembly of centriole subdistal appendages via centrosome binding proteins CCDC120 and CCDC68. Nature Communications *8*, 15057.

Huang, T.-W., Lin, C.-Y., and Kao, C.-Y. (2007). Reconstruction of human protein interolog network using evolutionary conserved network. BMC Bioinformatics *8*, 152.

Huerta-Cepas, J., Dopazo, H., Dopazo, J., and Gabaldón, T. (2007). The human phylome. Genome Biol *8*, R109.

Huerta-Cepas, J., Capella-Gutiérrez, S., Pryszcz, L.P., Marcet-Houben, M., and Gabaldón, T. (2014). PhylomeDB v4: zooming into the plurality of evolutionary histories of a genome. Nucleic Acids Res. *42*, D897-902.

Huerta-Cepas, J., Szklarczyk, D., Forslund, K., Cook, H., Heller, D., Walter, M.C., Rattei, T., Mende, D.R., Sunagawa, S., Kuhn, M., 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.

Hug, L.A., Baker, B.J., Anantharaman, K., Brown, C.T., Probst, A.J., Castelle, C.J., Butterfield, C.N., Hernsdorf, A.W., Amano, Y., Ise, K., et al. (2016). A new view of the tree of life. Nature Microbiology *1*, 16048.

Hughes, D. (2000). Evaluating genome dynamics: the constraints on rearrangements within bacterial genomes. Genome Biol *1*, reviews0006.1-reviews0006.8.

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

Huynen, M.A., and Bork, P. (1998). Measuring genome evolution. Proc Natl Acad Sci U S A *95*, 5849–5856.

Huynen, M., Dandekar, T., and Bork, P. (1998). Differential genome analysis applied to the species-specific features of Helicobacter pylori. FEBS Lett. *426*, 1–5.

Inglis, P.N., Boroevich, K.A., and Leroux, M.R. (2006). Piecing together a ciliome. Trends Genet. *22*, 491–500.

Jacob, F., and Monod, J. (1961). Genetic regulatory mechanisms in the synthesis of proteins. J. Mol. Biol. *3*, 318–356.

Jain, M., Koren, S., Miga, K.H., Quick, J., Rand, A.C., Sasani, T.A., Tyson, J.R., Beggs, A.D., Dilthey, A.T., Fiddes, I.T., et al. (2018). Nanopore sequencing and assembly of a human genome with ultra-long reads. Nat. Biotechnol. *36*, 338–345.

Jansen, R., Yu, H., Greenbaum, D., Kluger, Y., Krogan, N.J., Chung, S., Emili, A., Snyder, M., Greenblatt, J.F., and Gerstein, M. (2003). A Bayesian networks approach for predicting protein-protein interactions from genomic data. Science *302*, 449–453.

Jensen, L.J., Julien, P., Kuhn, M., von Mering, C., Muller, J., Doerks, T., and Bork, P. (2008). eggNOG: automated construction and annotation of orthologous groups of genes. Nucleic Acids Res. *36*, D250-254.

Jim, K., Parmar, K., Singh, M., and Tavazoie, S. (2004). A Cross-Genomic Approach for Systematic Mapping of Phenotypic Traits to Genes. Genome Res *14*, 109–115.

Johannsen, W. (1923). SOME REMARKS ABOUT UNITS IN HEREDITY.

Johnson, R.N., O'Meally, D., Chen, Z., Etherington, G.J., Ho, S.Y.W., Nash, W.J., Grueber, C.E., Cheng, Y., Whittington, C.M., Dennison, S., et al. (2018). Adaptation and conservation insights from the koala genome. Nature Genetics *50*, 1102–1111.

Jothi, R., Zotenko, E., Tasneem, A., and Przytycka, T.M. (2006). COCO-CL: hierarchical clustering of homology relations based on evolutionary correlations. Bioinformatics *22*, 779–788.

Jupp, S., Malone, J., Bolleman, J., Brandizi, M., Davies, M., Garcia, L., Gaulton, A., Gehant, S., Laibe, C., Redaschi, N., et al. (2014). The EBI RDF platform: linked open data for the life sciences. Bioinformatics *30*, 1338–1339.

Karnkowska, A., Vacek, V., Zubáčová, Z., Treitli, S.C., Petrželková, R., Eme, L., Novák, L., Žárský, V., Barlow, L.D., Herman, E.K., et al. (2016). A Eukaryote without a Mitochondrial Organelle. Current Biology *26*, 1274–1284.

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.

Ke, Y.-N., and Yang, W.-X. (2014). Primary cilium: an elaborate structure that blocks cell division? Gene *547*, 175–185.

Kensche, P.R., van Noort, V., Dutilh, B.E., and Huynen, M.A. (2008). Practical and theoretical advances in predicting the function of a protein by its phylogenetic distribution. J R Soc Interface *5*, 151–170.

Kharchenko, P., Chen, L., Freund, Y., Vitkup, D., and Church, G.M. (2006). Identifying metabolic enzymes with multiple types of association evidence. BMC Bioinformatics *7*, 177.

Kim, K.I., and Simon, R. (2014). Using single cell sequencing data to model the evolutionary history of a tumor. BMC Bioinformatics *15*, 27.

Kim, C.B., Moon, S.Y., Gelder, S.R., and Kim, W. (1996). Phylogenetic relationships of annelids, molluscs, and arthropods evidenced from molecules and morphology. J. Mol. Evol. *43*, 207–215.

Kim, K., Kim, W., and Kim, S. (2011). ReMark: an automatic program for clustering orthologs flexibly combining a Recursive and a Markov clustering algorithms. Bioinformatics *27*, 1731–1733.

Kocot, K.M., Citarella, M.R., Moroz, L.L., and Halanych, K.M. (2013). PhyloTreePruner: A Phylogenetic Tree-Based Approach for Selection of Orthologous Sequences for Phylogenomics. Evol Bioinform Online *9*, 429–435.

Köhler, S., Vasilevsky, N.A., Engelstad, M., Foster, E., McMurry, J., Aymé, S., Baynam, G., Bello, S.M., Boerkoel, C.F., Boycott, K.M., et al. (2017). The Human Phenotype Ontology in 2017. Nucleic Acids Res. *45*, D865–D876.

Kondrashov, F.A., Rogozin, I.B., Wolf, Y.I., and Koonin, E.V. (2002). Selection in the evolution of gene duplications. Genome Biol *3*, research0008.1-research0008.9.

Koonin, E.V. (2009). Evolution of Genome Architecture. Int J Biochem Cell Biol *41*, 298–306.

Koski, L.B., Morton, R.A., and Golding, G.B. (2001). Codon Bias and Base Composition Are Poor Indicators of Horizontally Transferred Genes. Mol Biol Evol *18*, 404–412.

Kriventseva, E.V., Rahman, N., Espinosa, O., and Zdobnov, E.M. (2008). OrthoDB: the hierarchical catalog of eukaryotic orthologs. Nucleic Acids Res. *36*, D271-275.

Kryuchkova-Mostacci, N., and Robinson-Rechavi, M. (2016). Tissue-Specificity of Gene Expression Diverges Slowly between Orthologs, and Rapidly between Paralogs. PLoS Comput. Biol. *12*, e1005274.

Lafond, M., Meghdari Miardan, M., and Sankoff, D. (2018). Accurate prediction of orthologs in the presence of divergence after duplication. Bioinformatics *34*, i366–i375.

Lamour, K.H., Win, J., and Kamoun, S. (2007). Oomycete genomics: new insights and future directions. FEMS Microbiology Letters *274*, 1–8.

Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., et al. (2001). Initial sequencing and analysis of the human genome. Nature *409*, 860–921.

Lawrence, J.G., and Ochman, H. (2002). Reconciling the many faces of lateral gene transfer. Trends Microbiol. *10*, 1–4.

Lechner, M., Findeiss, S., Steiner, L., Marz, M., Stadler, P.F., and Prohaska, S.J. (2011). Proteinortho: detection of (co-)orthologs in large-scale analysis. BMC Bioinformatics *12*, 124.

Lechtreck, K.F. (2015). IFT-Cargo Interactions and Protein Transport in Cilia. Trends Biochem. Sci. *40*, 765–778.

Lecompte, O., Ripp, R., Puzos-Barbe, V., Duprat, S., Heilig, R., Dietrich, J., Thierry, J.C., and Poch, O. (2001). Genome evolution at the genus level: comparison of three complete genomes of hyperthermophilic archaea. Genome Res. *11*, 981–993.

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

Lee, I., Date, S.V., Adai, A.T., and Marcotte, E.M. (2004). A probabilistic functional network of yeast genes. Science *306*, 1555–1558.

Lee, Y., Sultana, R., Pertea, G., Cho, J., Karamycheva, S., Tsai, J., Parvizi, B., Cheung, F., Antonescu, V., White, J., et al. (2002). Cross-Referencing Eukaryotic Genomes: TIGR Orthologous Gene Alignments (TOGA). Genome Res *12*, 493–502.

van der Lee, R., Feng, Q., Langereis, M.A., Ter Horst, R., Szklarczyk, R., Netea, M.G., Andeweg, A.C., van Kuppeveld, F.J.M., and Huynen, M.A. (2015). Integrative Genomics-Based Discovery of Novel Regulators of the Innate Antiviral Response. PLoS Comput. Biol. *11*, e1004553.

Levy, S., Sutton, G., Ng, P.C., Feuk, L., Halpern, A.L., Walenz, B.P., Axelrod, N., Huang, J., Kirkness, E.F., Denisov, G., et al. (2007). The diploid genome sequence of an individual human. PLoS Biol. *5*, e254.

Lewin, H.A., Robinson, G.E., Kress, W.J., Baker, W.J., Coddington, J., Crandall, K.A., Durbin, R., Edwards, S.V., Forest, F., Gilbert, M.T.P., et al. (2018). Earth BioGenome Project: Sequencing life for the future of life. PNAS *115*, 4325–4333.

Ley, R.E., Turnbaugh, P.J., Klein, S., and Gordon, J.I. (2006). Microbial ecology: human gut microbes associated with obesity. Nature *444*, 1022–1023.

Li, H., Coghlan, A., Ruan, J., Coin, L.J., Hériché, J.-K., Osmotherly, L., Li, R., Liu, T., Zhang, Z., Bolund, L., et al. (2006). TreeFam: a curated database of phylogenetic trees of animal gene families. Nucleic Acids Res. *34*, D572-580.

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, L., Stoeckert, C.J., and Roos, D.S. (2003). OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res. *13*, 2178–2189.

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.

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.

Liu, Y.J., Hodson, M.C., and Hall, B.D. (2006). Loss of the flagellum happened only once in the fungal lineage: phylogenetic structure of kingdom Fungi inferred from RNA polymerase II subunit genes. BMC Evol. Biol. *6*, 74.

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

Lucas, A. (2018). amap: Another Multidimensional Analysis Package.

Lukjancenko, O., Wassenaar, T.M., and Ussery, D.W. (2010). Comparison of 61 sequenced Escherichia coli genomes. Microb. Ecol. *60*, 708–720.

Lynch, M., and Conery, J.S. (2000). The evolutionary fate and consequences of duplicate genes. Science *290*, 1151–1155.

Makarova, K.S., Wolf, Y.I., and Koonin, E.V. (2003). Potential genomic determinants of hyperthermophily. Trends Genet. *19*, 172–176.

Marcotte, E.M. (2000). Computational genetics: finding protein function by nonhomology methods. Current Opinion in Structural Biology *10*, 359–365.

Marcotte, E.M., Pellegrini, M., Ng, H.L., Rice, D.W., Yeates, T.O., and Eisenberg, D. (1999). Detecting protein function and protein-protein interactions from genome sequences. Science *285*, 751–753.

Margulies, E.H., Blanchette, M., Haussler, D., and Green, E.D. (2003). Identification and Characterization of Multi-Species Conserved Sequences. Genome Res *13*, 2507–2518.

Margulies, M., Egholm, M., Altman, W.E., Attiya, S., Bader, J.S., Bemben, L.A., Berka, J., Braverman, M.S., Chen, Y.-J., Chen, Z., et al. (2005). Genome sequencing in microfabricated high-density picolitre reactors. Nature *437*, 376–380.

Martin, R.G., Matthaei, J.H., Jones, O.W., and Nirenberg, M.W. (1962). Ribonucleotide composition of the genetic code. Biochem. Biophys. Res. Commun. *6*, 410–414.

Maston, G.A., Evans, S.K., and Green, M.R. (2006). Transcriptional regulatory elements in the human genome. Annu Rev Genomics Hum Genet *7*, 29–59.

May-Simera, H., Nagel-Wolfrum, K., and Wolfrum, U. (2017). Cilia - The sensory antennae in the eye. Progress in Retinal and Eye Research *60*, 144–180.

Merchant, S.S., Prochnik, S.E., Vallon, O., Harris, E.H., Karpowicz, S.J., Witman, G.B., Terry, A., Salamov, A., Fritz-Laylin, L.K., Maréchal-Drouard, L., et al. (2007). The Chlamydomonas genome reveals the evolution of key animal and plant functions. Science *318*, 245–250.

von Mering, C., Huynen, M., Jaeggi, D., Schmidt, S., Bork, P., and Snel, B. (2003). STRING: a database of predicted functional associations between proteins. Nucleic Acids Res *31*, 258–261.

Merkeev, I.V., Novichkov, P.S., and Mironov, A.A. (2006). PHOG: a database of supergenomes built from proteome complements. BMC Evol Biol *6*, 52.

Mi, H., Dong, Q., Muruganujan, A., Gaudet, P., Lewis, S., and Thomas, P.D. (2010). PANTHER version 7: improved phylogenetic trees, orthologs and collaboration with the Gene Ontology Consortium. Nucleic Acids Res. *38*, D204-210.

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*, D336-342.

Miller, J.B., Pickett, B.D., and Ridge, P.G. (2018). JustOrthologs: A Fast, Accurate, and User-Friendly Ortholog Identification Algorithm. Bioinformatics.

Miñarro-Gimenez, J.A., Madrid, M., and Fernandez-Breis, J.T. (2009). OGO: an ontological approach for integrating knowledge about orthology. BMC Bioinformatics *10 Suppl 10*, S13.

Moran, N.A. (2003). Tracing the evolution of gene loss in obligate bacterial symbionts. Curr. Opin. Microbiol. *6*, 512–518.

Morris, K.V., and Mattick, J.S. (2014). The rise of regulatory RNA. Nat Rev Genet *15*, 423–437.

Mouse Genome Sequencing Consortium, Waterston, R.H., Lindblad-Toh, K., Birney, E., Rogers, J., Abril, J.F., Agarwal, P., Agarwala, R., Ainscough, R., Alexandersson, M., et al. (2002). Initial sequencing and comparative analysis of the mouse genome. Nature *420*, 520–562.

