École Doctorale des Sciences de la Vie et de la Santé

Thèse

présentée pour l'obtention du titre de

Docteur de l'Université d'Aix-Marseille

en Bioinformatique, Biochimie Structurale et Génomique

par

Jacques DAINAT

Étude du processus de perte de gènes et de pseudogénisation
Intégration et informatisation des concepts de l'évolution biologique
Application à la lignée humaine depuis l'origine des Eucaryotes

Soutenue publiquement le 16 octobre 2012

Membres du jury :

Rapporteurs :	M. Philippe MONGET, Directeur de Recherche INRA, Tours
	M. Hugues ROEST CROLLIUS, Directeur de Recherche CNRS, Paris
Examinateurs :	M. Michel LAURIN, Directeur de Recherche CNRS, Paris
	M. Pedro COUTINHO, Professeur des Universités, Marseille
Directeur de thèse :	M. Pierre PONTAROTTI, Directeur de Recherche CNRS, Marseille
Co-directeur de thèse :	M. Philippe GOURET, Ingénieur de Recherche, Marseille

Laboratoire d'Analyse, Topologie, Probabilités – Equipe Evolution Biologique & Modélisation
UMR CNRS 7353/IFR48/ECCOREV FR N°3098
Université d'Aix-Marseille - 13331 Marseille

École Doctorale des Sciences de la Vie et de la Santé

Thèse

présentée pour l'obtention du titre de

Docteur de l'Université d'Aix-Marseille

en Bioinformatique, Biochimie Structurale et Génomique

par

Jacques DAINAT

Étude du processus de perte de gènes et de pseudogénisation
Intégration et informatisation des concepts de l'évolution biologique
Application à la lignée humaine depuis l'origine des Eucaryotes

Soutenue publiquement le 16 octobre 2012

Membres du jury :

Rapporteurs :	M. Philippe MONGET, Directeur de Recherche INRA, Tours
	M. Hugues ROEST CROLLIUS, Directeur de Recherche CNRS, Paris
Examinateurs :	M. Michel LAURIN, Directeur de Recherche CNRS, Paris
	M. Pedro COUTINHO, Professeur des Universités, Marseille
Directeur de thèse :	M. Pierre PONTAROTTI, Directeur de Recherche CNRS, Marseille
Co-directeur de thèse :	M. Philippe GOURET, Ingénieur de Recherche, Marseille

Laboratoire d'Analyse, Topologie, Probabilités – Equipe Evolution Biologique & Modélisation
UMR CNRS 7353/IFR48/ECCOREV FR N°3098
Université d'Aix-Marseille - 13331 Marseille

Résumé

La biologie a connu une extraordinaire révolution avec l'arrivée de nombreux génomes entièrement séquencés. L'analyse de la quantité d'informations disponibles nécessite la création et l'utilisation d'outils informatiques automatisés. L'interprétation des données biologiques prend tout son sens à la lumière de l'évolution. En ce sens, les études évolutives sont incontestablement nécessaires pour donner un sens aux données biologiques. Dans ce contexte, le laboratoire développe des outils pour étudier l'évolution des génomes (et protéomes) à travers les mutations subies. Cette thèse porte sur l'étude spécifique des événements de pertes de gènes unitaires. Ces événements peuvent révéler des pertes de fonctions très instructives pour comprendre l'évolution des espèces. En premier lieu, j'ai développé l'outil GLADX qui mime l'expertise humaine afin d'étudier automatiquement et avec précision les événements de pertes de gènes unitaires. Ces études se basent sur la création et l'interprétation de données phylogénétiques, de BLAST, de prédictions protéiques, etc., dans un contexte automatisé. Ensuite, j'ai développé une stratégie utilisant l'outil GLADX pour étudier à grande échelle les pertes de gènes unitaires au cours de l'évolution du protéome humain. La stratégie utilise d'abord comme filtre l'analyse de groupes d'orthologues fabriqués par un outil de clustérisation à partir du protéome complet de nombreuses espèces. Cette analyse a permis de détecter 6237 pertes de gènes unitaires putatives dans la lignée humaine. L'étude approfondie de ces pertes avec GLADX a mis en évidence de nombreux problèmes liés à la qualité des données disponibles dans les bases de données. Elle a essentiellement permis de détecter 1318 pertes de gènes unitaires depuis l'ancêtre des Eucaryotes correspondants à près de 5% du protéome humain. Cette étude montre l'importance du phénomène de pertes de gènes unitaires dans l'évolution des génomes. La majorité des pertes identifiées sont décrites pour la première fois. Une des particularités de cette thèse est d'aborder et analyser aussi bien les pertes de gènes unitaires sans signaux des séquences d'origine, que celles liées à des pseudogènes. De plus, les événements de pertes et de pseudogénisations sont mis en évidence dans un contexte évolutif. L'apport d'informations fonctionnelles et des études comparatives permettront d'appréhender les phénomènes sous-jacents ayant induit ces pertes.

Mots clés : Perte de gènes unitaires, pseudogène unitaire, pseudogénisation, automatisation, phylogénie, homme

Abstract

Biology has undergone an extraordinary revolution with the appearance of numerous whole genomes sequenced. Analysis of the amount of information available requires creation and use of automated tools. The interpretation of biological data becomes meaningful in light of evolution. In view of all this, evolutionary studies are undoubtedly necessary to highlight the biological data. In this context, the laboratory develops tools to study the genomes (and proteomes) evolution through all the undergone mutations. The project of this thesis focuses specifically on the events of unitary gene losses. These events may reveal loss of functions very instructive for understanding the evolution of species. First, I developed the GLADX tool that mimics human expertise to automatically and accurately investigate the events of unitary gene losses. These studies are based on the creation and interpretation of phylogenetic data, BLAST, predictions of protein, etc., in an automated environment. Secondly, I developed a strategy using GLADX tool to study, at large-scale, the loss of unitary genes during the evolution of the human proteome. The strategy uses, in the first step, the analysis of orthologous groups produced by a clustering tool from complete proteomes of numerous species. This analysis used as a filter, allowed detecting 6237 putative losses in the human lineage. The study of these unitary gene loss cases has been deepened with GLADX and allowed to highlight many problems with the quality of available data in databases. It enabled to mainly identify 1318 unitary gene losses from the ancestor of Eukaryotes corresponding to about 5% of the human proteome. This study shows the importance of the loss phenomenon of well-established genes in the genome evolution. The majority of losses are described for the first time. One feature of this thesis is that losses without origin sequences signal and those associated with pseudogenes are both discussed and analyzed. Moreover, loss and pseudogenization events are highlighted in an evolutionary context. The provision of functional information and comparative studies could allow understanding the underlying phenomena of such losses.

Key words: unitary gene loss, unitary pseudogene, pseudogenization, automation, phylogeny, human

« Soyons sincères : l'homme est un champignon rêveur ;

un concombre qui a des visions ;

un salsifis qui souffre de marottes. »

Alexandre Vialatte (1901 - 1971)

Remerciements

En premier lieu je souhaiterais remercier Philippe Monget, Hugues Roest Crollius, tous deux Directeurs de recherche, qui me font l'honneur d'être rapporteurs de ce travail de thèse.

Merci à Michel Laurin, Directeur de Recherche et Pedro Coutinho Professeur des Universités pour avoir accepté de juger ce travail.

Un immense merci à Pierre Pontarotti, mon directeur de thèse, pour son écoute attentive et le temps qu'il m'a accordé durant ces trois années de thèse. Il a su m'apporter des connaissances essentielles et me guider dans ce travail de recherche. Je lui suis très reconnaissant pour la diversité et la richesse des questions scientifiques abordées, ainsi que leurs aspects philosophiques. Ces approches concrètes et les évasions intellectuelles m'ont permis de continuer à grandir ces trois dernières années. Merci enfin pour sa sympathie, sa simplicité et sa convivialité.

Je tiens à remercier vivement Philippe Gouret pour sa patience et son encadrement informatique efficace durant les longs mois de développement informatique : il a été très disponible et sympathique tout au long de ma thèse.

Un merci particulier à Julien Paganini : nous avons partagé le même bureau pendant presque trois ans. J'ai beaucoup appris à ses côtés. Son amabilité et sa bonne humeur ont permis de passer les longues journées dans une ambiance agréable. Je le remercie surtout pour m'avoir supporté, malgré de nombreuses interventions musicales, chantées, pas toujours faciles à écouter.

Je remercie toute l'équipe avec qui j'ai pu avoir des interactions enrichissantes, et particulièrement Olivier Chabrol qui a su m'épauler sur certains aspects informatiques, Manuela Carenzi pour son aide précieuse dans mes démarches mathématiques et Antonio Hernandez Lopez pour ses compétences linguistiques en Anglais.

Durant ma thèse j'ai eu l'occasion de rencontrer de nombreuses personnes extérieures qui m'ont été utiles et m’ont beaucoup appris. Je pense en particulier à Anthony Levasseur et Étienne Danchin.

Je n'oublie pas les stagiaires, Rami, Emmanuelle et Tharwa qui ont été précieux pour l'étude des nombreux résultats obtenus.

La thèse ne s'arrête pas aux frontières du laboratoire : je souhaite également remercier ma famille et mes amis qui ont su me soutenir, me motiver et être disponible dans les moments difficiles.

Merci à Juliette pour avoir embelli mes années de thèse, et surtout pour son aide précieuse en de nombreux domaines.

Je remercie tous les participants aux discussions passionnantes que nous avons eues au cours de ces années, sur la politique, sur des faits divers, scientifiques, voire philosophiques. La diversité des connaissances et des points de vue est essentielle pour s'épanouir dans la vie.

Glossaire & abréviations

BD : Base de données.

BLAST : Basic Local Alignment Search Tool. Programme informatique qui effectue des comparaisons de séquences.
Blast est entré dans le langage courant pour BLAST et signifie une recherche de séquences « similaires » de manière générale sans forcément définir le type de d'algorithme utilisé.

BLASTN : BLAST entre des séquences nucléotidiques.

BLASTP : BLAST entre des séquences protéiques.

BLAT : BLAST-like Alignment Tool. Programme rapide de recherche de similarité et d'alignement de séquences nucléotidiques.

Bootstrap : C'est une technique d'inférence statistique basée sur des ré-échantillonnages successifs des données. Elle est utilisée pour tester la robustesse de nœuds phylogénétiques. Plus la valeur calculée est proche de 100%, plus un nœud est robuste.

COG : Cluster of Orthologous Groups of proteins ⇔ cluster de groupes de protéines orthologues.

EST : Expressed Sequence Tag ⇔ Marqueur de séquence exprimée ; c'est une courte séquence d'ADN complémentaire.

Framework : Architecture logicielle permettant d'avoir des fondations et des règles de tout ou une partie d'un logiciel.

Gap : trou de séquence.

GO : Groupe de gènes orthologues.

Hit de BLAST : Séquence similaire retrouvé par blast formé d'un ou plusieurs HSP concaténés ensemble.

HSP : High Scoring Pair ⇔ Zone possédant une similarité significative dans le cadre de résultat de BLAST.

KOG : Cluster of euKayotic Orthologous Groups of proteins ⇔ cluster de groupes de protéines orthologues Eucaryotes.

LCA : Last Common Ancestor ⇔ dernier ancêtre commun.

ORF : Open Reading Frame ⇔ cadre ouvert de lecture : Séquences possédant un codon initiateur et stop pouvant coder une protéine.

Pipeline : Suite de scripts informatique automatisant des processus d'analyses.

Taxid Pipeline : Identifiant taxonomique unique.

TBLASTN : BLAST d'une protéine contre des séquences nucléotidiques (dans les six cadres de lecture).

THG : Transfert Horizontal de Gène.