Mukherjee, S., Stamatis, D., Bertsch, J., Ovchinnikova, G., Verezemska, O., Isbandi, M., Thomas, A.D., Ali, R., Sharma, K., Kyrpides, N.C., et al. (2017). Genomes OnLine Database (GOLD) v.6: data updates and feature enhancements. Nucleic Acids Res *45*, D446–D456.

Müller, M., Mentel, M., van Hellemond, J.J., Henze, K., Woehle, C., Gould, S.B., Yu, R.-Y., van der Giezen, M., Tielens, A.G.M., and Martin, W.F. (2012). Biochemistry and evolution of anaerobic energy metabolism in eukaryotes. Microbiol. Mol. Biol. Rev. *76*, 444–495.

Nakaya, A., Katayama, T., Itoh, M., Hiranuka, K., Kawashima, S., Moriya, Y., Okuda, S., Tanaka, M., Tokimatsu, T., Yamanishi, Y., et al. (2013). KEGG OC: a large-scale automatic construction of taxonomy-based ortholog clusters. Nucleic Acids Res. *41*, D353-357.

Needleman, S.B., and Wunsch, C.D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. J. Mol. Biol. *48*, 443–453.

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 Computational Biology *7*, e1002073.

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. (2018). OrthoInspector 3.0: open portal for comparative genomics. Nucleic Acids Res.

Niimura, Y., Matsui, A., and Touhara, K. (2014). Extreme expansion of the olfactory receptor gene repertoire in African elephants and evolutionary dynamics of orthologous gene groups in 13 placental mammals. Genome Res *24*, 1485–1496.

Niu, Y., Moghimyfiroozabad, S., Safaie, S., Yang, Y., Jonas, E.A., and Alavian, K.N. (2017). Phylogenetic Profiling of Mitochondrial Proteins and Integration Analysis of Bacterial Transcription Units Suggest Evolution of F1Fo ATP Synthase from Multiple Modules. Journal of Molecular Evolution *85*, 219.

Notredame, C., Higgins, D.G., and Heringa, J. (2000). T-Coffee: A novel method for fast and accurate multiple sequence alignment. J. Mol. Biol. *302*, 205–217.

O'Brien, K.P., Remm, M., and Sonnhammer, E.L.L. (2005). Inparanoid: a comprehensive database of eukaryotic orthologs. Nucleic Acids Res. *33*, D476-480.

Osterman, A., and Overbeek, R. (2003). Missing genes in metabolic pathways: a comparative genomics approach. Curr Opin Chem Biol *7*, 238–251.

Overbeek, R., Fonstein, M., D'Souza, M., Pusch, G.D., and Maltsev, N. (1999). The use of gene clusters to infer functional coupling. Proc. Natl. Acad. Sci. U.S.A. *96*, 2896–2901.

Owen, R. (1843). Lectures on the Comparative Anatomy and Physiology of the Invertebrate Animals: Delivered at the Royal College of Surgeons, in 1843 (Longman, Brown, Green, and Longmans).

Pabinger, S., Dander, A., Fischer, M., Snajder, R., Sperk, M., Efremova, M., Krabichler, B., Speicher, M.R., Zschocke, J., and Trajanoski, Z. (2014). A survey of tools for variant analysis of next-generation genome sequencing data. Brief Bioinform *15*, 256–278.

Pasek, S., Risler, J.-L., and Brézellec, P. (2006). Gene fusion/fission is a major contributor to evolution of multi-domain bacterial proteins. Bioinformatics *22*, 1418–1423.

Patel-King, R.S., and King, S.M. (2016). A prefoldin-associated WD-repeat protein (WDR92) is required for the correct architectural assembly of motile cilia. Mol. Biol. Cell *27*, 1204–1209.

Pellegrini, M. (2012). Using phylogenetic profiles to predict functional relationships. Methods Mol. Biol. *804*, 167–177.

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.

Penel, S., Arigon, A.-M., Dufayard, J.-F., Sertier, A.-S., Daubin, V., Duret, L., Gouy, M., and Perrière, G. (2009). Databases of homologous gene families for comparative genomics. BMC Bioinformatics *10 Suppl 6*, S3.

Penkett, C.J., Morris, J.A., Wood, V., and Bähler, J. (2006). YOGY: a web-based, integrated database to retrieve protein orthologs and associated Gene Ontology terms. Nucleic Acids Res *34*, W330–W334.

Pereira, C., Denise, A., and Lespinet, O. (2014). A meta-approach for improving the prediction and the functional annotation of ortholog groups. BMC Genomics *15 Suppl 6*, S16.

Petersen, M., Meusemann, K., Donath, A., Dowling, D., Liu, S., Peters, R.S., Podsiadlowski, L., Vasilikopoulos, A., Zhou, X., Misof, B., et al. (2017). Orthograph: a versatile tool for mapping coding nucleotide sequences to clusters of orthologous genes. BMC Bioinformatics *18*.

Peterson, M.E., Chen, F., Saven, J.G., Roos, D.S., Babbitt, P.C., and Sali, A. (2009). Evolutionary constraints on structural similarity in orthologs and paralogs. Protein Sci *18*, 1306–1315.

Pombert, J.-F., Blouin, N.A., Lane, C., Boucias, D., and Keeling, P.J. (2014). A lack of parasitic reduction in the obligate parasitic green alga Helicosporidium. PLoS Genet. *10*, e1004355.

Pombert, J.-F., Haag, K.L., Beidas, S., Ebert, D., and Keeling, P.J. (2015). The Ordospora colligata genome: Evolution of extreme reduction in microsporidia and host-to-parasite horizontal gene transfer. MBio *6*.

Ponting, C.P. (2008). The functional repertoires of metazoan genomes. Nat. Rev. Genet. *9*, 689–698.

Praveen, K., Davis, E.E., and Katsanis, N. (2015). Unique among ciliopathies: primary ciliary dyskinesia, a motile cilia disorder. F1000Prime Rep *7*, 36.

Pryszcz, L.P., Huerta-Cepas, J., and Gabaldón, T. (2011). MetaPhOrs: orthology and paralogy predictions from multiple phylogenetic evidence using a consistency-based confidence score. Nucleic Acids Res *39*, e32.

Puigbò, P., Lobkovsky, A.E., Kristensen, D.M., Wolf, Y.I., and Koonin, E.V. (2014). Genomes in turmoil: quantification of genome dynamics in prokaryote supergenomes. BMC Biol. *12*, 66.

Putnam, N.H., Srivastava, M., Hellsten, U., Dirks, B., Chapman, J., Salamov, A., Terry, A., Shapiro, H., Lindquist, E., Kapitonov, V.V., et al. (2007). Sea anemone genome reveals ancestral eumetazoan gene repertoire and genomic organization. Science *317*, 86–94.

Pyron, R.A. (2015). Post-molecular systematics and the future of phylogenetics. Trends Ecol. Evol. (Amst.) *30*, 384–389.

Pyron, R.A., Burbrink, F.T., and Wiens, J.J. (2013). A phylogeny and revised classification of Squamata, including 4161 species of lizards and snakes. BMC Evolutionary Biology *13*, 93.

Quince, C., Walker, A.W., Simpson, J.T., Loman, N.J., and Segata, N. (2017). Shotgun metagenomics, from sampling to analysis. Nat. Biotechnol. *35*, 833–844.

Rabbani, B., Tekin, M., and Mahdieh, N. (2014). The promise of whole-exome sequencing in medical genetics. Journal of Human Genetics *59*, 5–15.

Rands, C.M., Meader, S., Ponting, C.P., and Lunter, G. (2014). 8.2% of the Human Genome Is Constrained: Variation in Rates of Turnover across Functional Element Classes in the Human Lineage. PLoS Genet *10*.

Rane, R.V., Oakeshott, J.G., Nguyen, T., Hoffmann, A.A., and Lee, S.F. (2017). Orthonome - a new pipeline for predicting high quality orthologue gene sets applicable to complete and draft genomes. BMC Genomics *18*, 673.

Ranea, J.A.G., Yeats, C., Grant, A., and Orengo, C.A. (2007). Predicting Protein Function with Hierarchical Phylogenetic Profiles: The Gene3D Phylo-Tuner Method Applied to Eukaryotic Genomes. PLoS Computational Biology *3*.

Rees, J., and Cranston, K. (2017). Automated assembly of a reference taxonomy for phylogenetic data synthesis. Biodiversity Data Journal *5*, e12581.

Reiter, J.F., and Leroux, M.R. (2017). Genes and molecular pathways underpinning ciliopathies. Nature Reviews Molecular Cell Biology *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.

Remm, M., Storm, C.E., and Sonnhammer, E.L. (2001). Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. J. Mol. Biol. *314*, 1041–1052.

Reyes-Palomares, A., Rodríguez-López, R., Ranea, J.A.G., Jiménez, F.S., and Medina, M.A. (2013). Global Analysis of the Human Pathophenotypic Similarity Gene Network Merges Disease Module Components. PLoS One *8*.

van Rheenen, W., Shatunov, A., Dekker, A.M., McLaughlin, R.L., Diekstra, F.P., Pulit, S.L., van der Spek, R.A.A., Võsa, U., de Jong, S., Robinson, M.R., et al. (2016). Genome-wide association analyses identify new risk variants and the genetic architecture of amyotrophic lateral sclerosis. Nat Genet *48*, 1043–1048.

Rhodes, D.R., Tomlins, S.A., Varambally, S., Mahavisno, V., Barrette, T., Kalyana-Sundaram, S., Ghosh, D., Pandey, A., and Chinnaiyan, A.M. (2005). Probabilistic model of the human protein-protein interaction network. Nature Biotechnology *23*, 951–959.

Riley, D.R., Sieber, K.B., Robinson, K.M., White, J.R., Ganesan, A., Nourbakhsh, S., and Dunning Hotopp, J.C. (2013). Bacteria-human somatic cell lateral gene transfer is enriched in cancer samples. PLoS Comput. Biol. *9*, e1003107.

Roncaglia, P., van Dam, T.J.P., Christie, K.R., Nacheva, L., Toedt, G., Huynen, M.A., Huntley, R.P., Gibson, T.J., and Lomax, J. (2017). The Gene Ontology of eukaryotic cilia and flagella. Cilia *6*, 10.

Roth, A.C., Gonnet, G.H., and Dessimoz, C. (2008). Algorithm of OMA for large-scale orthology inference. BMC Bioinformatics *9*, 518.

Rouard, M., Guignon, V., Aluome, C., Laporte, M.-A., Droc, G., Walde, C., Zmasek, C.M., Périn, C., and Conte, M.G. (2011). GreenPhylDB v2.0: comparative and functional genomics in plants. Nucleic Acids Res. *39*, D1095-1102.

Sagan, L. (1967). On the origin of mitosing cells. Journal of Theoretical Biology *14*, 225-IN6.

Sanger, F., Air, G.M., Barrell, B.G., Brown, N.L., Coulson, A.R., Fiddes, C.A., Hutchison, C.A., Slocombe, P.M., and Smith, M. (1977). Nucleotide sequence of bacteriophage phi X174 DNA. Nature *265*, 687–695.

Satir, P., and Christensen, S.T. (2007). Overview of structure and function of mammalian cilia. Annu. Rev. Physiol. *69*, 377–400.

Sayers, E.W., Barrett, T., Benson, D.A., Bryant, S.H., Canese, K., Chetvernin, V., Church, D.M., DiCuccio, M., Edgar, R., Federhen, S., et al. (2009). Database resources of the National Center for Biotechnology Information. Nucleic Acids Res. *37*, D5-15.

Schirmer, M., Smeekens, S.P., Vlamakis, H., Jaeger, M., Oosting, M., Franzosa, E.A., Ter Horst, R., Jansen, T., Jacobs, L., Bonder, M.J., et al. (2016). Linking the Human Gut Microbiome to Inflammatory Cytokine Production Capacity. Cell *167*, 1125-1136.e8.

Schmitt, T., Messina, D.N., Schreiber, F., and Sonnhammer, E.L.L. (2011). Letter to the editor: SeqXML and OrthoXML: standards for sequence and orthology information. Brief. Bioinformatics *12*, 485–488.

Schnoes, A.M., Brown, S.D., Dodevski, I., and Babbitt, P.C. (2009). Annotation Error in Public Databases: Misannotation of Molecular Function in Enzyme Superfamilies. PLoS Comput Biol *5*.

Schreiber, F., and Sonnhammer, E.L.L. (2013). Hieranoid: hierarchical orthology inference. J. Mol. Biol. *425*, 2072–2081.

Schreiber, F., Pick, K., Erpenbeck, D., Wörheide, G., and Morgenstern, B. (2009). OrthoSelect: a protocol for selecting orthologous groups in phylogenomics. BMC Bioinformatics *10*, 219.

Schreiber, F., Patricio, M., Muffato, M., Pignatelli, M., and Bateman, A. (2014). TreeFam v9: a new website, more species and orthology-on-the-fly. Nucleic Acids Res. *42*, D922-925.

Shen, H., Li, J., Zhang, J., Xu, C., Jiang, Y., Wu, Z., Zhao, F., Liao, L., Chen, J., Lin, Y., et al. (2013). Comprehensive Characterization of Human Genome Variation by High Coverage Whole-Genome Sequencing of Forty Four Caucasians. PLOS ONE *8*, e59494.

Shi, G., Peng, M.-C., and Jiang, T. (2011). MultiMSOAR 2.0: An Accurate Tool to Identify Ortholog Groups among Multiple Genomes. PLoS One *6*.

Shin, J., and Lee, I. (2017). Construction of Functional Gene Networks Using Phylogenetic Profiles. Methods Mol. Biol. *1526*, 87–98.

Sievers, F., Wilm, A., Dineen, D., Gibson, T.J., Karplus, K., Li, W., Lopez, R., McWilliam, H., Remmert, M., Söding, J., et al. (2011). Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol. Syst. Biol. *7*, 539.

Simão, F.A., Waterhouse, R.M., Ioannidis, P., Kriventseva, E.V., and Zdobnov, E.M. (2015). BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics *31*, 3210–3212.

Škunca, N., and Dessimoz, C. (2015). Phylogenetic Profiling: How Much Input Data Is Enough? PLOS ONE *10*, e0114701.

Skunca, N., Bošnjak, M., Kriško, A., Panov, P., Džeroski, S., Smuc, T., and Supek, F. (2013). Phyletic profiling with cliques of orthologs is enhanced by signatures of paralogy relationships. PLoS Comput. Biol. *9*, e1002852.

Smith, T.F., and Waterman, M.S. (1981). Identification of common molecular subsequences. J. Mol. Biol. *147*, 195–197.

Snel, B., and Huynen, M.A. (2004). Quantifying Modularity in the Evolution of Biomolecular Systems. Genome Research *14*, 391.

Song, P., and Perkins, B.D. (2018). Developmental expression of the zebrafish Arf-like small GTPase paralogs arl13a and arl13b. Gene Expr. Patterns *29*, 82–87.