Table des matières

Index des illustrations

Index des tableaux

Index des articles

Index des annexes

Avant-propos

J'ai préparé cette thèse de doctorat sous la direction de Pierre Pontarotti, directeur de l'équipe « Evolution Biologique et Modélisation » (UMR 6632) hébergée à l'Université de Provence sur le site Saint Charles. Les études menées par l'équipe portent sur les cadres conceptuels de l'évolution à la croisée de la biologie, de l'informatique et des mathématiques. Le nombre des données biologiques qui s'accroît de jour en jour, nécessite une recherche informatique automatisée. Les grandes avancées en biologie et en informatique ont abouti conjointement au séquençage et à l'annotation de nombreux génomes avec de nombreuses informations au sein de phyla variés. Cet élan continue de s'accélérer. L'équipe cherche à tirer profit de l'ensemble des données disponibles et des outils performants mis au point au laboratoire, afin d'étudier les événements complexes subis par les génomes lors de leur évolution. Les événements en question se présentent sous la forme de gains, de duplications, d'échanges, de pertes de domaines et de gènes, de transferts latéraux de gènes, etc. En 2007, forte des outils développés au laboratoire, l'équipe a entrepris, en collaboration avec l'équipe de bioinformatique de l'IGBMC de Strasbourg dirigée par Olivier Poch, une étude phylogénomique du protéome humain dans le cadre du projet EvolHHuPro. C'est dans ce contexte que j'ai été recruté afin de préparer une thèse de doctorat sur la perte de gènes au sein des Eucaryotes. La perte de gènes bien établis est un événement riche d'informations pour comprendre l'évolution des espèces, et m'a conduit à développer une approche qui permet de les analyser spécifiquement. En effet, les gènes bien établis sont sûrement liés à des fonctions essentielles, et la disparition de leurs fonctions doit avoir un lien avec le mode de vie de l'espèce étudiée et de son environnement. Peu d'études existent sur la perte de gènes bien établis chez les Eucaryotes, et celles publiées sont souvent incomplètes. Cela s'explique par le fait que les approches fines sont fastidieuses et peu automatisées.

Après une modélisation conceptuelle des éléments nécessaires à l'étude des pertes de gènes, tel que le problème complexe de la pseudogénisation, j'ai développé au sein du laboratoire et sous l'encadrement de Philippe Gouret un module appelé GLADX intégré au framework DAGOBAH. DAGOBAH est le centre stratégique d'études multipartites et agrégatives. Le système multi-agents de DAGOBAH permet de développer des agents qui permettent alors de travailler sur des questions diverses et spécifiques et de produire des données. Ces agents

peuvent communiquer entre eux, partager des données et des résultats, et déduire des résultats de niveau supérieur. Après un long développement de GLADX je suis arrivé à mettre au point une version qui permet une analyse de qualité des pertes de gènes. Cette version de GLADX a permis d'étudier à une grande échelle, la perte de gènes chez l'homme et de compléter ainsi le projet du laboratoire sur l'étude de l'évolution du protéome humain. Après de longs mois de calcul, j'ai découvert des centaines de nouvelles pertes jamais décrites. L'ensemble des événements de pertes et de pseudogénisations a été analysé avec précision. J'ai ainsi trouvé de nombreuses mutations nucléotidiques participant à la pseudogénisation. De plus, l'analyse par phylogénie comparative et génomique comparative m'a permis d'annoter des séquences. Pour illustrer cette idée j'ai découvert des centaines de nouveaux gènes dans différentes espèces dont l'homme, non encore décrits dans les bases de données. Des annotations systématiques au sein des génomes m'ont permis d'avoir une vision plus complète des phénomènes évolutifs.

L'ensemble de ce travail est développé dans ce manuscrit.

Note : Afin de rendre intelligible le travail de recherche effectué au cours de cette thèse, la chronologie du travail de recherche n'a pas forcément été respectée. Comme c'est souvent le cas, le cheminement de ce travail de recherche a été épistémologiquement tortueux.
Malgré un long travail de développement informatique, j'ai fait le choix de ne pas aborder la partie algorithmique dans le manuscrit afin de simplifier la présentation de mon travail de recherche.

Participations scientifiques :

Une part du travail effectué au cours de la thèse a été soumise à publication. Ces publications sont incluses dans ce manuscrit de thèse et certains résultats sont résumés ou complétés dans les chapitres correspondants.

Les travaux effectués durant la thèse ont également été exposés lors de participations à différents congrès nationaux et internationaux. Voici la liste de ces participations scientifiques :

Jacques Dainat, Pierre Pontarotti, Philippe Gouret. Gene-loss and pseudogenization in human lineage throughout the Eukaryotes evolution. *15th Evolutionary Biology Meeting*, *Marseille, France, Septembre 2011.*

Jacques Dainat, Pierre Pontarotti, Philippe Gouret. Gene-loss and pseudogenization in human lineage throughout the Eukaryotes evolution. *Annual Meeting of the Society for Molecular Biology and Evolution*, *Kyoto, Japon, Juillet 2011.*

Jacques Dainat, Julie D. Thompson, Olivier Poch, Pierre Pontarotti, Philippe Gouret. Lineage-specific orthologous gene loss and pseudogenisation, automated analysis in Metazoans. *14th Evolutionary Biology Meeting, Marseille, France, Septembre 2010.*

Jacques Dainat, Julie D. Thompson, Olivier Poch, Pierre Pontarotti, Philippe Gouret. "GeneLoss": Automation of the study of lineage-specific gene loss and pseudogenisation. *Journées Ouvertes de Biologie, Informatique et Mathématiques, Montpellier, France, Septembre 2010.*

Jacques Dainat, Julie D. Thompson, Olivier Poch, Pierre Pontarotti, Philippe Gouret. Automation of the study of lineage-specific gene loss and pseudogenisation process. *XVIIIème Colloque de l'Ecole Doctorale des Sciences de la Vie et de la Santé, Marseille, France, Mai 2010.*

Jacques Dainat, Julie D. Thompson, Olivier Poch, Pierre Pontarotti, Philippe Gouret. Automation of the study of lineage-specific gene loss and pseudogenisation process. *13th Evolutionary Biology Meeting, Marseille, France, Septembre 2009.*

Chapitre I – Introduction

1 Préambule historique

1.1 Perception de l'Evolution jusqu'au Moyen Age

La transmission des caractères d'un individu à sa descendance est connue de l'homme au paléolithique supérieur (-30000 à -12000 ans), qui pratique déjà la domestication et la sélection. Par exemple les loups sont à l'origine des chiens aux environs de 31 700 ans av. J.-C. (Germonpre *et al.*, 2009 ; Pionnier-Capitan *et al.*, 2011). La domestication se généralise au néolithique (9000 à 3500 av. J.-C.) avec la domestication entre autres des herbivores. L'observation des phénomènes liés à la domestication est à l'origine de l'idée d'évolution, et la modification des espèces au cours du temps a permis d'appréhender l'origine de la diversité du vivant. La formalisation de l'évolution de la vie est déjà présente chez les philosophes grecs (Démocrite, Epicure, Lucrèce), mais la conception fixiste émise par Aristote (384-322 av. J.-C.) prédomine ; elle est en contradiction avec l'idée même d'évolution puisqu'elle affirme une existence éternelle des genres ou espèces. La génération spontanée qui est l'apparition d'êtres vivants sans ascendant est une théorie synthétisée par Aristote. Elle a longtemps fait partie du sens commun à cause de l'apparition d'êtres vivants là où on n'en avait pas repéré au préalable avec les méthodes d'observations courantes. La transmission des caractères acquis était déjà appréhendée par Aristote "*c'est l'homme qui engendre l'homme*" (Aristote, n.d.). Mais elle ne s'exprime sous cette forme qu'à partir de la fin du XIX^{ème} lorsque la distinction entre caractères innés et acquis apparaît.

1.2 Perception de l'Evolution du Moyen Age au Siècle des Lumières

Le paradigme entretenu par les religions et la théologie selon lequel les êtres vivants sont le fruit d'une intervention divine (créationnisme), reste omniprésent dans la pensée occidentale. Durant des siècles, le judaïsme et le christianisme se basent sur les récits bibliques pour estimer l'âge de la terre qui est évalué à quelques milliers d'années. Alphonse de Vignole (1649-1744) collecte au début du XVIII^{ème} les évaluations de l'âge de la terre et présente « plus de 200 calculs différents, dont le plus court ne compte que 3483 ans, [...] le plus long en compte 6984 » (Vignoles, 1738). Etant donné qu'à cette époque l'apparition de la vie ne remonte qu'à seulement quelques milliers d'années, il est très difficile d'imaginer un lien entre espèces et encore moins des évolutions lentes. Ainsi, pendant de longs siècles, la vision

hiérarchique des êtres vivants représentée par la *scala naturae* (Illustration 1), qui peut être rapprochée de l'arbre de la vie accepté de nos jours, n'est pas remise en question.

Illustration 1 : La scala naturæ
Alias l'échelle des êtres, c'est une conception de l'ordre de l'univers qui date de l'époque médiévale et qui se caractérise par une stricte hiérarchie entre les différents niveaux.

1.3 Le Siècle des Lumières

Au XVIII^{ème} siècle, les théories de la génération spontanée ainsi que celle de la transmission des caractères acquis sont ébranlées par les réflexions des philosophes ; elles seront définitivement invalidées à la fin du XIX^{ème} siècle. L'accumulation de nombreuses observations permet d'établir des liens entre les différentes espèces et de proposer l'idée d'évolution. On peut citer notamment la classification des organismes en botanique, basée le plus souvent sur l'anatomie et la morphologie des espèces qui existe depuis Théophraste (vers 372-287 av. J.-C.). De nombreuses méthodes de classifications sont utilisées et permettent fréquemment de faire apparaître les mêmes grandes familles. Le Suédois Carl von Linné (1707-1778) affirme que ces rapprochements en grandes familles ne peuvent pas être une simple coïncidence, et pense qu'il doit exister une classification unique dite Classification

Naturelle. Le travail qu'il entreprend permet une amélioration de la classification des espèces. Grâce à une hiérarchisation des taxons et une approche binomiale pour nommer les espèces (en latin), nom de genres et nom d'espèces, il donne les bases de la nomenclature scientifique et offre un langage commun par-delà les noms vernaculaires propres à chaque langue (Linnaeus, 1735). Un peu plus tard, Buffon (1707-1788) décrit dans ses travaux les ressemblances entre l'homme et le singe (Buffon & Daubenton, 1766) et remet en cause un âge récent de la terre. L'incroyable augmentation de l'âge de la terre défendue par Buffon permet un changement conceptuel dans la chronologie. C'est l'origine d'une étape importante dans le développement des théories de l'évolution (Illustration 2).

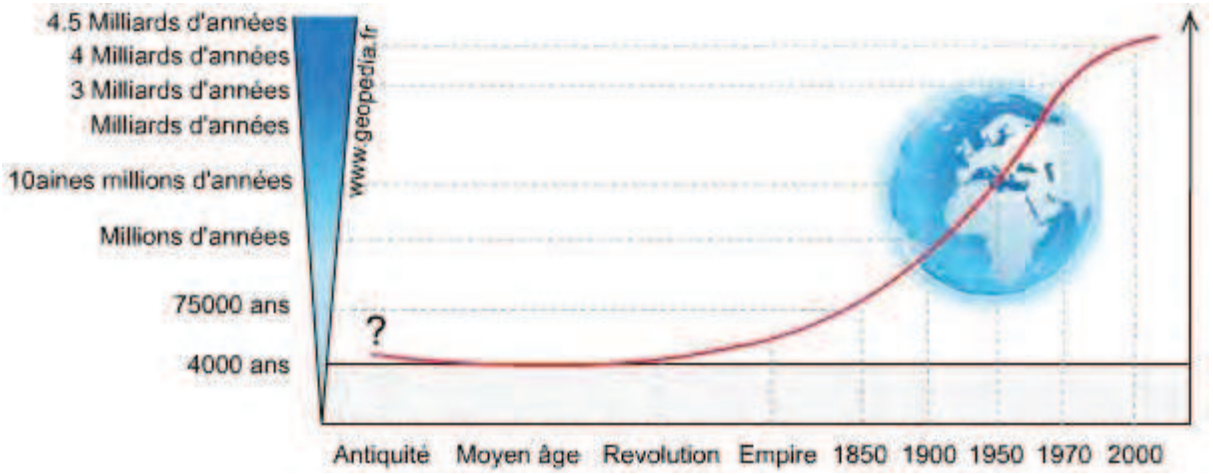


Illustration 2 : Perception de l'âge de la terre par l'homme au cours du temps

Source : http://www.geopedia.fr/age-terre.htm

Vers la fin du XVIII^{ème} siècle, le naturaliste Georges Cuvier (1769-1832) qui étudie les fossiles découverts depuis le XVI^{ème} siècle, fonde la paléontologie scientifique et permet son essor. A sa suite, d'autres chercheurs (Jean-Baptiste de Lamarck (1774-1829) par exemple) jusque-là partisans de la fixité des espèces comprennent l'importance de la paléontologie et la nécessité de classer les espèces pour mieux percevoir et observer le vivant.