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*, D234-239.

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

Soucy, S.M., Huang, J., and Gogarten, J.P. (2015). Horizontal gene transfer: building the web of life. Nat. Rev. Genet. *16*, 472–482.

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.

Spradling, A., Ganetsky, B., Hieter, P., Johnston, M., Olson, M., Orr-Weaver, T., Rossant, J., Sanchez, A., and Waterston, R. (2006). New Roles for Model Genetic Organisms in Understanding and Treating Human Disease: Report From The 2006 Genetics Society of America Meeting. Genetics *172*, 2025–2032.

Staub, E., Mackowiak, S., and Vingron, M. (2006). An inventory of yeast proteins associated with nucleolar and ribosomal components. Genome Biology *7*, R98.

Steinegger, M., and Söding, J. (2017). MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. Nat. Biotechnol. *35*, 1026–1028.

Stephens, Z.D., Lee, S.Y., Faghri, F., Campbell, R.H., Zhai, C., Efron, M.J., Iyer, R., Schatz, M.C., Sinha, S., and Robinson, G.E. (2015). Big Data: Astronomical or Genomical? PLOS Biology *13*, e1002195.

Storm, C.E.V., and Sonnhammer, E.L.L. (2002). Automated ortholog inference from phylogenetic trees and calculation of orthology reliability. Bioinformatics *18*, 92–99.

Studer, R.A., and Robinson-Rechavi, M. (2009). How confident can we be that orthologs are similar, but paralogs differ? Trends in Genetics *25*, 210–216.

Stunnenberg, H.G., Abrignani, S., Adams, D., Almeida, M. de, Altucci, L., Amin, V., Amit, I., Antonarakis, S.E., Aparicio, S., Arima, T., et al. (2016). The International Human Epigenome Consortium: A Blueprint for Scientific Collaboration and Discovery. Cell *167*, 1145–1149.

Sueoka, N. (1988). Directional mutation pressure and neutral molecular evolution. Proc. Natl. Acad. Sci. U.S.A. *85*, 2653–2657.

Sun, J., Xu, J., Liu, Z., Liu, Q., Zhao, A., Shi, T., and Li, Y. (2005). Refined phylogenetic profiles method for predicting protein–protein interactions. Bioinformatics *21*, 3409–3415.

Sunagawa, S., Coelho, L.P., Chaffron, S., Kultima, J.R., Labadie, K., Salazar, G., Djahanschiri, B., Zeller, G., Mende, D.R., Alberti, A., et al. (2015). Ocean plankton. Structure and function of the global ocean microbiome. Science *348*, 1261359.

Sutphin, G.L., Mahoney, J.M., Sheppard, K., Walton, D.O., and Korstanje, R. (2016). WORMHOLE: Novel Least Diverged Ortholog Prediction through Machine Learning. PLoS Comput. Biol. *12*, e1005182.

Sutton, W.S. (1903). The Chromosomes in Heredity. Biological Bulletin *4*, 231–251.

Szklarczyk, D., Franceschini, A., Wyder, S., Forslund, K., Heller, D., Huerta-Cepas, J., Simonovic, M., Roth, A., Santos, A., Tsafou, K.P., et al. (2015). STRING v10: protein-protein interaction networks, integrated over the tree of life. Nucleic Acids Res. *43*, D447-452.

Szklarczyk, D., Morris, J.H., Cook, H., Kuhn, M., Wyder, S., Simonovic, M., Santos, A., Doncheva, N.T., Roth, A., Bork, P., et al. (2017). The STRING database in 2017: quality-controlled protein–protein association networks, made broadly accessible. Nucleic Acids Res *45*, D362–D368.

Tabach, Y., Golan, T., Hernández-Hernández, A., Messer, A.R., Fukuda, T., Kouznetsova, A., Liu, J.-G., Lilienthal, I., Levy, C., and Ruvkun, G. (2013a). Human disease locus discovery and mapping to molecular pathways through phylogenetic profiling. Mol. Syst. Biol. *9*, 692.

Tabach, Y., Billi, A.C., Hayes, G.D., Newman, M.A., Zuk, O., Gabel, H., Kamath, R., Yacoby, K., Chapman, B., Garcia, S.M., et al. (2013b). Small RNA pathway genes identified by patterns of phylogenetic conservation and divergence. Nature *493*, 694.

Tagle, D.A., Koop, B.F., Goodman, M., Slightom, J.L., Hess, D.L., and Jones, R.T. (1988). Embryonic epsilon and gamma globin genes of a prosimian primate (Galago crassicaudatus). Nucleotide and amino acid sequences, developmental regulation and phylogenetic footprints. J. Mol. Biol. *203*, 439–455.

Takabayashi, A., Ishikawa, N., Obayashi, T., Ishida, S., Obokata, J., Endo, T., and Sato, F. (2009). Three novel subunits of Arabidopsis chloroplastic NAD(P)H dehydrogenase identified by bioinformatic and reverse genetic approaches. Plant J. *57*, 207–219.

Takahashi, T., McDougall, C., Troscianko, J., Chen, W.-C., Jayaraman-Nagarajan, A., Shimeld, S.M., and Ferrier, D.E.K. (2009). An EST screen from the annelid Pomatoceros lamarckii reveals patterns of gene loss and gain in animals. BMC Evol. Biol. *9*, 240.

Takeda, S., and Narita, K. (2012). Structure and function of vertebrate cilia, towards a new taxonomy. Differentiation *83*, S4-11.

Tamames, J. (2001). Evolution of gene order conservation in prokaryotes. Genome Biol. *2*, RESEARCH0020.

Tang, H., Bowers, J.E., Wang, X., Ming, R., Alam, M., and Paterson, A.H. (2008). Synteny and Collinearity in Plant Genomes. Science *320*, 486–488.

Tatusov, R.L., Koonin, E.V., and Lipman, D.J. (1997). A Genomic Perspective on Protein Families. Science *278*, 631–637.

Telenti, A., Pierce, L.C.T., Biggs, W.H., di Iulio, J., Wong, E.H.M., Fabani, M.M., Kirkness, E.F., Moustafa, A., Shah, N., Xie, C., et al. (2016). Deep sequencing of 10,000 human genomes. Proc Natl Acad Sci U S A *113*, 11901–11906.

The UniProt Consortium (2017). UniProt: the universal protein knowledgebase. Nucleic Acids Res. *45*, D158–D169.

Thomas, P.D., Campbell, M.J., Kejariwal, A., Mi, H., Karlak, B., Daverman, R., Diemer, K., Muruganujan, A., and Narechania, A. (2003). PANTHER: A Library of Protein Families and Subfamilies Indexed by Function. Genome Res *13*, 2129–2141.

Thompson, J.D., Higgins, D.G., and Gibson, T.J. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. *22*, 4673–4680.

Thompson, J.D., Plewniak, F., Thierry, J., 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.

Thornton, J.W. (2004). Resurrecting ancient genes: experimental analysis of extinct molecules. Nature Reviews Genetics *5*, 366–375.

Timmis, J.N., Ayliffe, M.A., Huang, C.Y., and Martin, W. (2004). Endosymbiotic gene transfer: organelle genomes forge eukaryotic chromosomes. Nat. Rev. Genet. *5*, 123–135.

Townsend, T., Larson, A., Louis, E., and Macey, J.R. (2004). Molecular phylogenetics of squamata: the position of snakes, amphisbaenians, and dibamids, and the root of the squamate tree. Syst. Biol. *53*, 735–757.

Trapnell, C., Williams, B.A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M.J., Salzberg, S.L., Wold, B.J., and Pachter, L. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat. Biotechnol. *28*, 511–515.

Treangen, T.J., and Rocha, E.P.C. (2011). Horizontal transfer, not duplication, drives the expansion of protein families in prokaryotes. PLoS Genet. *7*, e1001284.

Uchiyama, I. (2003). MBGD: microbial genome database for comparative analysis. Nucleic Acids Res *31*, 58–62.

Uchiyama, I., Mihara, M., Nishide, H., and Chiba, H. (2015). MBGD update 2015: microbial genome database for flexible ortholog analysis utilizing a diverse set of genomic data. Nucleic Acids Res. *43*, D270-276.

Van Bel, M., Diels, T., Vancaester, E., Kreft, L., Botzki, A., Van de Peer, Y., Coppens, F., and Vandepoele, K. (2018). PLAZA 4.0: an integrative resource for functional, evolutionary and comparative plant genomics. Nucleic Acids Res. *46*, D1190–D1196.

Van de Peer, Y., Maere, S., and Meyer, A. (2009). The evolutionary significance of ancient genome duplications. Nat. Rev. Genet. *10*, 725–732.

Van De Weghe, J.C., Rusterholz, T.D.S., Latour, B., Grout, M.E., Aldinger, K.A., Shaheen, R., Dempsey, J.C., Maddirevula, S., Cheng, Y.-H.H., Phelps, I.G., et al. (2017). Mutations in ARMC9, which Encodes a Basal Body Protein, Cause Joubert Syndrome in Humans and Ciliopathy Phenotypes in Zebrafish. Am. J. Hum. Genet. *101*, 23–36.

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.

Vanunu, O., Magger, O., Ruppin, E., Shlomi, T., and Sharan, R. (2010). Associating genes and protein complexes with disease via network propagation. PLoS Comput. Biol. *6*, e1000641.

Vert, J.-P. (2002). A tree kernel to analyse phylogenetic profiles. Bioinformatics *18 Suppl 1*, S276-284.

Vilella, A.J., Severin, J., Ureta-Vidal, A., Heng, L., Durbin, R., and Birney, E. (2009). EnsemblCompara GeneTrees: Complete, duplication-aware phylogenetic trees in vertebrates. Genome Res. *19*, 327–335.

Wagner, I., Volkmer, M., Sharan, M., Villaveces, J.M., Oswald, F., Surendranath, V., and Habermann, B.H. (2014). morFeus: a web-based program to detect remotely conserved orthologs using symmetrical best hits and orthology network scoring. BMC Bioinformatics *15*, 263.

Walhout, A.J., Sordella, R., Lu, X., Hartley, J.L., Temple, G.F., Brasch, M.A., Thierry-Mieg, N., and Vidal, M. (2000). Protein interaction mapping in C. elegans using proteins involved in vulval development. Science *287*, 116–122.

Wall, D.P., Fraser, H.B., and Hirsh, A.E. (2003). Detecting putative orthologs. Bioinformatics *19*, 1710–1711.

Wan, C., Borgeson, B., Phanse, S., Tu, F., Drew, K., Clark, G., Xiong, X., Kagan, O., Kwan, J., Bezginov, A., et al. (2015). Panorama of ancient metazoan macromolecular complexes. Nature *525*, 339–344.

Wang, L., and Jiang, T. (1994). On the complexity of multiple sequence alignment. J. Comput. Biol. *1*, 337–348.

Waters, E., Hohn, M.J., Ahel, I., Graham, D.E., Adams, M.D., Barnstead, M., Beeson, K.Y., Bibbs, L., Bolanos, R., Keller, M., et al. (2003). The genome of Nanoarchaeum equitans:

Insights into early archaeal evolution and derived parasitism. Proc Natl Acad Sci U S A *100*, 12984–12988.

Watson, J.D., and Crick, F.H. (1953). Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. Nature *171*, 737–738.

Weißenborn, S., and Walther, D. (2017). Metabolic Pathway Assignment of Plant Genes based on Phylogenetic Profiling–A Feasibility Study. Front Plant Sci *8*.

Wheeler, D.A., Srinivasan, M., Egholm, M., Shen, Y., Chen, L., McGuire, A., He, W., Chen, Y.-J., Makhijani, V., Roth, G.T., et al. (2008). The complete genome of an individual by massively parallel DNA sequencing. Nature *452*, 872–876.

Williams, D., Fournier, G.P., Lapierre, P., Swithers, K.S., Green, A.G., Andam, C.P., and Gogarten, J.P. (2011). A rooted net of life. Biol. Direct *6*, 45.

Williams, D., Gogarten, J.P., and Papke, R.T. (2012). Quantifying homologous replacement of loci between haloarchaeal species. Genome Biol Evol *4*, 1223–1244.

Wilson, E.B. (1905). THE CHROMOSOMES IN RELATION TO THE DETERMINATION OF SEX IN INSECTS. Science *22*, 500–502.

Woese, C.R., and Fox, G.E. (1977). Phylogenetic structure of the prokaryotic domain: the primary kingdoms. Proc Natl Acad Sci U S A *74*, 5088–5090.

Woese, C.R., Kandler, O., and Wheelis, M.L. (1990). Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. Proc Natl Acad Sci U S A *87*, 4576–4579.

Wolfe, K. (2000). Robustness--it's not where you think it is. Nat. Genet. *25*, 3–4.

Wolfe, K.H., and Shields, D.C. (1997). Molecular evidence for an ancient duplication of the entire yeast genome. Nature *387*, 708–713.

Wu, J., Kasif, S., and DeLisi, C. (2003). Identification of functional links between genes using phylogenetic profiles. Bioinformatics *19*, 1524–1530.

Wu, X., Liu, Q., and Jiang, R. (2009). Align human interactome with phenome to identify causative genes and networks underlying disease families. Bioinformatics *25*, 98–104.

Yamada, T., Kanehisa, M., and Goto, S. (2006). Extraction of phylogenetic network modules from the metabolic network. BMC Bioinformatics *7*, 130.

Yanai, I., Derti, A., and DeLisi, C. (2001). Genes linked by fusion events are generally of the same functional category: a systematic analysis of 30 microbial genomes. Proc. Natl. Acad. Sci. U.S.A. *98*, 7940–7945.

Yandell, M., and Ence, D. (2012). A beginner's guide to eukaryotic genome annotation. Nat. Rev. Genet. *13*, 329–342.

Yang, Z., and Rannala, B. (2012). Molecular phylogenetics: principles and practice. Nat. Rev. Genet. *13*, 303–314.

Yelton, A.P., Thomas, B.C., Simmons, S.L., Wilmes, P., Zemla, A., Thelen, M.P., Justice, N., and Banfield, J.F. (2011). A Semi-Quantitative, Synteny-Based Method to Improve Functional Predictions for Hypothetical and Poorly Annotated Bacterial and Archaeal Genes. PLoS Comput Biol *7*.

Yu, H., Luscombe, N.M., Lu, H.X., Zhu, X., Xia, Y., Han, J.-D.J., Bertin, N., Chung, S., Vidal, M., and Gerstein, M. (2004). Annotation Transfer Between Genomes: Protein–Protein Interologs and Protein–DNA Regulogs. Genome Res. *14*, 1107–1118.

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.

Zdobnov, E.M., von Mering, C., Letunic, I., and Bork, P. (2005). Consistency of genome-based methods in measuring Metazoan evolution. FEBS Lett. *579*, 3355–3361.