1.4 Les révolutions intellectuelles du XIX^{ème} siècle

Au début du XIX^{ème} siècle, Lamarck rompt avec la vision fixiste de la nature et change radicalement sa façon de la percevoir. Il élabore une théorie dite transformiste (Lamarck, 1809, 1822) qui vise à expliquer l'extinction des espèces. Il s'oppose à une classification hiérarchisée des êtres, et il pose les premières pierres de la construction de la théorie de l'évolution telle que nous la connaissons aujourd'hui. Il formule deux principes qui décrivent

le mécanisme fondamental du processus évolutif. Les organes se développent en fonction du besoin de l'organisme et de l'utilisation qui en découle. Par exemple la girafe a allongé son cou pour aller chercher sa nourriture en hauteur (Illustration 3). Il s'appuie sur l'idée de la transmission des caractères acquis : c'est ainsi que se transmettent les comportements ou les traits acquis à la génération suivante.

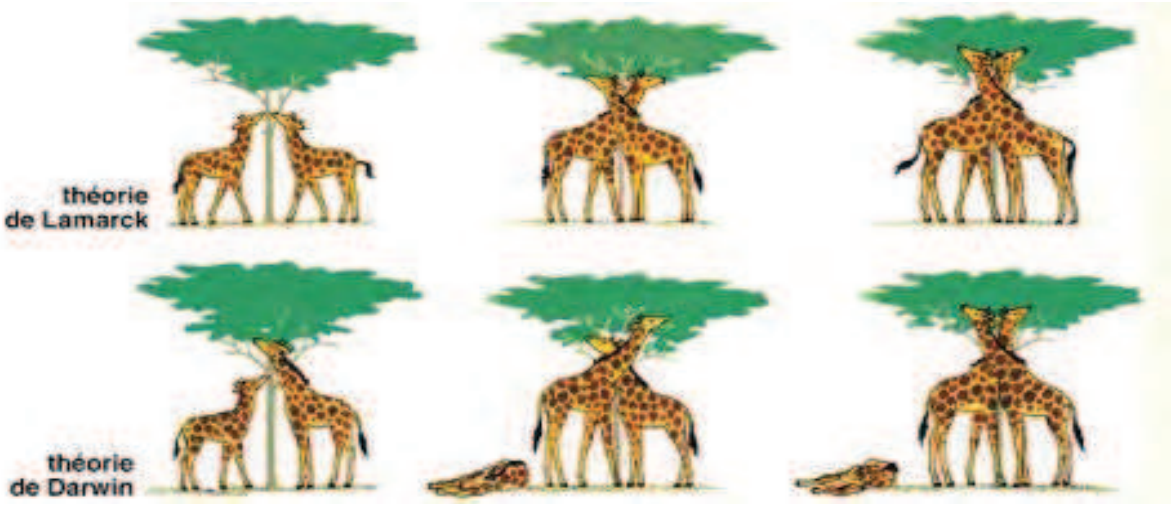


Illustration 3 : Évolution des espèces selon Lamarck et selon Darwin
Source : http://www.colvir.net/prof/serge.lapierre/LamarckDarwin.jpg

Au milieu du XIX^{ème} siècle, Charles Darwin (1809-1882) qui étudie la transmission des caractères, idée encore discutée, esquisse le premier arbre évolutif (Illustration 4). Cet arbre permet une matérialisation concrète du concept selon lequel la diversification des espèces que l'on observe provient d'un processus de divergences et de ramifications. Cette vision nouvelle va à l'encontre de l'idée d'une diversification hiérarchisée linéaire. Pendant des années Charles Darwin accumule de nombreuses preuves provenant d'observations sur l'élevage animal, l'anatomie, la morphologie, la biogéographie, etc. Toutes ses observations lui permettent d'élaborer une théorie de l'évolution qui explique de façon naturelle la complexité adaptative des êtres vivants, et leurs similitudes (Charles Darwin, 1859 ; Charles Darwin & Wallace, 1858). Cette théorie de l'évolution se base sur trois principes : le principe de variation, le principe d'adaptation et le principe d'hérédité. Ainsi, dans la variabilité des individus, les plus adaptés sont sélectionnés et transmettent leurs caractères à leurs descendances (Illustration 3). Le terme d'évolution n'est pas encore utilisé, et pour expliquer la sélection des individus les mieux adaptés il crée le terme de « sélection naturelle ». Darwin utilise toujours la théorie de la transmission des caractères acquis encore largement admise, et qui d'ailleurs ne s'oppose pas à la sélection naturelle car elle ne s'intéresse pas au mode

d'acquisition d'un caractère mais à sa sélection. Le Darwinisme admet et reconnaît des changements dus au hasard, à la différence du Lamarckisme qui est une théorie finaliste : c'est une grande différence entre eux. Ainsi les grandes variations observées chez les girafes sont dues au hasard, et le hasard cède la place à la nécessité en sélectionnant les individus les mieux adaptés.

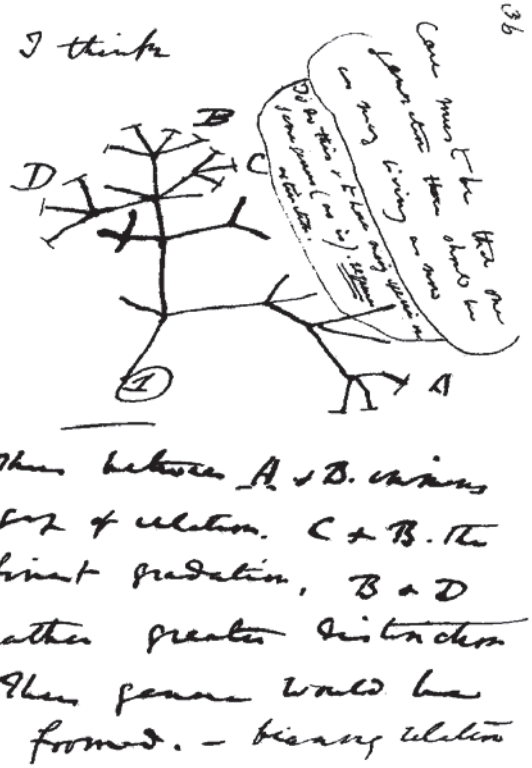


Illustration 4 : Première esquisse d'arbre évolutif faite par Darwin (C. R. Darwin, 1837)

Thomas Henry de Huxley, ami de Charles Darwin, publie dès 1863 un livre dans lequel il développe la thèse que les singes anthropoïdes sont nos proches parents (Huxley, 1863). Charles Darwin évite d'aborder le sujet de l'origine de l'homme « si encombré de préjugés » mais y contribue dans un ouvrage de 1871 (C. Darwin, 1871).

En 1868 Darwin essaie d'expliquer la transmission des caractères acquis en proposant l'hypothèse dite de la « pangenèse ». Cette hypothèse s'inspire de la théorie de Maupertuis (1698-1759), qui reprend l'idée antique d'Hippocrate (460-370 av. J.-C.). La pangenèse donnerait une mémoire aux organes sous forme de « gemmules » qui se rassembleraient dans les organes reproducteurs lors de la fécondation (Charles Darwin, 1868). D'autre part, August Weismann (1834-1914), dont les expériences portent sur la mutilation de rats, affirme que la

transmission des caractères acquis n'est pas valide (Weismann, 1892a). Malgré tout, cette théorie pangénique continue à faire couler de l'encre (Molinier, Ries, Zipfel, & Hohn, 2006). Contemporain de Darwin et Lamarck, Gregor Mendel (1822-1884), moine dans le monastère de Brno, considéré comme pionnier de la génétique, travaille sur la transmission des caractères morphologiques des pois sur plusieurs générations. De ses travaux il énonce des lois dites lois de Mendel, qui sont à la base de la génétique moderne, bien avant la découverte du support de ces informations (Johann Gregor Mendel, 1866). En s'appuyant notamment sur les travaux de Friedrich Miescher (1844-1895) qui a décrit la nucléine (Miescher, 1871), et de Van Beneden qui a découvert le phénomène de la méiose (Beneden, 1883), August Weisman élabore la théorie du « plasma germinatif ». Selon cette théorie, le « plasma germinatif » (que l'on peut associer au terme génome de nos jours) est le composant des cellules germinales responsables de l'hérédité (Weismann, 1892b). Du temps de Weismann on ne connaît presque rien de la chimie des cellules. Toutefois, grâce à certaines connaissances de la dynamique cellulaire, Walther Flemming (1843-1905) découvre en 1878 dans les cellules somatiques l'existence de structures, appelées plus tard **chromosomes**. Il découvre aussi que le nombre de chromosomes reste constant au cours de la mitose. Oscar Hertwig (1849-1922) et Eduard Adolf Strasburger (1844-1912) montrent en 1880 que la fusion de l'ovule et du spermatozoïde est un élément essentiel à la fécondation. Weismann propose alors l'existence d'un processus de réduction par deux du nombre de chromosomes dans les cellules germinales au cours de la méiose ; c'est au cours de la fécondation de l'ovule par le spermatozoïde que se reconstituent les paires de chromosomes présents dans les cellules somatiques. Weismann postule également que les chromosomes contiennent les « facteurs » de l'hérédité, dont Mendel avait prédit l'existence. Ces résultats, parmi d'autres, conduisent au rejet de la théorie de la transmission des caractères acquis, et permettent d'admettre la possibilité d'un réassortiment des gènes dans les cellules germinales à chaque génération. Weismann comprend que les cellules germinales sont d'une certaine façon immortelles ; transmises à la descendance par la reproduction, leur continuité dans le temps sous une forme matérielle est la source des phénomènes de l'hérédité. Les cellules germinales ont également la capacité de donner naissance à des cellules somatiques au potentiel de division limité (Illustration 5). "Le corps, le Soma, produit de ce point de vue et dans une certaine mesure, l'effet d'un appendice accessoire des véritables porteurs de la vie, les cellules reproductrices" (Weismann, 1892c).

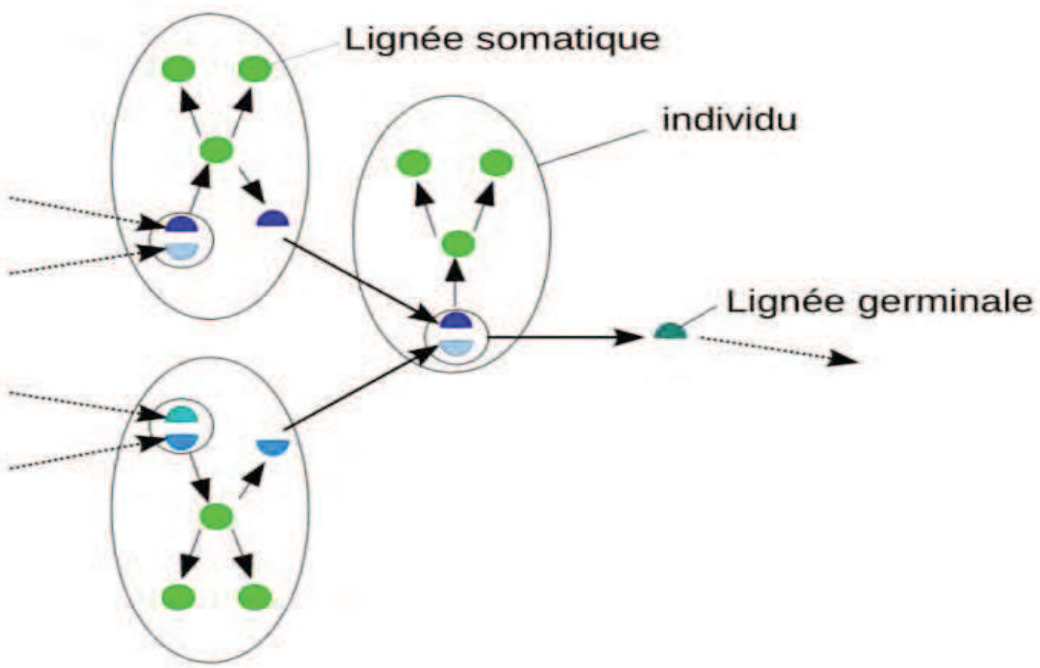


Illustration 5 : Transmission des caractères par la lignée germinale

1.5 Progrès expérimentaux et technologiques

Au début du XX^{ème} siècle, Theodor Boveri (1862-1915) montre sur l'oursin que les facteurs héréditaires se trouvent dans le noyau, et qu'un nombre anormal de chromosomes entraîne un développement anormal. De son côté Walter Sutton (1877-1916) démontre qu'un œuf diploïde fécondé est constitué de paires de chromosomes : les chromosomes d'une paire proviennent l'un du gamète femelle et l'autre du gamète mâle. Ces chercheurs établissent les bases cellulaires de la théorie chromosomique de l'hérédité, qui expliquent l'hérédité mendélienne. William Henry Bateson (1861-1926) comprend qu'il est nécessaire de nommer ce phénomène de variations héréditaires, et propose en 1906 le terme de génétique : "science qui étudie les mystères de l'hérédité et leurs variations". Trois ans plus tard, Wilhelm Ludvig Johannsen (1857-1927) propose le terme de **gène** pour la partie du chromosome qui code pour un caractère. Il introduit également les notions de génotype et de phénotype en 1911. Il faut attendre 1910 et les travaux de Thomas Hunt Morgan (1866-1945) pour démontrer que les chromosomes sont le support des gènes disposés de manière linéaire, et confirmer la théorie chromosomique de l'hérédité (Morgan, Sturtevant, Muller, & Bridges, 1915). Morgan découvre la liaison génétique (tendance de certains locus ou allèles d'être hérités ensemble) et reprend la découverte de Frans Alfons Janssens (1865-1924) sur les échanges de fragments de chromatides (**enjambements** ou *crossing-over*) pendant la méiose, pour dire que ces échanges

contribuent au brassage des gènes. Morgan et Alfred Henry Sturtevant (1891-1970) développent une méthode pour situer approximativement la position des gènes sur les chromosomes, ce qui permet de construire les premières cartes génétiques. Grâce à ces découvertes, la biologie jusque-là plutôt descriptive devient une science expérimentale et donne une image cohérente des processus de l'évolution.

1.5.1 La biologie moléculaire du XX^{ème} siècle avant l'ère informatique

Dans la première partie du XX^{ème} siècle, la découverte des chromosomes (support chimique de l'information génétique) et des lois qui gouvernent les informations qu'ils contiennent, donnent naissance à la biologie moléculaire. Cette discipline scientifique favorise une synthèse des connaissances en génétique, biochimie, biophysique et mathématique, et permet ainsi de mieux appréhender les mécanismes moléculaires du fonctionnement de la cellule. Le terme de **biologie moléculaire** est utilisé pour la première fois par Warren Weaver (1894-1978) en 1938 (Weaver, 1970). Pendant longtemps, il est admis que les protéines sont les supports de l'information génétique. C'est en 1944 qu'Oswald Theodore Avery (1877-1955) met en évidence la nature chimique des gènes. Ils sont constitués d'acides désoxyribonucléiques (ADN), chaînes de polynucléotides combinant 4 nucléotides : Adénie (A) Cytosine (C) Guanine (G), et Thymine (T) (Avery, MacLeod, & McCarty, 1944). Il faut attendre 1952, et l'expérience d'Alfred Hershey et Martha Chase, pour que l'hypothèse des protéines porteuses de l'information génétique soit définitivement invalidée.

La structure de l'ADN des chromosomes est complétée par la découverte en 1953 de sa structure spatiale en hélice faite par James Dewey Watson (1928 -) et Francis Harry Compton Crick (1916 -) (Watson & Crick, 1953). Toutes ces avancées sont dues à diverses découvertes technologiques : la diffraction aux rayons X mise au point par Max Von Laue (1879-1960) en 1912, le microscope électronique théorisé par Louis de Broglie (1892-1987) en 1924 et construit à partir de 1931 et l'électrophorèse inventée par Arne Tiselius (1902-1971) pour séparer les protéines en solution (Tiselius, 1930). Entre les années 1930 et 1970, les travaux en Biologie Moléculaire permettent d'élaborer un corpus de règles communes à tous les êtres vivants, basé sur la nature des gènes et les modalités de leurs traductions en caractères.

Après la découverte de la structure en double hélice de l'ADN, la biologie moléculaire connaît d'importantes avancées. La théorie fondamentale de la biologie moléculaire intitulée « *central dogma of molecular biology* » est énoncée par Francis Crick (F. H. Crick, 1958). Cet auteur définit l'ADN comme un support stable qui peut transmettre son information de

manière identique par réplication, et qui peut définir des fonctions biologiques par transcription en ARN (ARNr, ARNm, ARNt, etc.), l'ARNm pouvant à son tour être ou non traduit en protéine. Cette théorie s'enrichit au fur et à mesure de nombreuses découvertes, avec par exemple la découverte d'une possibilité d'un retour à la forme ADN à partir d'ARN.

L'ADN est constitué de quatre nucléotides différents (A,C,T,G), et les ARN de quatre nucléotides différents (A,C,U,G). Dans l'ARN, la Cytosine (C) est remplacée par l'Uridine (U). L'ADN est très stable alors que les ARNm sont très fragiles.

Ensuite, est découvert le code génétique qui permet la traduction en protéine d'une séquence de la molécule d'ADN appelé gène (Matthaei, Jones, Martin, & Nirenberg, 1962).

Un gène est donc une séquence de nucléotides de longueur variable avec une région régulatrice, et une région codante qui peut comprendre une succession d'exons et d'introns. La plupart des gènes bactériens ou de levure ne possèdent pas d'intron, alors que la plupart des gènes des organismes multicellulaires en contiennent. La séquence des introns est souvent beaucoup plus longue que celle des exons. Seul, les exons codent pour les protéines.

La traduction du gène en protéine (chaîne d'acides aminés) passe obligatoirement par une transcription de sa séquence nucléotidique en ARNm, séquence nucléotidique complémentaire. L'ARNm est un transcrit dit primaire lorsqu'il contient un ou plusieurs introns. Un transcrit primaire doit passer par une phase de maturation qui élimine par épissage les introns et peut donner un ou plusieurs ARNm dits monocistroniques qui seront traduits en protéines. La séquence nucléotidique de l'ARNm est lue par triplets de nucléotides consécutifs appelés codons qui chacun code pour un acide aminé. Il existe 64 codons différents dont plusieurs peuvent coder pour le même acide aminé et quelques-uns contrôlent la transcription. Les mécanismes du contrôle de l'expression des gènes s'affinent par de nouvelles découvertes et permettent un essor de la biologie moléculaire.

En 1953 Frederick Sanger (1918 -) séquence pour la première fois une protéine (l'insuline). La publication de nombreuses séquences de protéines, conduit les chercheurs à créer une banque de classement est d'archivage des données. En 1965 apparaît le premier recueil des séquences protéiques (Dayhoff, Eck, Chang, & Sochard, 1965). Dès cette époque, la biologie moléculaire s'institutionnalise en Europe, avec la création en 1959 du *Journal of Molecular Biology* et la création en 1964 de l'*European Molecular Biology Organization* (EMBO), suivie par celle de l'*European Molecular Biology Laboratory* (EMBL) en 1974 qui regroupe aujourd'hui plus de 20 pays.

1.5.2 Naissance de la théorie synthétique de l'évolution

En parallèle à l'apparition de la biologie moléculaire, le généticien Theodosius Dobzhansky (1900-1975) fait une synthèse en 1937 des théories de l'évolution (Dobzhansky, 1937). Il considère que les phénomènes évolutifs se déroulent sous l'action de la sélection naturelle, et permettent des changements de fréquences de gènes au sein des lignées. Cette hypothèse basée sur la théorie darwinienne intègre donc la théorie de l'hérédité mendélienne, les connaissances acquises en génétique des populations, en systématique (Mayr, 1942, 1963), et en paléontologie (Simpson, 1944). La synthèse de ces travaux explique l'absence d'intermédiaire entre différentes espèces, par les changements lents et continus des espèces ancestrales, mais aussi par le manque d'archives paléontologiques. Cette théorie donne naissance à une vision unitaire de l'évolution, appelé théorie **synthétique de l'évolution** ou **néo-darwinisme**. Cette vision constitue pour le biologiste un cadre conceptuel qui donne tout leur sens aux données scientifiques. Elle est traduite dans l'aphorisme de Theodosius Dobzhansky " Rien n'a de sens en biologie, si ce n'est à la lumière de l'évolution ". Au travers de la théorie synthétique de l'évolution, la vision de l'évolution de Darwin s'élargit et se comprend dès lors comme la transformation de groupes d'individus au sein d'une même espèce sous l'influence de la sélection naturelle. La fréquence d'une mutation ou d'un gène au sein d'une population est fonction de la valeur adaptative des caractères qu'elle amène aux individus en lien avec leur environnement. Des allèles silencieux dans une partie de la population peuvent devenir bénéfiques dans un nouveau biotope. Lorsque des populations se trouvent isolées, elles peuvent acquérir certains caractères particuliers, indépendamment des autres membres de l'espèce. La divergence donne alors des espèces bien distinctes lorsque les populations ne peuvent pas se reproduire entre elles. Cet événement est connu en biologie évolutive sous le nom de **spéciation**. L'isolation de populations pour une espèce donnée peut être de natures diverses : géographique, écologique, physiologique, éthologique, etc.

1.5.3 La biologie moléculaire à l'ère informatique

1.5.3.1 Développement de l'ère informatique

La deuxième partie du XX^{ème} siècle voit naître un ensemble de techniques, de protocoles, qui envahissent tous les champs de la biologie et rendent possible l'étude des gènes et des génomes à l'échelle moléculaire. Rappelons que les premières machines programmables sont les métiers à tisser de Joseph-Marie Jacquard (1752-1834) au tout début du XIX^{ème} siècle qui utilisent des cartes perforées. Il faut attendre les travaux conceptuels de Alan Mathison Turing

(1912-1954) pour que soit conçu un ordinateur programmable (Turing, 1937).