Zerbino, D.R., Achuthan, P., Akanni, W., Amode, M.R., Barrell, D., Bhai, J., Billis, K., Cummins, C., Gall, A., Girón, C.G., et al. (2018). Ensembl 2018. Nucleic Acids Res *46*, D754–D761.

Zmasek, C.M., and Eddy, S.R. (2002). RIO: analyzing proteomes by automated phylogenomics using resampled inference of orthologs. BMC Bioinformatics *3*, 14.

Annexes

Annexe 1: *OrthoInspector 3.0: open portal for comparative genomics*.

Annexe 2: *MyGeneFriends: A Social Network Linking Genes, Genetic Diseases, and Researchers*.

Annexe 1

OrthoInspector 3.0: open portal for comparative genomics.

Nucleic Acids Research, 2018 **1**
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

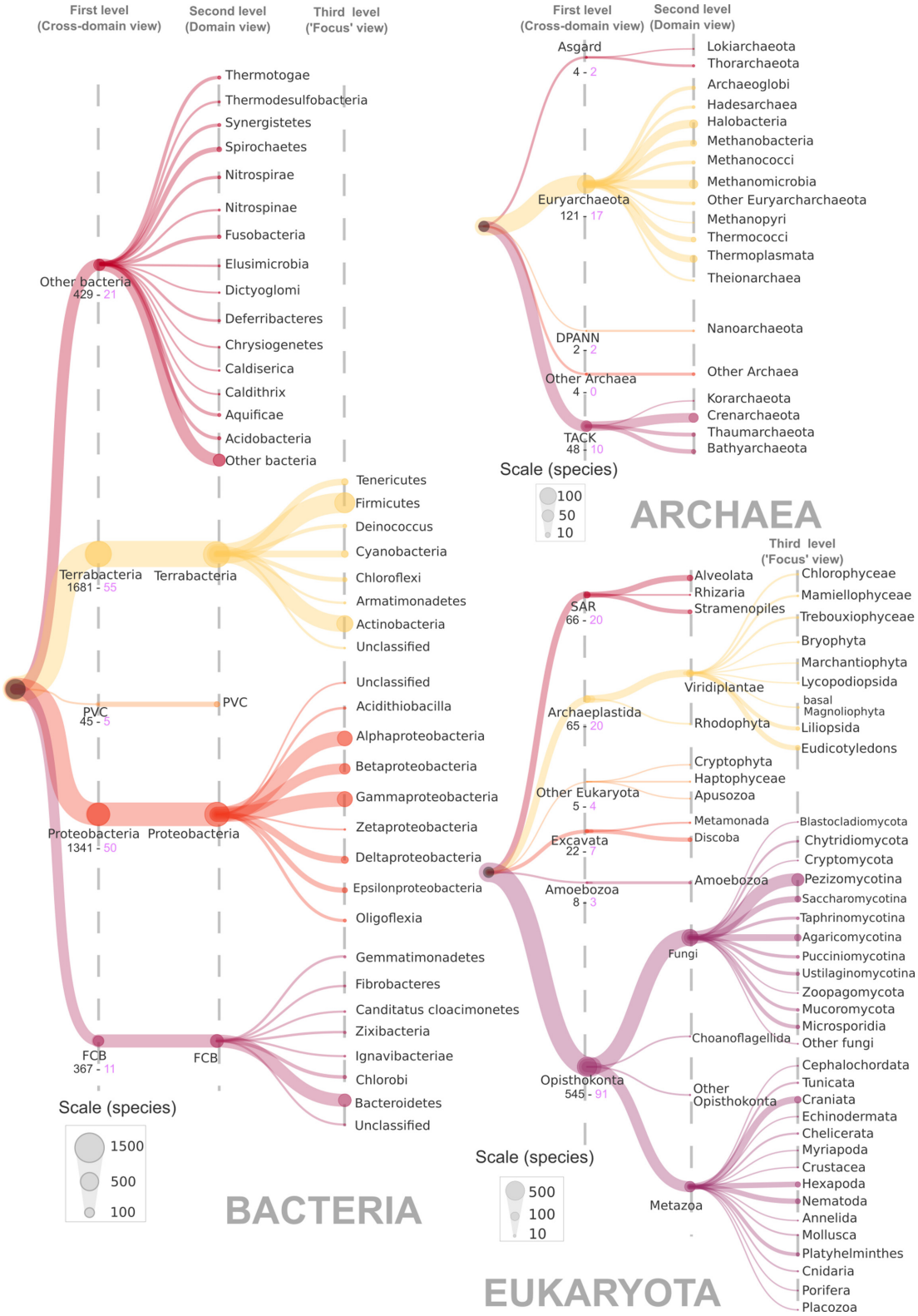


Figure 1. Taxonomic distribution of species represented in OrthoInspector. The domain trees are distributed on three 'levels'. The first level corresponds to the cross-domain taxonomic distribution heatmap shown when browsing the cross-domain database, the second level is shown on the heatmap for domain 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.

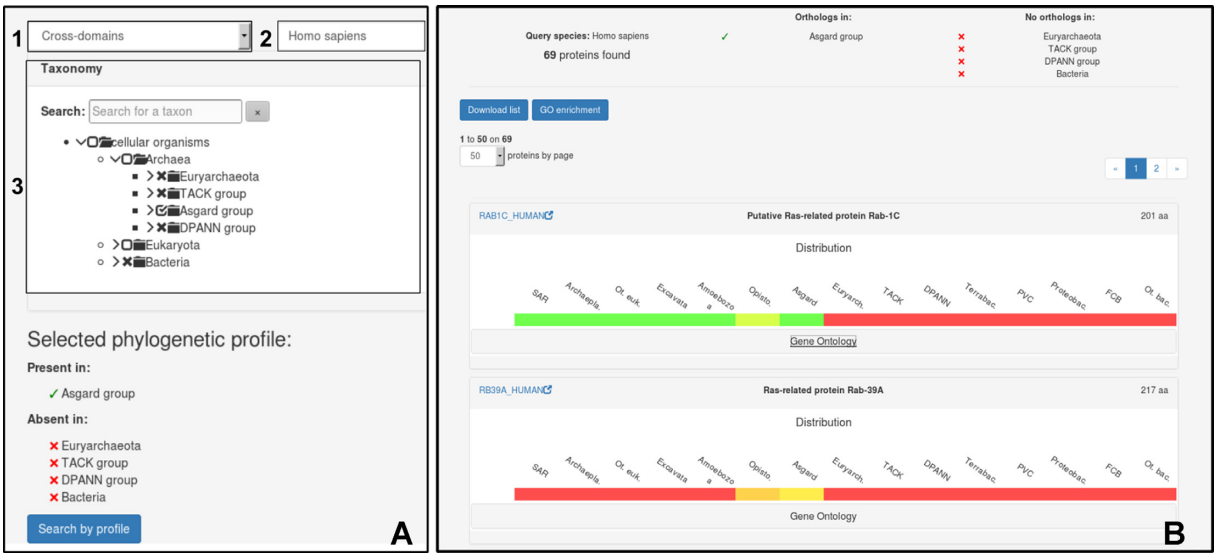


Figure 3. Phylogenetic profile search interface. (**A**) Definition of the phylogenetic profile. User selects: (1) the database, (2) the query species in the drop-down menu and (3) the presence/absence constraints using the phylogenetic tree. A summary of constraints is shown below the tree. Here, human 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 2

MyGeneFriends: A Social Network Linking Genes, Genetic Diseases, and Researchers

Original Paper

MyGeneFriends: A Social Network Linking Genes, Genetic Diseases, and Researchers

Alexis Allot, PhD; Kirsley Chennen, PhD; Yannis Nevers, MSc; Laetitia Poidevin, MSc; Arnaud Kress, MSc; Raymond Ripp, PhD; Julie Dawn Thompson, PhD; Olivier Poch, PhD; Odile Lecompte, PhD

ICUBE UMR 7357, Complex Systems and Translational Bioinformatics, Université de Strasbourg - CNRS - FMTS, Strasbourg, France

Corresponding Author:
Odile Lecompte, PhD
ICUBE UMR 7357
Complex Systems and Translational Bioinformatics
Université de Strasbourg - CNRS - FMTS
CSTB – ICUBE UMR7357
4 rue de Kirschleger
Strasbourg, 67085
France
Phone: 33 3 68 85 32 96
Fax: 33 3 68 85 35 18
Email: odile.lecompte@unistra.fr

Abstract

Background: The constant and massive increase of biological data offers unprecedented opportunities to decipher the function and evolution of genes and their roles in human diseases. However, the multiplicity of sources and flow of data mean that efficient access to useful information and knowledge production has become a major challenge. This challenge can be addressed by taking inspiration from Web 2.0 and particularly social networks, which are at the forefront of big data exploration and human-data interaction.

Objective: MyGeneFriends is a Web platform inspired by social networks, devoted to genetic disease analysis, and organized around three types of proactive agents: genes, humans, and genetic diseases. The aim of this study was to improve exploration and exploitation of biological, postgenomic era big data.

Methods: MyGeneFriends leverages conventions popularized by top social networks (Facebook, LinkedIn, etc), such as networks of friends, profile pages, friendship recommendations, affinity scores, news feeds, content recommendation, and data visualization.

Results: MyGeneFriends provides simple and intuitive interactions with data through evaluation and visualization of connections (friendships) between genes, humans, and diseases. The platform suggests new friends and publications and allows agents to follow the activity of their friends. It dynamically personalizes information depending on the user's specific interests and provides an efficient way to share information with collaborators. Furthermore, the user's behavior itself generates new information that constitutes an added value integrated in the network, which can be used to discover new connections between biological agents.

Conclusions: We have developed MyGeneFriends, a Web platform leveraging conventions from popular social networks to redefine the relationship between humans and biological big data and improve human processing of biomedical data. MyGeneFriends is available at lbgi.fr/mygenefriends.

(J Med Internet Res 2017;19(6):e212) doi:10.2196/jmir.6676

KEYWORDS

health care; social media; genetic variation; hereditary disease

Introduction

Social and Scientific Contexts

Web 2.0 and, particularly, social networks (Facebook, Google+, and LinkedIn), interconnect billions of users and manage terabytes of dynamic data flow [1]. They are at the forefront of the interactions and cooperation between humans and big data, and as such, they have established or popularized new conventions. A central concept in these innovations is the notion of an agent, representing an autonomous and active network member with various prerogatives. Notably, an agent can (1)

add new information, via micro-blogging for example; (2) spread information through the network via sharing [2]; (3) evaluate information with like, dislike, or vote reactions; (4) partition information using privacy settings; or (5) annotate information with comments. Agents play an active role in the evolution of the network structure by creating nodes (agent profile pages) and bidirectional (friendship) or unidirectional (follower) links between agents. They also partition agents into groups and connect agents to unstructured information (tagging). These actions are processed by specialized tools embedded in the network to create valuable feedback in the form of filtered and personalized information such as friendship suggestions, affinity scores between people, news feeds, targeted advertisements [3], merchandise suggestions [4], or real-time world observations [5].

The field of biology is evolving and adapting at a tremendous rate in response to the widespread use of high throughput methods and the rise of personal genomics [6]. For the end user of biological data, the paradigm shift initiated by the emergence of this big data [7] has led to important changes in the research landscape [8]. To keep up with the huge volumes of data and information, users need to easily and intuitively access, communicate, and network with useful information of personal interest. Therefore, data storage platforms and workflow infrastructures must evolve to integrate Web 2.0 and social network conventions.

Bioinformatics in the Web 2.0 Era

In this context, several major bioinformatics resources have introduced tools for personalized data flow management. The Online Mendelian Inheritance in Man (OMIM) [9] resource now proposes MIMmatch [10], a service allowing users to receive email notifications when entries for their favorite genes or diseases have changed. MyNCBI [11] retains user preferences to provide customized services for NCBI databases, whereas the Uniprot [12] website has been updated to allow users to select only categories of information they are interested in, to mask large-scale publications, and to use a basket to store proteins of interest. Similar efforts toward more efficient and personalized information management are also emerging in the exploitation of the increasing publication flow. Bibsonomy [13] allows a researcher to collect and manage publications and collaborate with colleagues, whereas PubChase and ReadCube recommend new publications depending on the content of an existing library. BioTextQuest+ [14] provides an interactive exploration platform for PubMed [15] and OMIM, and facilitates knowledge extraction by document clustering and bioentity recognition. GoPubMed [16] proposes pertinent publication searches by using background knowledge in the form of ontologies (gene ontology [GO], Medical Subject Headings [MeSH], etc) that take into account the user's keywords, but also synonyms and child concepts, whereas DeepQA4PA [17] returns GO concepts associated with publications related to a specific question. After identifying a gene or list of genes of interest, GeneMania [18] and GenesLikeMe [19] identify and score related genes that may also interest the user based on ontologies, disorders, compounds, phenotypes, expression levels, domains, sequences, and other data.

Important efforts have also been devoted to contextualizing entities by connecting them to their network. For instance, FACTA+ [20], Pubtator [21], or PAML-IST [22] return publications and their links to various biological entities such as compounds, drugs, enzymes, genes, diseases, symptoms, mutations, species, and others. EuropePMC [23] adds connections to GO and experimental factor ontology (EFO), and iHOP [24] highlights the most recent publications linked to a protein. Interaction with this complex data has been facilitated by the progressive democratization of visualization techniques. For instance, Javascript libraries like BioJS [25] provide reusable components for visualization of biological information (3D structures, phylogenetic trees, proteomes, pathways, and multiple sequence alignments), contributed by users and stored in a registry. Visualization techniques facilitate understanding of information updates, clarify links between entities and groups of entities, and highlight metadata information such as data sources, confidence estimates, and so on. For example, the ExAC browser [26] provides clear visualization of variations in a gene, the Semantic Body Browser [27] shows gene expression in a human and a mouse with a heat map on a schematized body, and NetGestalt [28] introduces 1-dimensional visualization of network modules to facilitate network comparisons.

Conversely, other tools aim to extract relationships between entities. For example, Chilibot [29] searches interaction (stimulation, inhibition, etc) or parallel (studied together, coexistence, homology, etc) relationships between user-submitted genes or proteins. EvexDB [30] extracts specific events: regulatory control, coregulation, or binding to a given gene.

Finally, some bioinformatics resources have introduced specific collaborative and social components, with wiki-inspired approaches like Proteopedia [31] or WikiGene [32], collaborative sequence annotations such as WebApollo [33], or voting for medical relevance and scientific evidence of variations with GeneTalk [34]. Recent initiatives such as Coremine or MAGI [35] combine these trends. Coremine allows exploration of various biomedical concepts and connections between them, addition of private or public comments, alerts on new articles or connections, and bookmarking. MAGI combines public and private cancer genomics datasets with sharing and collaborative annotation features as well as with interactive visualizations of variants, gene expression, and protein-protein interactions.