L'informatique naît en 1946 avec le premier ordinateur (ENIAC) qui remplace le calcul analogique par le calcul numérique. Il ne possède pas encore de mémoire de masse, et sa mémoire vive est formée de près de 18 000 tubes à vide. La programmation est externe et se réalise à l'aide d'un câblage complexe. Les entrées et sorties se font à l'aide de cartes perforées utilisées jusqu'au début des années 1980. L'ordinateur évolue avec l'apparition des transistors grâce aux travaux de John Bardeen (1908-1991), William Shockley (1910-1989) et Walter Brattain (1902-1987), et des bandes magnétiques qui permettent de stocker l'information.

En 1951 apparaissent les premiers ordinateurs commercialisés (UNIVAC). Ils contiennent un programme interne qui permet de transformer les instructions du programmeur en instructions machine. L'entrée des données se fait à l'aide de cartes perforées ou de bandes magnétiques. Les défauts majeurs de ces machines sont liés à la fragilité des tubes à vide, et au langage machine qui est une suite de nombres binaires correspondant aux états électroniques de la machine. A partir de 1954, les mémoires magnétiques remplacent les autres formes de mémoire et en 1956 IBM commercialise les premiers disques durs.

Une deuxième génération d'ordinateurs voit le jour dans les années 1960 grâce à l'utilisation de transistors à la place des lampes à vide ce qui permet une miniaturisation des ordinateurs, une consommation moindre d'électricité et une meilleure fiabilité. Le développement des langages de troisième génération tel que FORTRAN ou BASIC permet d'écrire un programme de traitement des données indépendamment de la machine et d'éviter l'utilisation du langage machine ou le langage assembleur.

La troisième génération d'ordinateurs permet de faire plusieurs tâches à la fois. Elle repose sur l'utilisation de circuits intégrés qui doublent les fonctions des transistors fixés dessus, en réduisent le nombre et permettent de diminuer le coût. Entre 1960 et 1970 le nombre de transistors sur une puce passe de 10 à plus de 5 000, ce qui développe à proportion la puissance des ordinateurs.

ARPANET le précurseur d'Internet, commence à être développé en 1960, et en 1969 il relie pour la première fois 4 ordinateurs. A cette époque les interfaces télétypes remplacent les cartes perforées et permettent de communiquer avec un ordinateur grâce au texte. A la fin des années 1960, la souris et la visualisation des données sur écran inventées par Douglas Engelbart (1925 -) font leur apparition, et permettent d'améliorer l'interface homme-machine.

Dans les années 1970 apparaît la quatrième génération d'ordinateurs. Le module de contrôle tient désormais sur une seule puce la *Central Processing Unit* (CPU) avec plus de 5 000 transistors. Les télétypes sont progressivement remplacés par les écrans vidéo comme terminaux d'affichage. Dans un premier temps, l'interface est en ligne de commande, puis dans les années 1980 apparaissent les premières interfaces graphiques inventées par Xerox. Dans les années qui suivent, l'innovation consiste surtout en une miniaturisation des processeurs et l'augmentation de leurs capacités de calcul suivant la loi de Moore qui prédit un doublement de leurs performances tous les deux ans, mais qui commence à être confrontée aux limites physiques. Nous sommes ainsi passés de 2 300 transistors sur le processeur Intel 4004 en 1971 à 1 170 000 000 sur un processeur Intel Core i7 en 2010, ce qui permet de passer de 0,06 à 147 600 millions d'instructions par seconde.

L'amélioration des technologies de stockage magnétique permet de passer du stockage de quelques kilos octets par seconde à plusieurs téraoctets (10^9 kilos octets). Dans les années 1990 le Web fait son apparition et connaît une croissance exceptionnelle. Nous atteignons aujourd'hui plus d'un milliard d'utilisateurs de par le monde. Les vitesses de communication n'ont pas cessé de s'améliorer, et l'augmentation des capacités de calculs et de stockage, permet de développer les outils, du plus simple au plus compliqué, nécessaires et utiles à la recherche scientifique.

1.5.3.2 **Biologie moléculaire et génie génétique, apparition de la bioinformatique**

Au début des années 1970, la structure de la molécule d'ADN est connue, ainsi que son mode de réplication et d'expression. Malgré la possibilité de définir à peu près la position des gènes sur les chromosomes, on ne sait toujours pas isoler un gène, séquencer l'ADN qui le constitue, et en faire une étude approfondie. C'est alors que survient une découverte capitale faite par le microbiologiste suisse Werner Arber : celle des **enzymes de restriction** appelés aussi « ciseaux à ADN ». Il découvre que la bactérie E. coli produit une enzyme spécifique (EcoB) capable de morceler de l'ADN étranger provenant de bactériophage. Il propose de considérer que ces bactéries sont capables de se protéger de l'ADN étranger en le fragmentant par un mécanisme qui identifie des sites spécifiques sur cet ADN (Arber & Linn, 1969). Il faut peu de temps aux scientifiques pour découvrir une seconde enzyme de restriction appelée EcoK (Meselson & Yuan, 1968). D'autres études, par exemple celle de Hamilton O. Smith, permettent de découvrir des enzymes de restriction chez d'autres espèces bactériennes et confirment les hypothèses d'Arber en montrant que ces enzymes sont extraordinairement

sélectives dans le choix de la position des coupures (H. O. Smith & Kelly, 1970 ; H. O. Smith & Wilcox, 1970). Les enzymes de restriction cherchent une suite spécifique de bases (de 4 à 6 bases de long) sur la molécule d'ADN pour y effectuer une coupure. Plus de 800 enzymes de restriction différentes sont découvertes chez les bactéries, qui reconnaissent et coupent plus de 100 sites de restriction différents. Ainsi, la découverte au début des années 1970 des enzymes de restriction ou « ciseaux à ADN » capables de découper avec précision une molécule d'ADN, et des ligases capables de recoller avec précision des fragments d'ADN, donne naissance au génie génétique. Ces outils permettent d'établir une cartographie précise des génomes.

L'Américain Paul Berg utilise les enzymes de restriction et les ligases de manière ciblée et réussit à lier en laboratoire des morceaux d'ADN d'une bactérie avec celui d'un virus. Il fabrique ainsi, pour la première fois dans l'histoire, de l'ADN dit recombinant qui rend possible la **transgénèse** et ouvre la voie à la création d'organismes génétiquement modifiés (**OGM**). Le génie génétique comporte aujourd'hui de nombreuses techniques de manipulation d'acides nucléiques (ADN et ARN).

L'apparition du séquençage des protéines au milieu des années 1950 a conduit à élaborer les premières matrices permettant la compréhension des relations évolutives (M. O. Dayhoff & Ledley, 1962). C'est l'élément fondateur du domaine de la bioinformatique. Ensuite apparaît la construction des arbres phylogénétiques (W. M. Fitch & Margoliash, 1967). Le développement d'algorithmes informatiques a permis de construire l'alignement global de deux protéines (Needleman & Wunsch, 1970). Le stockage des informations des protéines, initié par Dayhoff s'étoffe très vite.

En 1977 apparaissent les méthodes de séquençages de l'ADN. Ces nouvelles techniques offrent des perspectives impressionnantes. Les connaissances au niveau des gènes gagnent en précision et nécessitent de nouvelles méthodes d'analyses. Les nombreuses données de séquençage sont archivées dans des bases de données informatisées. Le développement du séquençage apporte un souffle nouveau à la biologie moléculaire et laisse présager une augmentation très rapide des connaissances sur l'information stockée au niveau des gènes. La discipline de la « bio-informatique » croît rapidement, notamment par la publication des séquences rendue publique dès 1982, par l'utilisation d'algorithmes qui permettent des alignements locaux (T. F. Smith & Waterman, 1981), et par la mise en place d'algorithmes de recherche dans les bases de données.

La technique de la PCR découverte quelques années plus tard accroît les perspectives ouvertes par la découverte du séquençage ; elle a un effet démultiplicateur sur le potentiel de décodage de l'information génétique.

1.5.4 L'ère de la génétique

La connaissance de la position des gènes dans les génomes et de leurs séquences grâce au développement du séquençage, permet d'observer dans le détail les bases élémentaires à l'origine de la vie. Ces nouvelles connaissances appuyées sur des observations concrètes produisent une grande effervescence intellectuelle. Elles permettent d'entrevoir l'élucidation rapide des nombreuses questions que se posent les biologistes, et en particulier celle de l'origine de la vie.

1.5.4.1 ADN et hérédité

Chez les Eucaryotes (*Eukaryota*) le support de l'hérédité est la molécule d'ADN sur laquelle sont disposés les gènes qui contiennent les informations, transcrites en ARNm puis en protéines par la machinerie cellulaire. La molécule d'ADN est une molécule stable qui peut être répliquée quasiment à l'identique dans la lignée des cellules somatiques lors du phénomène de mitose. Dans la lignée des cellules germinales, l'ADN permet la transmission des caractères biologiques à la descendance. Trois mécanismes sont proposés pour décrire cette transmission :

- La transmission mendélienne : Elle permet grâce à la méiose, un brassage intra et interchromosomique, des chromosomes homologues produisant une très grande diversité du matériel génétique contenu dans les gamètes.
- La transmission dite non mendélienne : elle permet la transmission de séquences d'ADN provenant d'un seul parent. Par conséquent il n'y a pas de modification de la séquence parentale lors de la transmission. Les exemples les plus classiques sont la transmission d'organites comme les chloroplastes et les mitochondries.
- La transmission de caractères épigénétiques : elle est encore largement débattue et de grandes revues scientifiques s'en font l'écho. Le mécanisme encore inconnu consiste en une transmission de modifications du phénotype, sans modification de la séquence d'ADN mais avec modification probable de sa structure par méthylation des cytosines ou des protéines histones.

1.5.4.2 La transmission mendélienne, moteur de diversité

La méiose a une importance considérable dans le brassage chromosomique et la transmission de l'hérédité dans toute sa diversité. En amont, la reproduction sexuée qui trouve ses origines dans la méiose, permet à ce niveau un brassage génétique, et de fait, donne tout son sens au phénomène de méiose.

Pour bien comprendre la diversité qui existe d'une génération à l'autre, prenons l'exemple de l'homme et le brassage chromosomique dans son ensemble. Chez l'homme, la fécondation est le résultat de l'union de deux cellules germinales haploïdes appelées gamètes (ovule + spermatozoïde) contenant chacune 23 chromosomes. Ces gamètes sont formés dans les organes génitaux, par un mécanisme appelé méiose à partir de cellules diploïdes (zygotes) formés de 46 chromosomes ou 23 paires (23 chromosomes de la mère et 23 du père). Pour former une cellule haploïde de 23 chromosomes, chacun de ces chromosomes est sélectionné au hasard parmi les deux chromosomes homologues parentaux possibles. Cette sélection correspond à un brassage interchromosomique avec un potentiel de 2^{23} gamètes différents. Au cours de la fécondation, la variabilité augmente encore. En effet, l'union d'un ovule parmi 2^{23} possibles et d'un spermatozoïde parmi 2^{23}, porte le nombre de zygotes différents possible à plus de 2^{46}, soit plus de 70 000 milliards de variations possibles.

D'autre part, le phénomène d'enjambement (*crossing-over*) qui se déroule durant la prophase I de la méiose accentue considérablement cette variabilité et assure un brassage intrachromosomique. Chez l'homme le nombre moyen d'enjambements est de 42 pour les femmes et de 27 pour les hommes (Broman, Murray, Sheffield, White, & Weber, 1998 ; Lynn, Ashley, & Hassold, 2004). Ce qui donne plus d'un enjambement par bivalent (2 chromosomes homologues, 4 chromatides). Théoriquement le nombre de recombinaisons possible est pratiquement infini, mais il faut remarquer que les chromosomes ont des zones préférentielles, appelées « *hotspot* », où le taux de recombinaisons est jusqu'à mille fois plus élevé que celui des zones environnantes. Le nombre d'enjambements varie beaucoup selon les individus et selon les espèces. La méiose est déterminante pour l'augmentation de la variabilité génétique des populations.

1.5.4.3 Les mutations comme moteur de l'évolution

Malgré la grande stabilité de l'ADN et notamment sa structure en double hélice, les séquences d'ADN peuvent être modifiées. Ces modifications appelées **mutations** sont le moteur de la diversité, et sont à l'origine de la biodiversité. Les mutations par le biais de la sélection

naturelle, se fixent ou non au sein de la descendance.

Les mutations peuvent résulter du processus d'enjambement lors de la méiose, ou de cause diverses, par exemple des erreurs de l'ADN polymérase lors de la réplication de l'ADN au cours du processus de la mitose. D'autres phénomènes mutationnels existent tels que les mutations spontanées, les transferts horizontaux de gènes étrangers ou les rétropositions, et ne sont pas soumis aux processus de mitose ou de méiose. Toutes ces mutations, lorsqu'elles atteignent les cellules germinales peuvent être transmises à la descendance. Voyons plus en détail quels sont les types de mutations, leurs causes et leurs conséquences.

1.5.4.3.1 Les mutations lors de la méiose

La méiose qui est à l'origine d'un brassage génétique est un moteur mutationnel important. En effet, lors de la méiose deux groupes de mutations peuvent être observés : des mutations impliquant le nombre de chromosomes transmis lors du brassage interchromosomique, et des mutations concernant le contenu génique des chromosomes lors du brassage intrachromosomique.

Des erreurs lors de la répartition des chromosomes peuvent entraîner des trisomies partielles, ou totales lorsque tous les chromosomes sont représentés de manière surnuméraire dans le zygote, par une duplication complète du génome. La duplication complète du génome chez les Eucaryotes semble avoir joué un rôle primordial dans leur évolution. On retrouve ainsi une polyploïdie dans de nombreuses plantes tel l'épeautre qui est hexaploïde. Il semble que la polyploïdie de l'ancêtre des dicotylédones soit à l'origine de leur radiation (Jaillon *et al.*, 2007). Des polyploïdisations sont également survenues chez les métazoaires comme chez la levure il y a plus de 100 millions d'années (Kellis, Birren, & Lander, 2004), ou à deux reprises chez l'ancêtre des Vertébrés (*Vertebrata*) (Makalowski, 2001). Chez ces derniers la polyploïdisation est suivie d'une diploïdisation ramenant le nombre des chromosomes à 2n chromosomes. Malgré la diploïdisation, de nombreux gènes plus ou moins modifiés se sont fixés durant l'état transitoire de polyploïdisation.

Lors du brassage intrachromosomique, les recombinaisons peuvent entraîner de nombreux types de mutations. Elles peuvent entraîner des duplications (redoublements), des inversions, des délétions (suppressions) lors de recombinaisons homologues inégales, et des insertions et translocations (échanges) lors de recombinaisons de chromosomes non homologues (Illustration 6). La taille des séquences impliquées est variable et les séquences peuvent contenir de nombreux gènes.

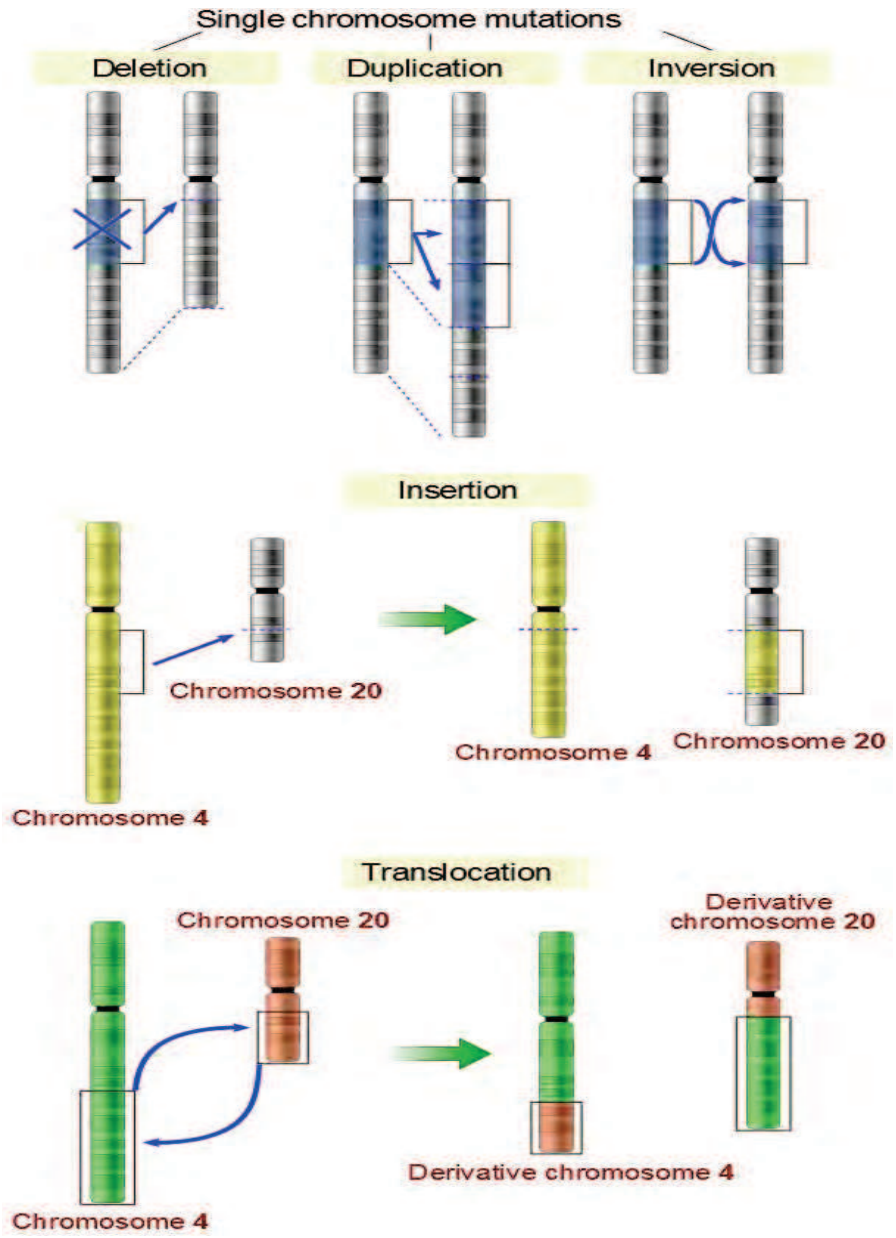


Illustration 6 : Mutations possibles lors de recombinaisons

Source : http://upload.wikimedia.org/wikipedia/commons/2/26/Chromosomes_mutations-en.svg

1.5.4.3.2 **Les mutations ponctuelles**

Les mutations ponctuelles peuvent être le résultat d'erreurs de réplication. La réplication est très fidèle, mais on compte environ 1 erreur pour 100 000 nucléotides répliqués. Grâce à différents systèmes de réparation, la fidélité de réplication s'élève finalement à 1 erreur pour 10 000 000 nucléotides répliqués. Ces erreurs peuvent engendrer des insertions, des délétions ou des substitutions. De courtes séquences répétées appelées séquences **microsatellites** peuvent induire des erreurs lors de la réplication et se traduire par l'insertion de ces séquences en très grand nombre dans certaines zones du génome.

1.5.4.3.3 **Les mutations spontanées**

Les mutations spontanées peuvent se fixer après réparation de séquences altérées. Ces altérations sont causées par une instabilité moléculaire et peuvent être dues à des réactions avec des éléments naturellement présents dans l'organisme. Il peut y avoir des pertes de bases menant à la création de sites apurinique ou apyrimidique, des désaminations, des alkylations et des méthylations de bases azotées. Malgré un système de réparation performant, ces altérations peuvent engendrer des mutations, comme des substitutions ou des délétions.

1.5.4.3.4 **Les mutations induites**

Les agressions induites par des substances chimiques ou des radiations ionisantes, engendrent de nombreuses altérations comme l'hydroxylation de bases azotées ou des coupures de brins d'ADN. Les rayons UV peuvent également engendrer la création de dimères. Ces altérations peuvent être à l'origine de substitutions, d'insertions et de délétions.

1.5.4.3.5 **Les autre types de mutations**

Un autre type de mutation appelé transposition passe par des enzymes comme les **intégrases** ou les **transposases** qui permettent à certains éléments appelés **transposons** de se déplacer ou de se dupliquer et de s'insérer dans les génomes. Il existe trois types d'éléments transposables dont les deux principaux sont les transposons de type I et ceux de type II.

Les transposons de type I appelés également **rétrotransposons** sont formés par une reverse transcriptase à la manière d'un « copié collé » alors que les transposons de type II sont formés sur le type « couper coller ». Leurs caractéristiques sont liées à la présence d'enzymes nécessaires aux rétrotranspositions dans leurs séquences. Les rétrovirus sont des rétrotransposons qui se différencient par la possession de gènes spécifiques comprenant notamment des gènes codants pour les protéines des capsides. Il existe également de courts éléments transposables non autonomes qui utilisent le matériel d'autres éléments transposables pour se multiplier et se déplacer. Ce matériel peut également affecter des séquences non assimilées à des éléments transposables. Lorsque ce phénomène affecte un ARN épissé, sa réinsertion sous forme d'ADN dans le génome donne naissance à des gènes dits processés situés la plupart du temps en dehors du locus d'origine. De manière générale, les éléments transposables contribuent à des réarrangements de portions du génome de types inversions, translocations et duplications. Lorsqu'une séquence d'un organisme est déplacée et insérée dans un autre organisme de même espèce ou d'espèce différente on parle alors de transfert latéral.

1.5.4.4 **Conséquences des mutations**

L'évolution se manifeste par l'apparition de mutations dans les cellules germinales transmises aux descendants. L'évolution est régie par la concomitance de mutations aléatoires et de la sélection naturelle à laquelle ces mutations sont soumises. Par le biais de la sélection naturelle, les mutations qui procurent un avantage sélectif tendent à se fixer, alors que celles qui sont défavorables voire délétères seront éliminées. Les mutations qui n'influencent pas la valeur sélective des individus sont appelées **mutations neutres** et dérivent sans sélection de façon aléatoire. Elles peuvent ainsi se fixer ou disparaître.

Les conséquences des mutations peuvent être variées et présenter des effets différents selon le niveau où est faite l'observation. Par exemple, une erreur de réplication peut causer la substitution d'un nucléotide au sein d'un gène. Cette substitution peut être analysée comme la cause d'un changement d'acide aminé ; celui-ci peut être caractérisé comme la cause d'un changement fonctionnel de la protéine issue du gène ; le changement fonctionnel peut être vu, à son tour, comme la cause d'un changement comportemental de l'individu. L'erreur de réplication initiale peut se traduire en bout de chaîne par un changement comportemental. Ainsi, selon le niveau d'observation, un même événement peut être interprété différemment. Pour bien comprendre les lois qui régissent la biologie, il est important de mener des observations à différents niveaux. Tout particulièrement dans les recherches sur l'évolution, où les observations portent sur une échelle de temps souvent considérable, et sur des phénomènes de niveaux différents (par exemple l'apparition ou la disparition d'un gène, d'un organe, d'une espèce).

Les mutations les plus étudiées sont celles qui ont des conséquences phénotypiques. Elles peuvent s'observer aux niveaux éthologique, morphologique, physiologique, moléculaire, etc. Les effets de certaines mutations ne peuvent se voir que sous des conditions particulières de température, d'hydratation, etc. D'un autre côté, il existe des mutations silencieuses qui n'ont aucun effet phénotypique observable.

L'étude des mutations sur une séquence entière précise, permet de déterminer l'importance et le type de la séquence étudiée, de connaître l'importance des trinucléotides au sein des gènes. Elle permet également d'analyser la dérive génétique, de déterminer les horloges moléculaires, etc. Le Tableau 1 ci-dessous récapitule les mutations qui peuvent être observées au niveau des séquences, avec les causes qui peuvent leur être imputées.