MyGeneFriends

Building on these advances, we have developed MyGeneFriends, a Web platform inspired by social networks, to redefine and enhance the relationship between humans and biological big data. By leveraging and combining conventions and practices arising from popular social networks, it provides more intuitive interactions with biological data and simplifies access to complex information by organizing it around three agents: genetic diseases, genes, and humans. This allows MyGeneFriends to be used not only by researchers and clinicians but also by the public, including empowered patients.

We focused on human genetic diseases (closely connected to genes and human users), as they represent major clinical

challenges and provide a simplified context to shed light on major common diseases. MyGeneFriends allows retrieval, management, contextualization, and annotation of information related to genes (expression, localization, and so on), genetic diseases (phenotypes, variations, and so on), and humans (interests, publications, and so on). The platform leverages user behavior and networking to personalize data visualization and the flood of information for each human user's needs, and allows project-oriented collaborations. Publication and friendship suggestions facilitate the identification of new relevant genes and diseases. Finally, we capitalize on the global social network to extract additional knowledge. MyGeneFriends was used during its development by members of our laboratory that provided continuous feedback. Additional feedback was collected from clinicians and researchers of the Medical Genetics Laboratory of Strasbourg and from colleagues from other laboratories that was particularly useful for improving the visualization of variations linked to a disease.

The aim of this paper was to introduce readers to MyGeneFriends and describe how practices from social networks can be applied to improve access to biological data.

Methods

Platform Architecture

The MyGeneFriends platform integrates multiple services (Multimedia Appendix 1) to extract and integrate large amounts of heterogeneous data. The data are stored and managed in a Postgres database, with a backup copy produced daily and stored on an external server. Elasticsearch [36] is used for powerful, complex, and fast plain text queries of publications and is synchronized daily with the MyGeneFriends database. The website is based on a stateless framework (Play framework) that includes many useful features such as error handling, build-in support for Json, WebServices, WebSockets, CoffeeScript, EBean object-relational mapper (ORM), localization, logging, and WebJars. The Play framework ensures easy horizontal scaling and scalability for increasing website traffic.

To execute local scripts and programs, a Web service has been developed using the Flask framework, which is called by MyGeneFriends using REST requests to run analysis or integration tasks. Data integration scripts are written in python, using peewee as the ORM.

Data Sources

Gene-related data including gene symbol, short description, type, and protein sequence are mainly obtained from the Ensembl [37] database. UCSC provides simple access to RefSeq [38] annotations for transcripts. To map gene identifiers between Ensembl and NCBI, we combine mappings performed by Ensembl and NCBI, together with gene symbol mapping, and extract one-to-one relationships. Gene expression data are obtained from the Human Genome Atlas microarray data [39] available in the gene expression omnibus (GEO) [40] database and validated using in-house statistical methods. In addition, relative signal intensities are calculated for heat map visualization using log signal intensities normalized in the range (0-1). Cellular localization of gene products is based on cellular component terms from GO [41]. Phylogenetic distributions for human genes and 100 eukaryotic species are retrieved from the OrthoInspector database [42] and used to categorize genes according to their evolutionary profile.

The relationships between genes and publications are defined using the gene2pubmed file from the NCBI. Publication abstracts are downloaded from Pubmed and integrated in the MyGeneFriends database. The python natural language toolkit (NLTK) [43] library is used to extract keywords from textual data linked to genes and diseases. It tokenizes the text into phrases and words, stems words in order to retrieve a canonical form, and filters words on the basis of the NCBI list of stop words (words that occur frequently in texts but are not informative) and in-house filters for word size, numbers, and special characters. Then, we take advantage of the gensim [44] library to calculate the Inverse Document Frequency (IDF) of the keywords and the TF*IDF (Term Frequency * Inverse Document Frequency). The IDF is used as a specificity score, and the TF*IDF is used to weight the relationship between a keyword and a gene or disease.

The main disease-related data are obtained from OMIM and Orphanet. In order to take into account differences in disease definitions from different data sources and propose a unified view of the current disease knowledge, an integration process was developed with two simple rules (Multimedia Appendix 2). After integration, diseases are linked to phenotypes using human phenotype ontology (HPO) [45] data files (hp.obo and phenotype_annotations.tab) containing phenotypes and phenotype-disease relationships. Variations and variation-disease relationships are extracted from the curated set provided by ClinVar [46] in the variant call format (VCF) file (limited to records with an rs# identifier). Each line is parsed and a variant entry is integrated into MyGeneFriends as a couple of genomic position and allele, allowing precise definition of the relationships between diseases and mutations. Variant effect predictor (VEP) [47] is used to link variations retrieved from ClinVar to Ensembl transcripts and to estimate their effect. The effects are then automatically classified into more general categories using the sequence ontology [48] data.

Data Flow Management

The data flow management involves the integration of data from diverse sources (databases, FTP servers, and local files) into the MyGeneFriends database. After cleaning and parsing mined data, additional analyses are automatically processed, such as keyword extraction from biological text or generation of links between variants and transcripts (mentioned previously). Then, MyGeneFriends compares remote and local data to generate news events. One or more fields from each item is used as a unique identifier. If a remote item has an identifier (one or several selected fields from an item) that is absent from the local database, it is considered to be a "new" event. If a local item has an identifier that is not present in the remote source, it is considered to be a "delete" event. If the identifier is present in both remote and local sources, the items are compared field by field to generate "update" events. Once these events are

generated, the local database is synchronized with the remote source.

Finally, the way the news item is presented to the user depends on the biological context of the considered element. When an agent is linked to a publication that is not available in the MyGeneFriends database, the publication is downloaded and made accessible directly from the news panel. When a sequence is updated, a sequence alignment is generated using ClustalW [49]. When a textual information changes, such as the description of a disease, the google-diff [50] python library is used to compare both versions of the text and highlight the differences.

Data Display as Word Clouds

Word cloud representations are used in the visualization panel of an agent to display the cellular localization of the protein encoded by a gene and the phenotypes associated with a disease. Specific terms are considered as more informative and emphasized in the word cloud. The specificity of a term (cellular component in GO and phenotype in HPO) describing an agent is estimated using the information content (IC) metric [51]. The IC is defined as the negative natural logarithm of the probability of a term t:

IC (t)=−logP (t),

where P (t) is based on the frequency of the term in the considered ontology.

The specificity is then defined as the IC normalized in the range (0-1), where 0 corresponds to the minimal font size and 1 to the maximal font size during word cloud rendering.

Friendships

The MyGeneFriends network is based on friendships between agents. Human friendships are defined by users, whereas gene-gene, disease-disease, and gene-disease links are automatically built from external sources (search tool for recurring instances of neighboring genes, STRING [52]; and HPO) or inferred from the MyGeneFriends network (Error: Reference source not found). STRING global scores (higher than 0.7, corresponding to “high confidence” in STRING) are used as a metric of friendship between genes based on protein-protein interaction data. Causative genes mined by HPO from OMIM and Orphanet are exploited to link genes and diseases.

In addition to these external sources, MyGeneFriends establish links based on common properties. Diseases sharing phenotypes are related to each other with a score defined as the sum of specificity scores of phenotypes common to both diseases, divided by the sum of specificity scores of all phenotypes related to both diseases. Similarly, genes sharing GO [41] terms are connected according to two different metrics. The first metric (“GO simple”) is based on the number of shared GO terms between 2 genes, whereas the second corresponds to the functional semantic similarity (FSS) [51]. Genes and diseases related to the same variant(s) are also linked. Moreover, genes are evolutionarily linked when applicable, on the basis of the Jaccard distance calculated between in-house phylogenetic profiles produced by OrthoInspector [42].

Finally, nonhuman agents can become friends based on social connections emerging from the network itself: genes sharing human or disease friends are connected, as well as diseases with common human or gene friends.

Suggestions and Affinity Score

To suggest new gene or disease friends or new publications to a user, MyGeneFriends relies on the content of the user’s active Topic. For friendship suggestions, each nonhuman agent from MyGeneFriends (a_c) is scored relative to the user’s active Topic and the top 10 candidates are suggested as new friends. The score of an agent (gene or disease) given a Topic is the sum of scores (S) between this agent and all agents of the same type in the active Topic (a_t):

$score(a_c) = \Sigma^{N}_{t=0} S(a_t, a_c)$To score genes, we use the global STRING score, whereas the score between two diseases d1 and d2 is calculated using the Information Content (IC) of the related phenotypes (P) as:

$Score(d1, d2) = (\Sigma\, IC\,(P \in d1 \cap d2))/\,(\Sigma\, IC\,(P \in d1 \cup d2))$

In addition, we provide an affinity score (a_{aff}) reflecting the proximity between an agent and the content of the user Topic and thus, the relevance of befriending this agent. It is displayed on the gene and disease profile pages when the agent can be related to the content of the Topic. The affinity score is defined as:

$a_{aff} = a_c$ /max ac X 100

To suggest pertinent publications, MyGeneFriends uses keywords associated with the active Topic. These keywords have been either added manually by the user or automatically inferred (see formula below). The keywords are weighted and used to query the Elasticsearch server to retrieve pertinent publications. For Elasticsearch, weights between 0 and 1 reduce the relevance of a term, and weights higher than 1 increase it. Therefore, the weight for each keyword given the content of the Topic (k_t) is defined as:

$weight(k_t) = 1 + \Sigma^{N}_{t=0}\, ka_t + k_h$,

where ka_t is the score describing the relationship between the keyword and an agent from the Topic, and k_h is a factor applied if the user has explicitly added this keyword to the Topic.

Results

Overview of the Platform

MyGeneFriends is a new social network leveraging conventions from Web 2.0 and interconnecting three kinds of autonomous and active agents: human genes, humans, and genetic diseases. All genetic disorders including malformations, groups of phenotypes, etc are included in the network, as well as all types of human genes (coding and noncoding) in agreement with the growing evidence concerning the importance of noncoding genes in biological processes and diseases [53-55].

All agent-related data is accessible via standardized profile pages. Daily data mining and integration processes have been

developed to maintain the "nonhuman" agents (more than 63,000 human genes and 14,000 genetic diseases) up to date and generate a news flow (more than 1 million news items were created in the last year) by exploiting public (Ensembl [37], NCBI, Uniprot [12], HPO [45], OMIM [10], Orphanet [56], OrthoInspector [42], etc) and in-house data resources. All data retrieved or processed by MyGeneFriends and related to genes and diseases are "public," whereas data submitted by humans are "private" (visible only by the owner) by default, unless the human decides to make it "protected" (visible by owner and selected collaborators) or "public" (visible to anyone).

The MyGeneFriends network arises from several millions of connections (called "friendships") between agents, resulting from automated dynamic data mining processes combined with human actions (Figure 1). Assessment of gene-gene, gene-disease, and disease-disease connections (nonhuman friendships) are based on automated mining of bibliographic, evolutionary, functional, phenotypic, or social data. Human friendships with genes, diseases, and other humans are defined by the user through gene targeting, definition of research interests (Topics), or user targeting (groups). Human friendships with genes or diseases can be private, protected, or public, although they are public by default to encourage networking. This data privacy management [57] is crucial to keep essential data private, while being open enough to "attract" new information and collaborators.

By exploiting human actions, MyGeneFriends can automatically (1) personalize information and visualization by highlighting and filtering pertinent data, (2) suggest new publications and friends (gene or disease), and (3) provide subnetworks for collaborations on defined research interests.

Agent Profiles

Each agent in MyGeneFriends has a profile (Figure 2) that provides a unified architecture and organization to ensure intuitive navigation through the network and access to relevant and personalized information about agents. These profiles contain 4 major sections: "header," "basic information," "friends," and "news."

Figure 1. Ontology of friendships between agents in MyGeneFriends. Agents are linked by numerous friendships (corresponding to green boxes) of different kinds (blue boxes). First, we separate decision-driven friendships (agent actions) from naturally occurring friendships (mined). Then, we split natural friendships into those due to direct contact between agents, and those influenced by an external factor. This external factor mimics the human tendency of befriending people with the same interests (represented here as phenotypes, annotations, variants, and phylogenetic distributions) or common friends (genes, humans, and diseases).