<table>
<tr><th rowspan="2">Causes mutationnelles</th><th colspan="2">Mutations</th></tr>
<tr><th>Conséquences sur les séquences</th><th>Conséquences au niveau des gènes ou des domaines</th></tr>
<tr><td>Erreurs de réplication
Modifications spontanées
Agressions (substances chimiques ou radiations)</td><td>Substitution, insertion, délétion</td><td rowspan="2">Echange, transfert, fusion, perte, gain</td></tr>
<tr><td>Recombinaisons (enjambement inégal, rétroposition, transfert horizontal)
Agressions par radiation (cassure des brins d’ADN)</td><td>Duplication, inversion, délétion, insertion, translocation</td></tr>
</table>

Tableau 1 : Causes possibles des mutations et leurs effets au niveau des séquences et des gènes

L’étude du comportement évolutif des séquences d’ADN au travers de leurs mutations est très instructive. Les travaux effectués au sein du laboratoire se concentrent actuellement sur les séquences géniques qui semblent être les éléments apportant le plus d’informations à l’étude de l’évolution. Nous cherchons à comprendre les modalités de l’évolution des gènes (et des protéines lorsqu’elles sont traduites), à travers l’étude des mutations qui ont conduit à leurs modifications, leurs apparitions, leurs disparitions (Tableau 1), ainsi que des conséquences qui en découlent aux différents niveaux d’un organisme vivant. Dans de nombreux cas il est impossible de connaître les causes initiales de l’apparition des mutations.

1.5.4.4.1 Gain de gènes

Dans les années 1930, les travaux de cytologie menés sur les chromosomes par Haldane (Haldane, 1933) et Muller (Muller, 1935) amènent ces chercheurs à formuler l'hypothèse que de nouveaux gènes, apportant de nouvelles fonctions, pourraient apparaître par remodelage de copies de gènes. Cette nouvelle notion de recombinaison, qui conduit à l’émergence de la duplication des gènes, a été largement confirmée ensuite par des travaux de types moléculaires. Les connaissances s’affinent, et il est mis en évidence que les recombinaisons provoquées par un enjambement inégal peuvent engendrer des duplications de fragments de gènes, de gènes complets, de longs fragments d'ADN, voire des génomes complets par polyploïdisation (Conant & Wolfe, 2008 ; Van de Peer, Maere, & Meyer, 2009). Ces phénomènes qui créent de nouveaux gènes existent chez les procaryotes (Spencer, Susko, &

Roger, 2006) et les Eucaryotes (Wolf, Novichkov, Karev, Koonin, & Lipman, 2009). Depuis le travail d'Ohno (Ohno, 1970) il est largement admis que la duplication des gènes peut permettre à l'une des deux copies, d'acquérir une nouvelle fonction par néofonctionnalisation. Le duplicata peut également être préservé et augmenter l'expression du gène. Il est également proposé qu'après duplication d'un gène, sa fonction ancestrale soit partagée par les deux gènes par subfonctionnalisation (Force *et al.*, 1999). La néofonctionnalisation et la subfonctionnalisation sont souvent modelées par la sélection naturelle ou induites par une évolution neutre (Conant & Wolfe, 2008 ; Force *et al.*, 1999 ; Innan & Kondrashov, 2010).

L'étude des mécanismes de duplication de l'ADN a permis en 1991 de découvrir un autre mécanisme de duplication appelé **rétroposition**. Ce mécanisme utilise un ARN messager mature provenant de la transcription d'un gène « source », qui par transcription reverse donne un ADN complémentaire (ADNc) qui par la suite s'insère dans le génome. Ces séquences appelées **rétroposons** sont facilement reconnaissables car elles ne possèdent pas d'introns et peuvent comporter une queue polyA. Les enzymes nécessaires pour la rétroposition, sont codés par différents éléments rétroposables existant chez les espèces. Par exemple, chez les Mammifères (*Mammalia*) ces enzymes peuvent provenir des rétroposons LINE-1. D'autres mécanismes induisent l'apparition de nouveaux gènes. Ils peuvent naître par co-option de génomes parasites génomiques (Feschotte & Pritham, 2007 ; Volff, 2006) comme ceux des rétrovirus endogènes, ou même, comme le montre une étude récente (Heinen, Staubach, Häming, & Tautz, 2009), à partir de séquences précédemment non fonctionnelles.

Il existe également le **transfert latéral de gènes**, appelé aussi **transfert horizontal de gènes** (THG), processus par lequel un organisme incorpore le matériel génétique d'un autre organisme. Depuis la découverte des THG (Ochiai, Yamanaka, Kimura, & Sawada, 1959), des THG ont été décrits dans de nombreux phyla et on a pu mesurer leur importance dans les mécanismes de l'évolution. L'importance de ce phénomène est bien établie chez les bactéries (Boucher *et al.*, 2003) et chez les Archées (Garcia-Vallvé, Romeu, & Palau, 2000). Chez les Eucaryotes, la perception de ce phénomène est encore cantonnée à des événements liés à une endosymbiose (transfert de gènes mitochondrial, chloroplastique ou plasmidique au génome nucléaire), ou à du parasitisme (transfert du gène d'un parasite à son hôte). Mais de récentes études montrent que le THG est un phénomène bien présent chez les Eucaryotes et se retrouve dans différentes espèces qui n'ont pas forcément un lien hôte-parasite. Des interactions plus simples peuvent suffire comme chez le mollusque *Elysia* lors de la consommation d'une algue (Rumpho *et al.*, 2008), ou chez les plantes lors de greffes (Stegemann & Bock, 2009).

Certaines études décrivent de nombreux THG chez des organismes provenant de partenaires de taxa divers, comme les rotifères bdélloides qui ont acquis des gènes bactériens, de plantes, de champignons (Gladyshev, Meselson, & Arkhipova, 2008), ou les nématodes qui ont acquis des gènes provenant de microorganismes avec lesquels ils ont des interactions trophiques ou symbiotiques (Haegeman, Jones, & Danchin, 2011). Chez les Eucaryotes multicellulaires les THG semblent improbables à cause de la mise à l'abri des cellules germinales (Keeling & Palmer, 2008). Mais la mise en évidence de THG chez la souris (Pace, Gilbert, Clark, & Feschotte, 2008) et chez les rotifères bdélloides (Gladyshev *et al.*, 2008) jette une lumière nouvelle sur l'évolution des Eucaryotes unicellulaires. L'importance des THG dans l'évolution des Eucaryotes pluricellulaires demeure encore obscure et mérite des investigations supplémentaires.

1.5.4.4.2 **Perte de gènes**

La perte de gènes a longtemps été un événement non intuitif. L'élément déterminant qui a déclenché l'intérêt de ce type d'événement est la découverte des **duplications** de gènes. En effet, quelques années après la description des duplications, Ohno (Ohno, 1972) propose qu'une partie des gènes dupliqués peut devenir de l'ADN non codant par désactivation de ces gènes alors appelés **pseudogènes**. Le processus de formation de pseudogènes est appelé **pseudogénisation**. Un pseudogène est donc un gène non fonctionnel dont la séquence est voisine de celle du gène fonctionnel.

Le **pseudogène** peut être défini comme une copie non fonctionnelle d'un gène que des mutations empêchent de s'exprimer et qui subsiste dans le génome. Cette hypothèse permet d'expliquer en partie la présence d'ADN non codant au sein des génomes. Pour la première fois la perte de gènes n'est plus synonyme d'un impact grave sur la pérennité d'un organisme. Les gènes fonctionnels perdus par un tel processus sont considérés comme des gènes non essentiels aux organismes parce que leurs fonctions sont redondantes. Il faut attendre le début des années 90 pour observer les premières pseudogénisations touchant des gènes connus pour jouer un rôle important au sein des organismes dans lesquels ils sont encore présents. Chez les espèces concernées, la perte des fonctions associées à ces pseudogènes (gènes perdus) se manifeste par des changements phénotypiques (Nishikimi, Fukuyama, Minoshima, Shimizu, & Yagi, 1994 ; Wu, Lee, Muzny, & Caskey, 1989). Les études d'**endosymbiontes** mettent en évidence de nombreuses pertes de gènes et de nombreux pseudogènes de gènes anciennement fonctionnels (Wakasugi *et al.*, 1994).

Les progrès techniques permettent de séquencer de nombreux génomes entiers. La

comparaison des génomes complets permet de mettre en évidences de nombreuses pertes de gènes pouvant se traduire par la réduction des génomes au travers d'événements amenant à des délétions. Ainsi, l'étude de pertes de gènes anciennement fonctionnels devient primordiale pour comprendre la plasticité des génomes et l'impact de ces événements sur l'évolution des espèces. La compréhension de ces événements est le sujet de ma thèse.

1.5.5 L'ère de la génomique

Le séquençage de génomes entier a été initié en 1976 par celui du génome ARN d'un bactériophage (Fiers *et al.*, 1976), et a pris un grand essor à partir de la fin des années 90 avec le séquençage de génomes ADN « complexes ». Ainsi, le génome de nombreux organismes dans les principales branches du vivants a été séquencé : génomes de Bactéries (*Bacteria*) (Fleischmann *et al.*, 1995), d'Archées (*Archaea*) (Bult *et al.*, 1996) et d'Eucaryotes (Goffeau *et al.*, 1996).

Les avancées techniques en matière de séquençage ont permis une baisse importante de leur coût et a contribué à un large développement de projets de séquençage de génomes complets. Le nombre de séquençages a explosé dans le monde (Illustration 7). Actuellement, près de 2907 génomes ont été séquencés et leurs séquences déposées dans des bases de données (BD) publiques (Pagani *et al.*, 2011). Il existe également 8 565 projets de séquençage de génomes en cours dans diverses structures, par exemple le *Joint Genome Institute* (JGI), le *J. Craig Venter Institute* (JCVI), le *Wellcome Trust Sanger Institute*, ou encore le Genoscope qui est le centre national de séquençage pour la France. A titre d'exemple, le Génoscope a multiplié son volume de séquençage annuel par 40 et divisé ses coûts par 30 au cours de ces dix dernières années. En 2009, 57 génomes de Vertébrés sont disponibles et des projets sont proposés afin de séquencer dans un futur proche plus de 10 000 génomes de Vertébrés représentatifs de cette lignée (Haussler *et al.*, 2009). D'autres projets cherchent à déterminer le génome de nombreux individus d'une même espèce (The 1000 Genomes Project Consortium, 2010 ; Weigel & Mott, 2009).