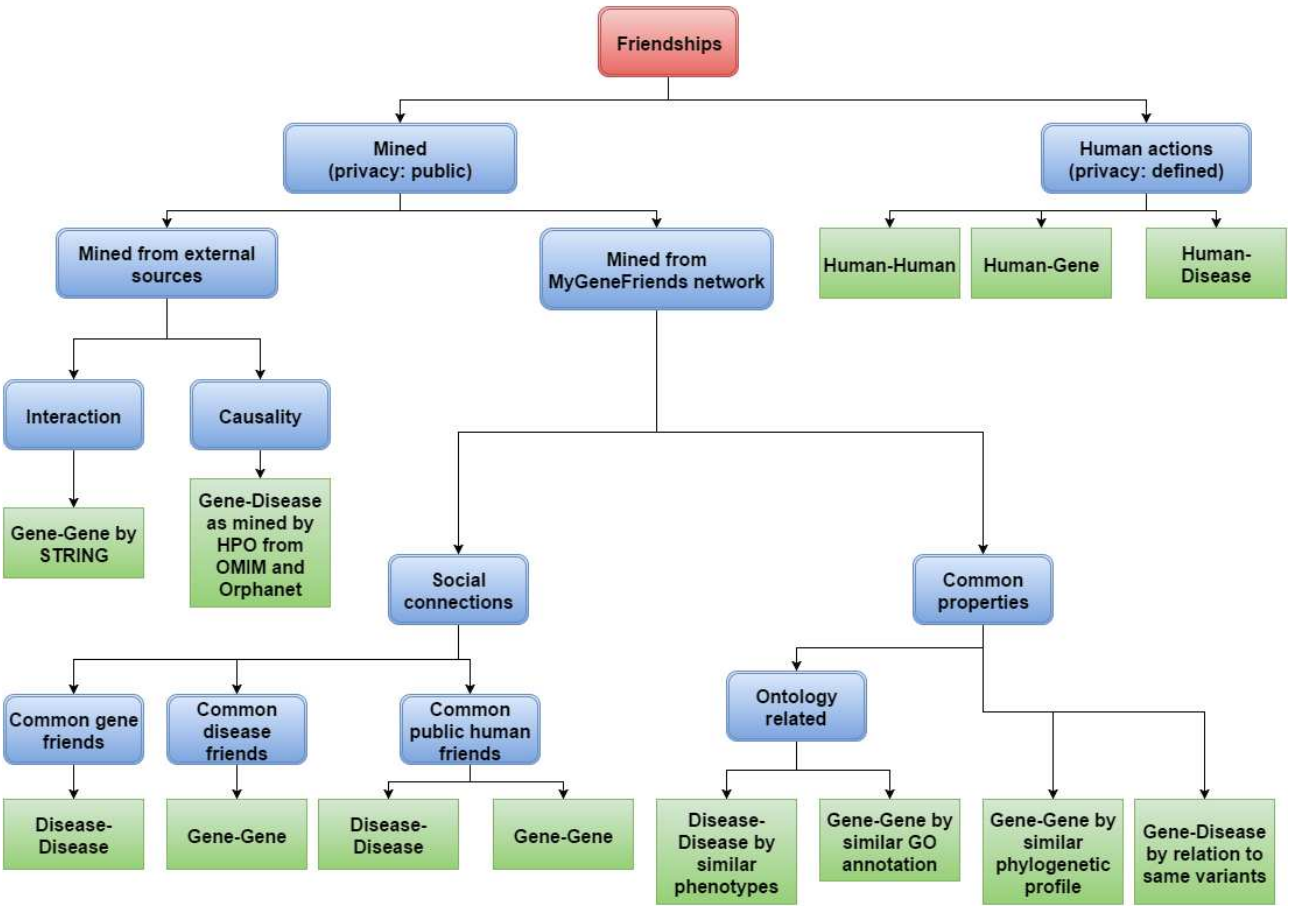
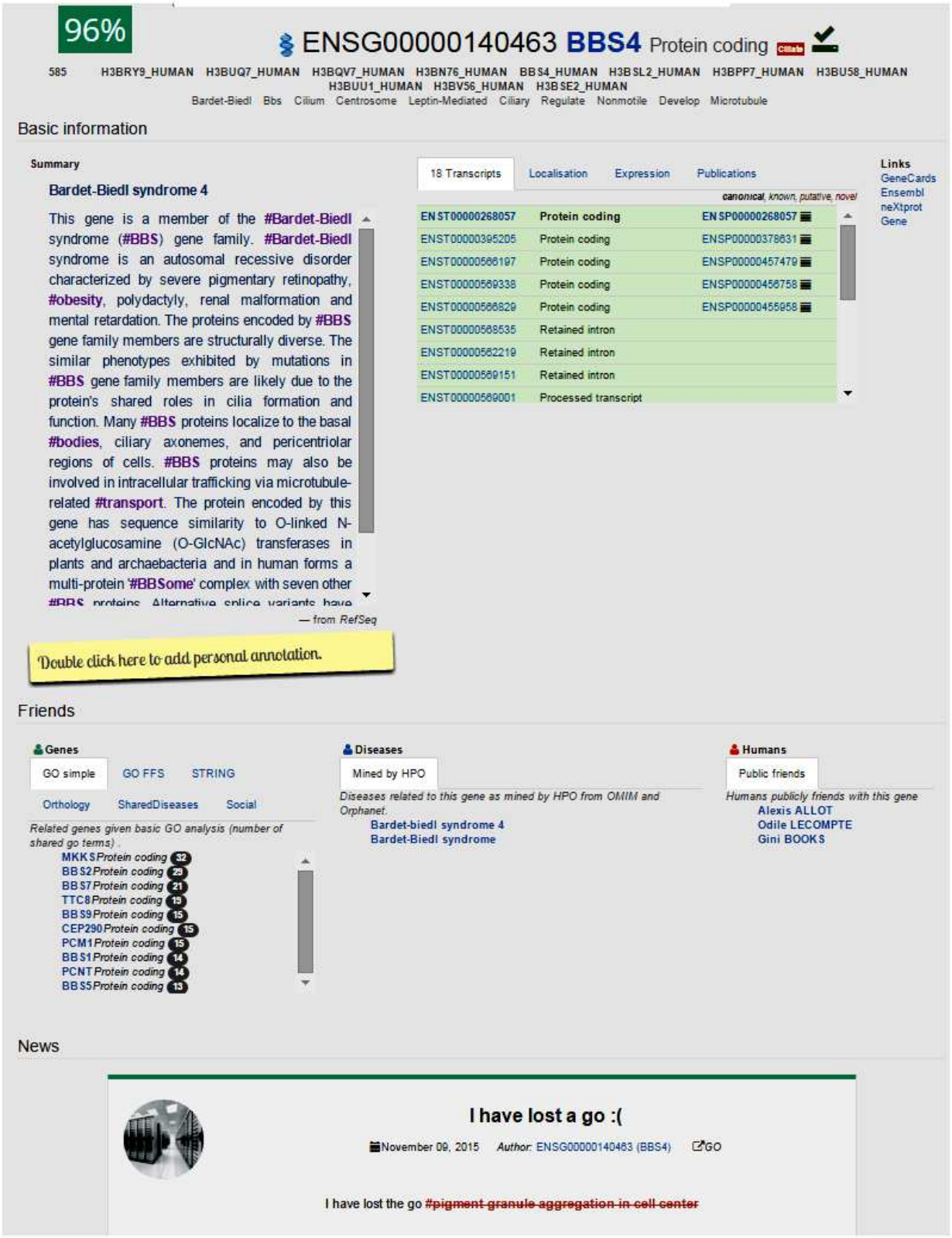


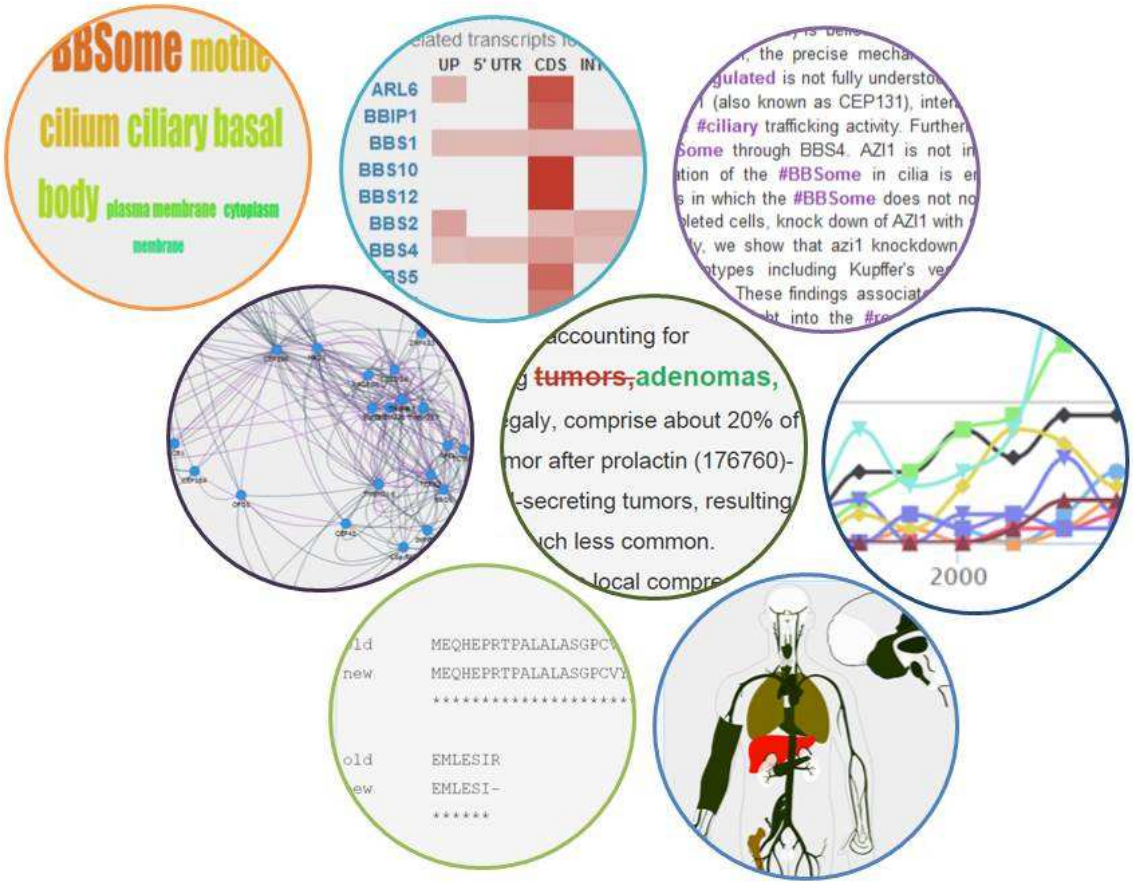
Figure 2. Representative profile page of a MyGeneFriends agent (here the gene BBS4). Four distinct sections are shown. The "header" section briefly introduces the agent, displays a list of synonyms, and allows friendship management. It shows the affinity score (here 96%) estimating how interesting this agent could be for the user. The "basic Information" section shows more detailed information about the agent: a description, different visualizations describing the agent, links to external sources, and a personal annotation from the user. The "friends" section allows navigation through the "friends of friends" network by displaying public friends of the agent, grouped according to their type. Finally, the "news" section displays all the news related to the agent.



The top agent-related keywords are displayed in the header to briefly introduce the agent, whereas the summary in the "basic information" section provides a more detailed description. Humans can expand the official description of a nonhuman agent by adding personal annotations or unpublished results that can then be accessed at any time and shared with collaborators (Figure 2). Exploration panels give access to the most important information using visualization techniques (see Figure 3) to highlight specific information for genes and diseases as described below. The "friends" section of the profile displays links to public friends (genes, diseases, or humans) of the agent, allowing further networking with potentially interesting agents. Finally, the news feed is an intuitive way to track changes in information related to an agent.

To personalize the profile view, the keywords inferred to be important for the viewer are highlighted in the agent description. For example, if the user is friends with cilia-related genes, the word cilium is highlighted in the description of the other agents (human, gene, or disease). Moreover, if a nonhuman agent is related to the user's current collaborators, an affinity score is shown, inviting the user to befriend this agent. If the agent already collaborates with the user, the score reflects how close it is to other collaborators.

Figure 3. MyGeneFriends uses various visualization techniques to optimize the display of biological information: (1) word clouds highlight the most specific ontology terms, (2) barcodes offer a synaptic and interactive view of the density of variations related to regions of a gene or effect on a protein, (3) highlighting words in text identifies the most pertinent paragraphs for a given human user, (4) networks of friends help to understand the connections between agents and identify groups of highly connected agents, (5) colors highlight modifications in textual information related to agents, (6) timelines show the evolution of the popularity of gene collaborators in a Topic, (7) pairwise alignments identify the differences between two versions of protein sequences, and (8) heat maps on schemas of the human body, brain, and fetus allow easy analysis of the expression pattern of a given gene.



Gene Profile

Gene profiles use RefSeq summaries to describe agents and connect these agents to external resources via links to Ensembl, GeneCards, NCBI, and neXtprot websites.

Exploration panels display the most important aspects of the gene. The first panel presents gene transcripts with their properties: sequence, type (protein coding, miRNA, etc), reliability (known, putative, and novel), and corresponding protein sequence, if any. The second panel shows the subcellular localization(s) of the encoded proteins, defined by the GO cellular component ontology, as a word cloud (Figure 3). The third panel shows the gene expression for protein coding genes as a heat map in more than 40 tissues, through an interactive schematic view of the human body (male and female), brain, and foetus (Figure 3). Pan and zoom capabilities (jquery.panzoom.js) allow users to navigate through the schematic view and visualize even the smallest tissues. Additional information such as tissue description and probe set signal intensities are available. In addition to the visualization

of gene expression, the "Expression filter" tool allows users to find genes of interest based on their expression or absence of expression in a defined set of tissues. Publications associated with the gene are displayed with their abstract and can be liked, disliked, or marked as valuable. The number of all genes related to the publication, as well as the count of likes and dislikes, help to estimate the relevance of the publication for the considered gene. Moreover, genes related to a publication can be visualized as an interactive graph, allowing further networking and identification of additional genes of interest.

Genetic Disease Profile

Diseases are extracted from the OMIM (all entries except explicit genes) and Orphanet (all entries, including groups of phenotypes) databases. The preference of exhaustivity over specificity is motivated by the inherent difficulty in defining a disease. We use expert created links between Orphanet and OMIM entries (displayed on Orphanet entries) as the main data source to merge diseases. When a disease is not linked to any other, or when a clear one-to-one mapping can be made between an Orphanet and an OMIM entry, the entries from both databases are fused into a single one (see Multimedia Appendix 2). Once this process is complete, we use the remaining one-to-many connections (eg, one entry for "Bardet-Biedl syndrome" in Orphanet corresponds to multiple entries in OMIM for each Bardet Biedl syndrome subtype) to create groups of highly connected diseases, which we call "metadiseases."

Two main features have been selected to characterize a disease on the disease profile panel: (1) variations explaining the causes of a disease, and (2) phenotypes describing its consequences. Phenotypes are represented by a word cloud highlighting rare HPO phenotypes associated with the disease. The description of variants is generated by the integration of more than 100,000 ClinVar [46] curated variations (single-nucleotide variants and small insertions and deletions) directly linked to diseases.

As the effects of the variants can differ per considered transcript, MyGeneFriends uses the Ensembl VEP [47] script to create more than a million links between variants and Ensembl transcripts stored in the MyGeneFriends database. To describe the complex relationships between variants, transcripts, and disease-causing genes, we have developed three synoptic and interactive views with variants grouped per affected gene. With this synthetic barcode representation (Figure 3), the human user has a rapid overview of the characteristics of the known variants associated with the disease and can easily identify variants exhibiting specific features, for instance, synonymous variants affecting a splicing region. The third view focuses on variants differentially affecting protein coding transcripts (Multimedia Appendix 3). Such variants can generate a mix of affected and unaffected proteins depending on the tissue or developmental stage and often result in puzzling phenotypes.

Metadiseases have special profile pages on MyGeneFriends, summarizing the main properties of nested diseases, displaying nested diseases as a network, and highlighting the most representative gene friends and phenotypes of the concerned diseases. To date, MyGeneFriends has information on 725 metadiseases, representing 3418 diseases.

Human Profile

The third agent in MyGeneFriends is the human user, who must register on the website (registration is free, and a demo account is available for testing purposes). The user's profile page contains information provided by the owner: his affiliation, geographic localization, a list of authored publications, and a short description. Even if no description is provided, MyGeneFriends introduces the human to other users by automatically extracting best scored keywords associated with public gene and disease friends of the human and displaying them on his profile.

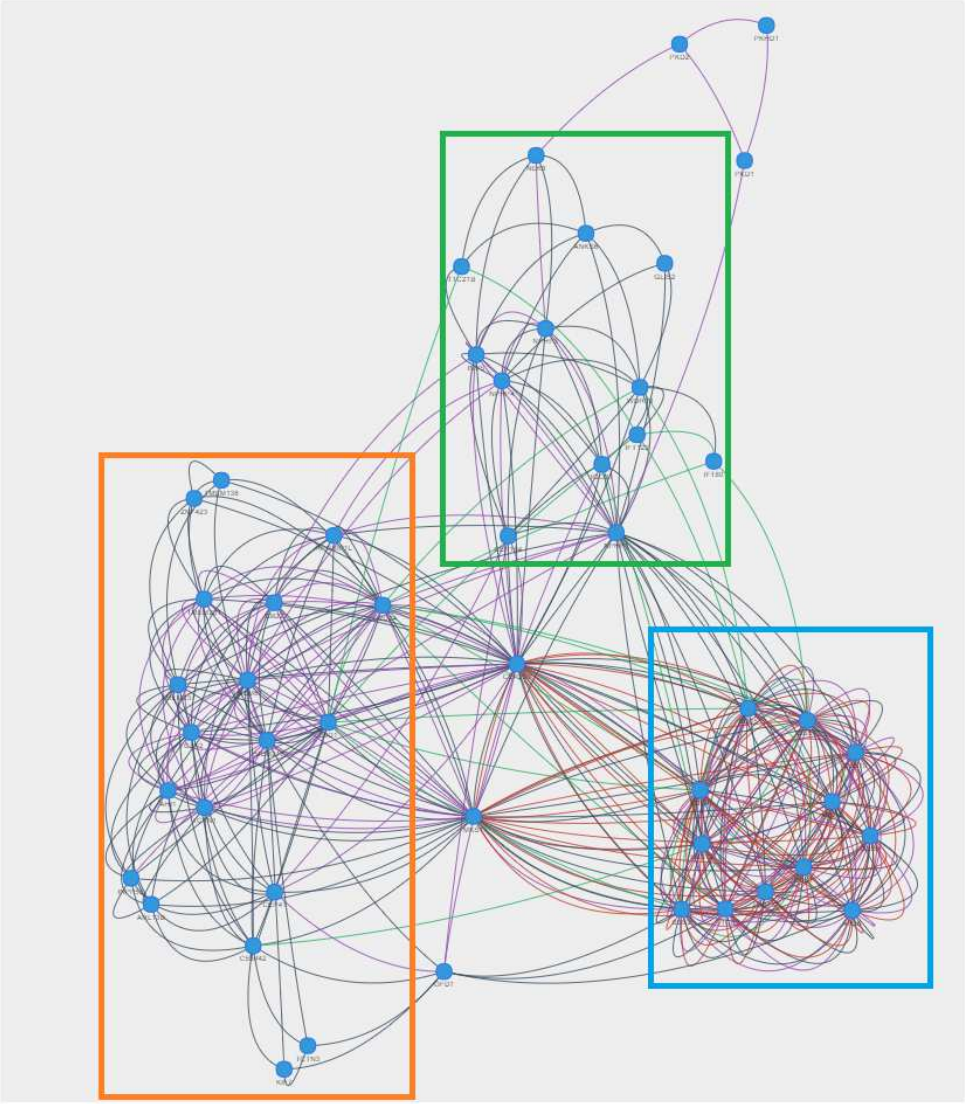
The private view of the profile page allows humans to create and manage groups of collaborators related to research projects, called Topics. All Topics owned by the user are shown in the "My Topics" section. The Topic selected as "active" is used for personalization and suggestion processes. A second section called "My collaborations" allows the user to monitor Topics from the other users with whom he collaborates.

Friendships and Networking

Friendships are an essential concept in MyGeneFriends, since on the one hand, they allow networking through friends and evaluation of the relatedness of 2 agents, and on the other hand, they are used to suggest interesting agents as new friends. Some friendships are automatically generated based on data mining, whereas others result from human activity. Friendships offer different and complementary points of view on the close environment of an agent in terms of protein interactions, function and localization, implication in research projects or diseases, and many others (Figure 1).

Exploitation of the friendship network in MyGeneFriends leverages mined and user-created connections to discover highly connected clusters. Interactive graph views with repulsion physics (using the vis.js library) allow intuitive visualization of friendships within a group of agents (genes from a publication, diseases from a metadisease, or agents associated with a Topic), leading to selection and observation of different types of friendships (common friends, common features, cooccurrence, and so on). Highly connected agents will naturally form subgroups corresponding to biologically relevant categories as exemplified by the Congenital Hepatic Fibrosis [58] gene network (Figure 4).

Figure 4. Dynamic network visualization of relationships between actors. Network of 52 genes related to Congenital Hepatic Fibrosis (CHF), a developmental disorder most frequently associated with ciliopathies. Red links represent shared public human friends, grey links represent shared diseases, violet links represent STRING relationships, and green links represent similar evolutionary profiles. Each link type can be removed or added to the network in real time. Moreover, in the dynamic network view provided by MyGeneFriends, highly connected genes are clustered automatically to form subgroups. In this example, 3 main subnetworks (highlighted by rectangles) emerge corresponding to genes associated with a distinct ciliopathy: Bardet-Biedl Syndrome (blue rectangle), Joubert and Meckel syndromes (orange rectangle), and Senior-Loken syndrome and nephronophthisis (green rectangle).



Topic: Interactive Collaborative Unit

On their profile pages, users can create groups of agents (called Topics). Each group centralizes information around a research project and links to agents collaborating with it (Multimedia Appendix 4), thus presenting a subjective view of biological information from a given research perspective.

This allows MyGeneFriends to display a personalized news feed, providing a technological watch of bibliographical and public database updates related to gene and disease friends that collaborate in the user's Topics. News items include various subjects such as new or lost friendships between diseases and genes, updated symbols, synonyms, descriptions, new or lost GO or HPO annotations, protein sequence updates, or presence in a new publication.

Several tools are provided for the analysis of Topic related agents. The network visualization facilitates the evaluation of the heterogeneity of the Topic's content (Multimedia Appendix 5), and the identification of highly linked subgroups of agents and relationships between these groups. The timeline visualization (Figure 3) places the Topic in a global research perspective, presenting the annual evolution of the number of publications associated with the genes in a Topic.

Finally, in addition to serving as a basis for friendships and publications suggestions (see Methods), information mined from Topics allows the enhancement of the reading experience of an agent's descriptions and publication abstracts by automatically highlighting the keywords most representative of the user's interests (Multimedia Appendix 6).

Discussion

Principal Findings

By leveraging conventions and practices used in popular social networks, MyGeneFriends aims to challenge the way we interact with data by providing a first step toward a system where biological entities such as genes and genetic diseases are no longer passive concepts, but are instead proactive agents of the research process, helping and collaborating with human counterparts.

In mainstream social networks, humans can create a representation of themselves in the form of a profile, then interact with the network by writing posts, adding commentaries or likes, making new friends, and sharing and spreading information. To transpose this concept to MyGeneFriends, we had to create a network that could reflect current research efforts in genetics and medicine. To populate the network, we focused on human genetic diseases because of their broad interest, and their more direct links to genes and genomic variations compared with infectious diseases or cancer. With humans and human genetic diseases selected, the choice of the third agent was obvious as many publications and bioinformatics resources structure their information in a gene-centric manner. To interconnect the network, we adopted two of the main characteristics of real-world friendships: commonality (common friends, qualities, and interactions), and group membership (family, coworkers, and hobbies).

Compared with existing Web services, MyGeneFriends can (1) leverage user behavior to provide personalized profiles and news feeds, given each user's specific research interests; and (2) consider user behavior as valid biological information integrated in the biological data network to be mined and influencing the discovery of connections between genes and diseases.

Conclusions

The development of MyGeneFriends lies at the frontier between bioinformatics and the emerging science of human-data interaction, and in the future, we plan to extend the functionalities in both areas. First, genes from other model species (mouse, zebrafish, etc) will be added and connected by friendship links based on orthology. Additional friendships will be incorporated to provide a regulatory context such as friendships based on transcription factors or miRNA. Second, we believe that while humans remain special agents in this first version of MyGeneFriends, in the future the three agents will interact on the same level, with more independent and proactive genes and diseases. Research will be facilitated by better communication between different agents, with each agent able to produce and transmit new, relevant data and knowledge. A gene could, for example, find itself linked to a new disease or ask to be sequenced by his friend, the sequencer. With this increased autonomy of nonhuman agents and an independent flow of information, the role of the human in the network must clearly evolve. This evolution can be viewed either as a danger or as a source of new collaborations and opportunities.

Acknowledgments

This work was funded by ANR Investissement d'Avenir Bioinformatique BIP: BIP (ANR10-BINF03-05).

The authors are thankful to Alexia Rohmer for her help with the interactive schematic view of the human body, and are also grateful to Dr. Ioannis Xenarios (Swiss Institute of Bioinformatics), Dr. Frédéric Chalmel (Université de Rennes 1), Dr. Laurent Vallar (Luxembourg Institute of Health), and Prof. Pierre Gançarski (University of Strasbourg) for helpful discussions.

Authors' Contributions

AA developed MyGeneFriends with contributions from KC and YN, and wrote the manuscript. AK and RR managed hardware infrastructure. JL wrote the manuscript. OP and OL supervised the project and wrote the manuscript. All authors read and approved the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1

MyGeneFriends relies on multiple servers, scripts, and databases. Jenkins is used to periodically execute python integration scripts, which maintain the MyGeneFriends database up to date. A backup copy of the database is created daily, and the publications table is synchronized daily with an Elasticsearch server. The MyGeneFriends website is built using the Play framework and calls an API server built with Flask to execute command line programs and download publications into the local database. Bug reports are sent to the YouTrack server.

[PPTX File, 103KB - jmir_v19i6e212_app1.pptx]

Multimedia Appendix 2

Diseases from different, contradictory data sources are merged using two rules. First, disease A is merged with disease B if and only if B has a single link to disease A, and no other diseases have a single link to A. Second, if a disease is linked to several other diseases, a disease group called Metadisease is created.

[PPTX File, 274KB - jmir_v19i6e212_app2.pptx]

Multimedia Appendix 3

The "differential effect on proteins" view for variants on the "Bardet-biedl syndrome 6" disease profile (a) shows that while variants (here rs74315398, rs28937875, rs587777827, rs74315399, rs74315397, rs74315396) linked to this disease affect the coding DNA sequence (CDS) of 2 transcripts of the gene MKKS, one protein coding transcript (green rectangle on Ensembl Genome Browser [b], and 1 of 3 labels in the view [a]) is not affected in the CDS.

[PPTX File, 71KB - jmir_v19i6e212_app3.pptx]

Multimedia Appendix 4

A Topic groups genes, diseases, and humans collaborating on a shared research interest. When a human becomes friends with a new gene or disease, it is added to the active Topic. Human collaborators see all protected friends and annotations related to the Topic.

[PPTX File, 113KB - jmir_v19i6e212_app4.pptx]

Multimedia Appendix 5

Network visualization of agents related to a Topic (here a set of Bardet-Biedl syndrome related genes on the left and a set of muscular fiber related genes on the right) help to understand how many potentially different research interests a Topic contains. The network displays highly connected agents automatically grouped together. Purple links represent STRING based relationships, green links are based on evolutionary profile similarity, red links indicate shared public human friends, and grey links shared disease friends.

[PPTX File, 61KB - jmir_v19i6e212_app5.pptx]

Multimedia Appendix 6

Keywords most related to agents from active Topics are highlighted in (1) publication abstracts, and (2) descriptions on agent profiles. This helps to quickly identify paragraphs that may interest the reader.

[PPTX File, 146KB - jmir_v19i6e212_app6.pptx]

References

1. Allot A, Lecompte O. LBGI. MyGeneFriends website URL: http://lbgi.fr/mygenefriends [accessed 2017-06-01] [WebCite Cache ID 6qtiI2y1c]
2. Bakshy E, Rosenn I, Marlow C, Adamic L. The role of social networks in information diffusion. 2012 Apr 16 Presented at: Proceedings of the 21st International Conference on World Wide Web; 2012; Lyon.
3. King SP, Burgener C, Paretti CT, Davis ME. Google Patents. 2010. System and method for contextual advertising based on status messages URL: https://docs.google.com/viewer?url=patentimages.storage.googleapis.com/pdfs/US20100228582.pdf [WebCite Cache ID 6kdH8PZdj]
4. Ma H, Zhou D, Liu C, Lyu M, King I. Recommender systems with social regularization. 2011 Presented at: Proceedings of the Fourth ACM International Conference on Web Search and Data Mining; 2011; Hong Kong p. 287-296. [doi: 10.1145/1935826.1935877]
5. Zhao S, Zhong L, Wickramasuriya J. Human as real-time sensors of social and physical events: a case study of twitter and sports games. arXiv 2011:1106-4300.
6. Hendlisz A. Of art and science: is personalized medicine getting personal enough? Curr Opin Oncol 2015 Jul;27(4):349-350. [doi: 10.1097/CCO.0000000000000205] [Medline: 26049276]
7. Hey T. The fourth paradigm – data-intensive scientific discovery. In: Kurbanoğlu S, Al U, Erdoğan PL, Tonta Y, Uçak N, editors. E-Science and Information Management. IMCW 2012. Communications in Computer and Information Science. Berlin, Heidelberg: Springer; 2012.
8. Kitchin R. Big Data, new epistemologies and paradigm shifts. Big Data Soc 2014 Jul 10;1(1):1-12. [doi: 10.1177/2053951714528481]

9. Hamosh A, Scott AF, Amberger J, Valle D, McKusick VA. Online mendelian inheritance in man (OMIM). Hum Mutat 2000;15(1):57-61. [doi: 10.1002/(SICI)1098-1004(200001)15:1<57::AID-HUMU12>3.0.CO;2-G] [Medline: 10612823]
10. Amberger JS, Bocchini CA, Schiettecatte F, Scott AF, Hamosh A. OMIM.org: online mendelian inheritance in man (OMIM(R)), an online catalog of human genes and genetic disorders. Nucleic Acids Res 2015 Jan;43(Database issue):D789-D798 [FREE Full text] [doi: 10.1093/nar/gku1205] [Medline: 25428349]
11. Giglia E. PubMed in progress: latest changes in MeSH and MyNCBI. Eur J Phys Rehabil Med 2011 Sep;47(3):525-528 [FREE Full text] [Medline: 21946409]
12. UniProt C. UniProt: a hub for protein information. Nucleic Acids Res 2015 Jan;43(Database issue):D204-D212 [FREE Full text] [doi: 10.1093/nar/gku989] [Medline: 25348405]
13. Benz D, Eisterlehner F, Hotho A, Jaschke R, Krause B, Stumme G. Managing publications and bookmarks with BibSonomy. 2009 Presented at: 20th Acm Conference on Hypertext and Hypermedia (Hypertext 2009); 2009; Torino p. 323-324. [doi: 10.1145/1557914.1557969]
14. Papanikolaou N, Pavlopoulos GA, Pafilis E, Theodosiou T, Schneider R, Satagopam VP, et al. BioTextQuest(+): a knowledge integration platform for literature mining and concept discovery. Bioinformatics 2014 Nov 15;30(22):3249-3256 [FREE Full text] [doi: 10.1093/bioinformatics/btu524] [Medline: 25100685]
15. Giglia E, Spinelli O. PubMed reloaded: new interface, enhanced discovery. Eur J Phys Rehabil Med 2009 Dec;45(4):631-636 [FREE Full text] [Medline: 20032922]
16. Doms A, Schroeder M. GoPubMed: exploring PubMed with the gene ontology. Nucleic Acids Res 2005 Jul 1;33(Web Server issue):W783-W786 [FREE Full text] [doi: 10.1093/nar/gki470] [Medline: 15980585]
17. Gobeill J, Gaudinat A, Pasche E, Vishnyakova D, Gaudet P, Bairoch A, et al. Deep question answering for protein annotation. Database (Oxford) 2015;2015:bav081 [FREE Full text] [doi: 10.1093/database/bav081] [Medline: 26384372]
18. Zuberi K, Franz M, Rodriguez H, Montojo J, Lopes CT, Bader GD, et al. GeneMANIA prediction server 2013 update. Nucleic Acids Res 2013 Jul;41(Web Server issue):W115-W122 [FREE Full text] [doi: 10.1093/nar/gkt533] [Medline: 23794635]
19. Stelzer G, Inger A, Olender T, Iny-Stein T, Dalah I, Harel A, et al. GeneDecks: paralog hunting and gene-set distillation with GeneCards annotation. OMICS 2009 Dec;13(6):477-487. [doi: 10.1089/omi.2009.0069] [Medline: 20001862]
20. Tsuruoka Y, Tsujii J, Ananiadou S. FACTA: a text search engine for finding associated biomedical concepts. Bioinformatics 2008 Nov 1;24(21):2559-2560 [FREE Full text] [doi: 10.1093/bioinformatics/btn469] [Medline: 18772154]
21. Wei C, Kao H, Lu Z. PubTator: a web-based text mining tool for assisting biocuration. Nucleic Acids Res 2013 Jul;41(Web Server issue):W518-W522 [FREE Full text] [doi: 10.1093/nar/gkt441] [Medline: 23703206]
22. Mandloi S, Chakrabarti S. PALM-IST: pathway assembly from literature mining--an information search tool. Sci Rep 2015;5:10021 [FREE Full text] [doi: 10.1038/srep10021] [Medline: 25989388]
23. Europe PMC Consortium. Europe PMC: a full-text literature database for the life sciences and platform for innovation. Nucleic Acids Res 2015 Jan;43(Database issue):D1042-D1048 [FREE Full text] [doi: 10.1093/nar/gku1061] [Medline: 25378340]
24. Hoffmann R, Valencia A. A gene network for navigating the literature. Nat Genet 2004 Jul;36(7):664. [doi: 10.1038/ng0704-664] [Medline: 15226743]
25. Yachdav G, Goldberg T, Wilzbach S, Dao D, Shih I, Choudhary S, et al. Anatomy of BioJS, an open source community for the life sciences. Elife 2015;4:e07009 [FREE Full text] [doi: 10.7554/eLife.07009] [Medline: 26153621]
26. Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T, et al. Analysis of protein-coding genetic variation in 60,706 humans. Nature 2016 Aug 18;536(7616):285-291. [doi: 10.1038/nature19057] [Medline: 27535533]
27. Lekschas F, Stachelscheid H, Seltmann S, Kurtz A. Semantic Body Browser: graphical exploration of an organism and spatially resolved expression data visualization. Bioinformatics 2015 Mar 1;31(5):794-796 [FREE Full text] [doi: 10.1093/bioinformatics/btu707] [Medline: 25344497]
28. Shi Z, Wang J, Zhang B. NetGestalt: integrating multidimensional omics data over biological networks. Nat Methods 2013 Jul;10(7):597-598 [FREE Full text] [doi: 10.1038/nmeth.2517] [Medline: 23807191]
29. Chen H, Sharp BM. Content-rich biological network constructed by mining PubMed abstracts. BMC Bioinformatics 2004 Oct 8;5:147 [FREE Full text] [doi: 10.1186/1471-2105-5-147] [Medline: 15473905]
30. Van Landeghem LS, Björne J, Wei C, Hakala K, Pyysalo S, Ananiadou S, et al. Large-scale event extraction from literature with multi-level gene normalization. PLoS One 2013;8(4):e55814 [FREE Full text] [doi: 10.1371/journal.pone.0055814] [Medline: 23613707]
31. Hodis E, Prilusky J, Sussman JL. Proteopedia: A collaborative, virtual 3D web-resource for protein and biomolecule structure and function. Biochem Mol Biol Educ 2010 Sep;38(5):341-342 [FREE Full text] [doi: 10.1002/bmb.20431] [Medline: 21567857]
32. Hoffmann R. A wiki for the life sciences where authorship matters. Nat Genet 2008 Sep;40(9):1047-1051. [doi: 10.1038/ng.f.217] [Medline: 18728691]
33. Lee E, Helt GA, Reese JT, Munoz-Torres MC, Childers CP, Buels RM, et al. Web Apollo: a web-based genomic annotation editing platform. Genome Biol 2013;14(8):R93 [FREE Full text] [doi: 10.1186/gb-2013-14-8-r93] [Medline: 24000942]

34. Kamphans T, Krawitz PM. GeneTalk: an expert exchange platform for assessing rare sequence variants in personal genomes. Bioinformatics 2012 Oct 1;28(19):2515-2516 [FREE Full text] [doi: 10.1093/bioinformatics/bts462] [Medline: 22826540]
35. Leiserson MD, Gramazio CC, Hu J, Wu H, Laidlaw DH, Raphael BJ. MAGI: visualization and collaborative annotation of genomic aberrations. Nat Methods 2015 Jun;12(6):483-484. [doi: 10.1038/nmeth.3412] [Medline: 26020500]
36. Gormley C, Tong Z. Elasticsearch: The Definitive Guide. California: O'Reilly Media; 2015.
37. Cunningham F, Amode MR, Barrell D, Beal K, Billis K, Brent S, Searle Stephen M J, et al. Ensembl 2015. Nucleic Acids Res 2015 Jan;43(Database issue):D662-D669 [FREE Full text] [doi: 10.1093/nar/gku1010] [Medline: 25352552]
38. Pruitt KD, Brown GR, Hiatt SM, Thibaud-Nissen F, Astashyn A, Ermolaeva O, et al. RefSeq: an update on mammalian reference sequences. Nucleic Acids Res 2014 Jan;42(Database issue):D756-D763 [FREE Full text] [doi: 10.1093/nar/gkt1114] [Medline: 24259432]
39. Su AI, Wiltshire T, Batalov S, Lapp H, Ching KA, Block D, et al. A gene atlas of the mouse and human protein-encoding transcriptomes. Proc Natl Acad Sci U S A 2004 Apr 20;101(16):6062-6067 [FREE Full text] [doi: 10.1073/pnas.0400782101] [Medline: 15075390]
40. Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, et al. NCBI GEO: archive for functional genomics data sets--update. Nucleic Acids Res 2013 Jan;41(Database issue):D991-D995 [FREE Full text] [doi: 10.1093/nar/gks1193] [Medline: 23193258]
41. Gene Ontology Consortium. Gene Ontology Consortium: going forward. Nucleic Acids Res 2015 Jan;43(Database issue):D1049-D1056 [FREE Full text] [doi: 10.1093/nar/gku1179] [Medline: 25428369]
42. Linard B, Allot A, Schneider R, Morel C, Ripp R, Bigler M, et al. OrthoInspector 2.0: software and database updates. Bioinformatics 2015 Feb 1;31(3):447-448 [FREE Full text] [doi: 10.1093/bioinformatics/btu642] [Medline: 25273105]
43. Bird S. NLTK. 2006 Presented at: Proceedings of the COLING/ACL on Interactive presentation sessions; 2006; Sydney. [doi: 10.3115/1225403.1225421]
44. Rehurek R, Sojka. Software framework for topic modelling with large corpora. 2010 Presented at: LREC 2010 workshop New Challenges for NLP Frameworks; 2010; Valetta p. 46-50.
45. Köhler S, Doelken SC, Mungall CJ, Bauer S, Firth HV, Bailleul-Forestier I, et al. The human phenotype ontology project: linking molecular biology and disease through phenotype data. Nucleic Acids Res 2014 Jan;42(Database issue):D966-D974 [FREE Full text] [doi: 10.1093/nar/gkt1026] [Medline: 24217912]
46. Landrum MJ, Lee JM, Riley GR, Jang W, Rubinstein WS, Church DM, et al. ClinVar: public archive of relationships among sequence variation and human phenotype. Nucleic Acids Res 2014 Jan;42(Database issue):D980-D985 [FREE Full text] [doi: 10.1093/nar/gkt1113] [Medline: 24234437]
47. McLaren W, Pritchard B, Rios D, Chen Y, Flicek P, Cunningham F. Deriving the consequences of genomic variants with the ensembl API and SNP effect predictor. Bioinformatics 2010 Aug 15;26(16):2069-2070 [FREE Full text] [doi: 10.1093/bioinformatics/btq330] [Medline: 20562413]
48. Cunningham F, Moore B, Ruiz-Schultz N, Ritchie GR, Eilbeck K. Improving the sequence ontology terminology for genomic variant annotation. J Biomed Semantics 2015;6:32 [FREE Full text] [doi: 10.1186/s13326-015-0030-4] [Medline: 26229585]
49. Thompson JD, Higgins DG, Gibson TJ. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 1994 Nov 11;22(22):4673-4680. [Medline: 7984417]
50. Fraser N. Google. 2012. google-diff-match-patch URL: https://code.google.com/p/google-diff-match-patch/ [accessed 2016-09-19] [WebCite Cache ID 6kdGfLcD8]
51. Reyes-Palomares A, Rodríguez-López R, Ranea JA, Sánchez-Jiménez F, Sánchez JF, Medina MA. Global analysis of the human pathophenotypic similarity gene network merges disease module components. PLoS One 2013;8(2):e56653 [FREE Full text] [doi: 10.1371/journal.pone.0056653] [Medline: 23437198]
52. Szklarczyk D, Franceschini A, Wyder S, Forslund K, Heller D, Huerta-Cepas J, et al. STRING v10: protein-protein interaction networks, integrated over the tree of life. Nucleic Acids Res 2015 Jan;43(Database issue):D447-D452 [FREE Full text] [doi: 10.1093/nar/gku1003] [Medline: 25352553]
53. Khorkova O, Hsiao J, Wahlestedt C. Basic biology and therapeutic implications of lncRNA. Adv Drug Deliv Rev 2015 Jun 29;87:15-24 [FREE Full text] [doi: 10.1016/j.addr.2015.05.012] [Medline: 26024979]
54. Luo H, Sun Y, Wei G, Luo J, Yang X, Liu W, et al. Functional characterization of long noncoding RNA Lnc_bc060912 in human lung carcinoma cells. Biochemistry 2015 May 12;54(18):2895-2902. [doi: 10.1021/acs.biochem.5b00259] [Medline: 25848691]
55. Sun J, Ding W, Zhi J, Chen W. MiR-200 suppresses metastases of colorectal cancer through ZEB1. Tumour Biol 2015:- Epub ahead of print. [doi: 10.1007/s13277-015-3822-3] [Medline: 26242262]
56. Maiella S, Rath A, Angin C, Mousson F, Kremp O. [Orphanet and its consortium: where to find expert-validated information on rare diseases]. Rev Neurol (Paris) 2013 Feb;169(Suppl 1):S3-S8. [doi: 10.1016/S0035-3787(13)70052-3] [Medline: 23452769]
57. Jee K, Kim G. Potentiality of big data in the medical sector: focus on how to reshape the healthcare system. Healthc Inform Res 2013 Jun;19(2):79-85 [FREE Full text] [doi: 10.4258/hir.2013.19.2.79] [Medline: 23882412]

58. Gunay-Aygun M, Gahl W, Heller T. Congenital hepatic fibrosis overview. In: GeneReviews. Seattle: University of Washington, Seattle; 2014.

Abbreviations

CDS: coding DNA sequence
EFO: experimental factor ontology
FSS: functional semantic similarity
FTP: file transfer protocol
GEO: gene expression omnibus
GO: gene ontology
HPO: human phenotype ontology
IC: information content
IDF: inverse document frequency
MeSH: MEdical Subject Headings
NCBI: National Center for Biotechnology Information
NLTK: natural language toolkit
OMIM: Online Mendelian Inheritance in Man
ORM: object-relational mapping
PDF: portable document format
REST: representational state transfer
STRING: search tool for recurring instances of neighboring genes
TF: term frequency
UCSC: University of California, Santa Cruz
UTR: untranslated region
VCF: variant call format
VEP: variant effect predictor

Edited by G Eysenbach; submitted 05.10.16; peer-reviewed by M Mazzucato, J Wang; comments to author 17.11.16; revised version received 21.12.16; accepted 04.03.17; published 16.06.17

Please cite as:
Allot A, Chennen K, Nevers Y, Poidevin L, Kress A, Ripp R, Thompson JD, Poch O, Lecompte O
MyGeneFriends: A Social Network Linking Genes, Genetic Diseases, and Researchers
J Med Internet Res 2017;19(6):e212
URL: http://www.jmir.org/2017/6/e212/
doi:10.2196/jmir.6676
PMID:28623182

Yannis NEVERS

Exploitation de marqueurs évolutifs pour l'étude des relations génotype/phénotype.

Application aux ciliopathies.

Résumé

A l'ère des omiques, l'étude des relations génotype-phénotype repose sur l'intégration de données diverses décrivant des aspects complémentaires des systèmes biologiques. La génomique comparative offre un angle d'approche original, celui de l'évolution, qui permet d'exploiter la grande diversité phénotypique du Vivant. Dans ce contexte, mes travaux de thèse ont porté sur la conception de marqueurs évolutifs décrivant les gènes selon leur histoire évolutive. Dans un premier temps, j'ai construit une ressource d'orthologie complète, OrthoInspector 3.0 pour extraire une information évolutive synthétique des données génomiques. J'ai ensuite développé des outils d'exploration de ces marqueurs en relation avec les données fonctionnelles et/ou phénotypiques. Ces méthodes ont été intégrées à la ressource OrthoInspector ainsi qu'au réseau social MyGeneFriends et appliquées à l'étude des ciliopathies, conduisant à l'identification de 87 nouveaux gènes ciliaires.

Mots-clé : relations génotype-phénotype, évolution, orthologie, bases de données, ciliopathies, biologie des systèmes

Summary

In the omics era, the study of genotype-phenotype relations requires the integration of a wide variety of data to describe diverse aspects of biological systems. Comparative genomics provides an original perspective, that of evolution, allowing the exploitation of the wide phenotypic diversity of living species. My thesis focused on the design of evolutionary markers to describe genes according to their evolutionary history. First, I built an exhaustive orthology resource, called OrthoInspector 3.0, to extract synthetic evolutionary information from genomic data. I then developed methods to explore the markers in relation to functional or phenotypic data. These methods have been incorporated in the OrthoInspector resource, as well as in the MyGeneFriends social network and applied to the study of ciliopathies, leading to the identification of 87 new ciliary genes.

Keywords: genotype-phenotype relations, evolution, orthology, databases, ciliopathies, systems biology