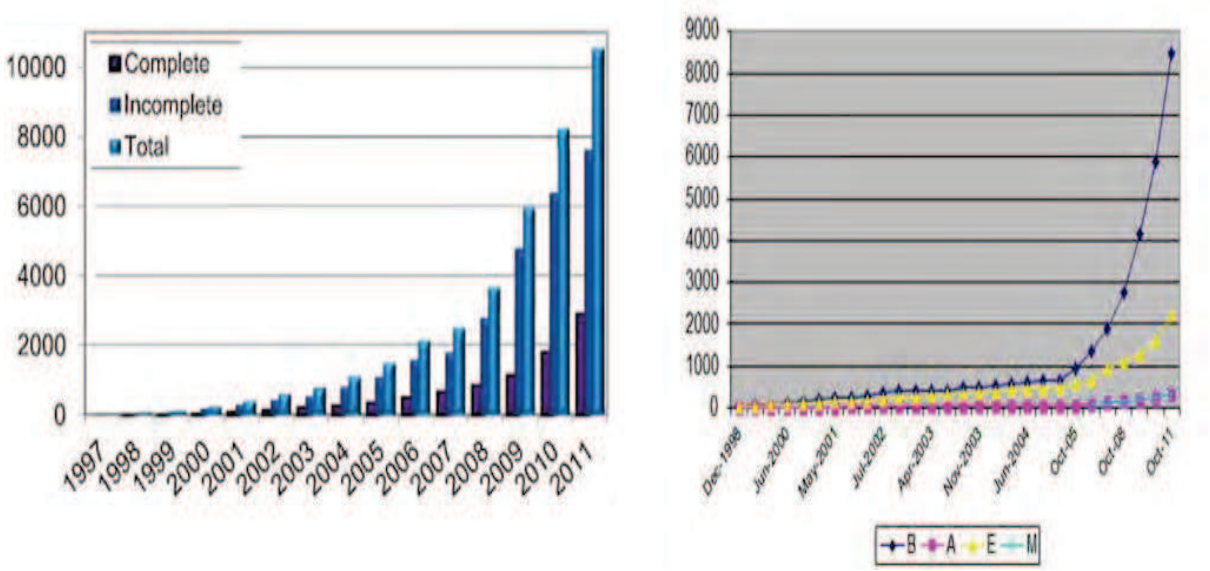


Illustration 7 : Statistique des génomes séquencés

Ces observations proviennent de The Genomes OnLine Database (GOLD) (Pagani *et al.*, 2011) *et représentent la période de la fin des années 90 à Septembre 2011*

A) Evolution du nombre de projets de séquençages

B) Evolution du nombre de projets de séquençages en fonction des groupes phylétiques.

B = Bactérie ; A=Archée ; E = Eucaryote ; M = Métagénome

Source : http://www.genomesonline.org/images/statistics/gold_s3.gif

Le séquençage à grande échelle a vu naître de manière concomitante des outils informatiques adaptés au traitement de l'énorme quantité d'informations produite. Des grandes structures se sont développées et organisées afin de stocker et pouvoir interroger les données produites. Ainsi les instituts du *National Institutes of Health* (NIH), de *l'European Molecular Biology Laboratory* (EMBL), et le *National Institute of Genetics* (NIG), se sont rapprochés pour créer l'*International Nucleotide Sequence Database* Collaboration (INSDC) qui permet le regroupement de toutes les séquences publiques existant au sein de leurs banques de données respectives : GenBank (Américain), EMBL (Européen) et DDBJ (Japonais) (Illustration 8). Ce regroupement permet une harmonisation du format des fichiers utilisés et des échanges réguliers entre les bases pour mettre à jour les données. En conséquence, l'interrogation de toutes les séquences disponibles dans les différents instituts est possible à partir de leurs interfaces Web respectives. Chaque groupe gère les mises à jour des séquences qu'il a créées. Toutefois, les serveurs des instituts n'ont pas la même présentation des données et les outils mis à disposition pour les l'analyser peuvent différer.

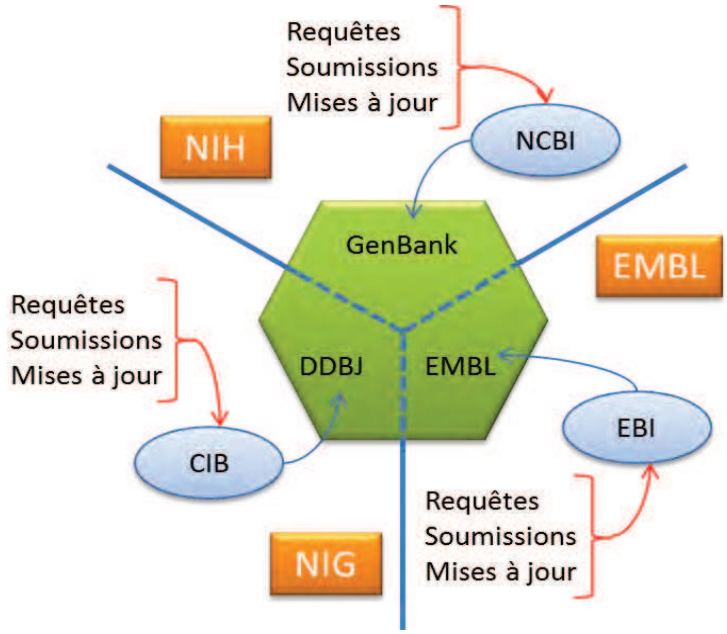


Illustration 8 : Interaction des trois grandes structures au sein de l'INSDC qui gère les séquences publiques

De nombreuses banques de données spécialisées (nucléotide, protéique, EST, etc.) ont vu le jour, et proposent de grandes quantités de données hétérogènes (séquences, annotations, etc). Ainsi, outre les banques de données de séquences nucléiques (GenBank, EMBL, DDBJ), il existe des banques qui s'orientent spécifiquement vers la collecte de séquences protéiques (SWISSPROT, UNIPROT, etc.), vers les séquences génomique (Ensembl, etc.), vers la structure 3D des macromolécules (PDB, etc.), etc. Toutes ces banques de données sont accessibles en ligne et mettent à disposition des outils permettant d'analyser et d'extraire les données intéressantes pour le biologiste. Elles ont permis à la bio-informatique de prendre son essor. Dans le monde entier, les laboratoires ont développé de nouveaux outils, et automatisé certains processus, pour aborder aux mieux leur sujet de recherche.

Le séquençage des génomes a suscité beaucoup d'engouement. Il doit apporter de nombreuses réponses aux différentes branches de la biologie. Mais l'analyse des génomes a mis à jour une grande complexité sous-jacente, faite de mutations, d'interactions, d'équilibres, etc. Il semble que nous soyons au début d'une période riche en développement des connaissances et en compréhension du fonctionnement du vivant. Cela ne sera possible qu'à travers le développement d'outils informatiques puissants et adaptés pour trier la montagne d'informations présentes dans les génomes. Les études globales au niveau du génome nous éclairent progressivement, peu à peu, sur la complexité de la vie et de l'évolution.

1.5.6 Les méthodes d'analyses

L'étude de l'évolution se base sur les connaissances contemporaines des génomes et implique une analyse descendante des données pour reconstruire les génomes ancestraux et inférer les caractères et les événements qu'ils ont connus. Les méthodes d'analyse comparative permettent d'effectuer ce travail. Elles se basent en particulier sur la phylogénie et la reconstruction de caractères ancestraux. Ces approches indispensables sont les biais utilisés dans l'analyse ascendante qui caractérise l'évolution.

La phylogénie

La phylogénie cherche à retracer les relations entre les êtres vivants actuels. Dans les arbres phylogénétiques les caractères ancestraux peuvent être définis à partir des caractères des feuilles, à partir de méthodes de reconstructions. Ainsi, il est possible de suivre l'histoire évolutive des caractères.

Les méthodes de reconstruction

Les processus de reconstruction initiés par Hennig (Hennig, 1966) utilisent la méthode de **parcimonie** qui repose sur le principe d'une analyse cladistique et qui utilise la notion de descendance et non de ressemblance. Cette méthode implique qu'au niveau des nœuds phylogénétiques existent des états ancestraux et dérivés. Depuis, de nombreuses méthodes de parcimonie ont été publiées telles celles de Dollo (Farris, 1977), de Sankoff (Sankoff, 1975 ; Sankoff & Rousseau, 1975) et de Mirkin (Mirkin, Fenner, Galperin, & Koonin, 2003).

Il existe également de nombreuses méthodes probabilistes pour la reconstruction. Par exemple la méthode du maximum de vraisemblance. Les méthodes probabilistes s'appliquent à des caractères pour lesquels une probabilité de transitions entre divers états est définie. Les méthodes statistiques sont mieux adaptées pour les études comportant de nombreux événements.

Ces méthodes d'analyses permettent la reconstruction et l'inférence :

- de structures protéiques et nucléotidiques
- d'expression
- de fonctions
- de caractéristiques physico-chimiques
- de caractéristiques anatomiques, morphologiques

Les connaissances des états dérivés et ancestraux permettent donc d'étudier des convergences

- phénotypiques
- transcriptionnelles
- d'expressions
- …

La génomique comparative

Avec l'arrivée du séquençage massif de ces dernières années sont apparues des méthodes de génomique comparative. Les méthodes de génomique comparative comparent des états, par exemple Présent/Absent. La comparaison de séquences génomiques amène par exemple à la détection d'indel (insertion ou délétion). Indel définit un trou dans une séquence dont on ignore précisément l'origine, le type d'événement qui l'a provoqué. La génomique comparative est utilisée pour la prédiction de fonctions, l'identification de sites fonctionnels, l'inférence de phylogénies, etc.

La phylogénomique comparative

La génomique comparative abrite entre autres la phylogénomique comparative. Les méthodes de phylogénomique comparative utilisent des méthodes de phylogénies comparatives qui existent depuis longtemps (Harvey & Pagel, 1991). Elles permettent de faire des corrélations de lien fonctionnel, de changements évolutifs, mais aussi de déterminer si un caractère est réellement un signal phylogénétique. Ainsi, lorsque des gènes apparaissent ou disparaissent plusieurs fois de manière indépendante, dans différentes espèces, il peut s'agir de gènes ayant un lien fonctionnel fort (Barker, Meade, & Pagel, 2007 ; Barker & Pagel, 2005).

La « génomique évolutive »

Les méthodes de « génomique évolutive » ont vu le jour plus récemment et permettent en utilisant les caractères Présents/Absents, de comparer non plus des états mais des événements. Par conséquent, au lieu d'observer des caractères Présents/Absents, il est possible d'observer par exemple des événements de pertes et des gains. Les approches évolutives cherchent à définir des liens entre différents événements apparus à des dates différentes. L'étude des événements et la compréhension de leurs temporalités nécessitent des analyses phylogénétiques. L'approche évolutive permet d'être plus proche de la réalité et d'observer avec précision les événements et leurs conséquences. Par exemple, lors de l'analyse de séquences génomiques, l'approche évolutive permet de définir non seulement un indel mais de gagner en précision et de définir s'il s'agit d'une insertion ou d'une délétion, et si cet

événement semble être la cause d'une pseudogénisation. La « génomique évolutive » nécessite l'utilisation des mathématiques pour comparer les probabilités d'événements. Au sein du laboratoire, nous voulons avancer dans les études évolutives, et c'est dans cette optique que l'équipe est liée à un laboratoire de mathématique.

1.5.7 Les relations entre les séquences

Pour comprendre l'évolution il faut pouvoir définir les relations qui existent entre les séquences. C'est un point central. Différents termes et leur définition permettent de décrire les relations entre les gènes :

Similarité : Des gènes sont dits similaires si leurs séquences se ressemblent. Lorsqu'aucun lien de parenté ne peut être mis en évidence entre des gènes, on dit qu'ils sont similaires.

Homologie : Des gènes sont homologues, lorsqu'ils partagent une origine commune.

Les termes qui suivent permettent de spécifier des relations plus complexes entre gènes homologues.

Orthologie : Les gènes sont orthologues lorsqu'ils sont issus d'un ancêtre commun. Ce sont donc des gènes provenant d'événements de spéciation.

Coorthologie : Des gènes issus de duplications dans une lignée spécifique et qui sont orthologues à un ou plusieurs gènes dans une autre lignée sont dit coorthologues. Ces coorthologues seront vus comme des inparalogues

Paralogie : Des gènes sont dit paralogues lorsqu'ils sont issus d'un événement de duplication.

Inparalogie : Les gènes qui ont subi une duplication dans une lignée spécifique après un événement de spéciation défini sont dits inparalogues.

Outparalogie : Les gènes résultant d'une duplication avant un événement de spéciation défini sont dits outparalogues.

Les définitions d'inparalogie et d'outparalogie sont relatives à un événement de spéciation défini, et n'ont par conséquent pas de signification absolue.

Xénologie : Des gènes homologues sont xénologues lorsqu'ils résultent d'un transfert horizontal de gènes.

Ohnologue : Des gènes paralogues sont ohnologues quand ils résultent d´un événement de duplication complète du génome.

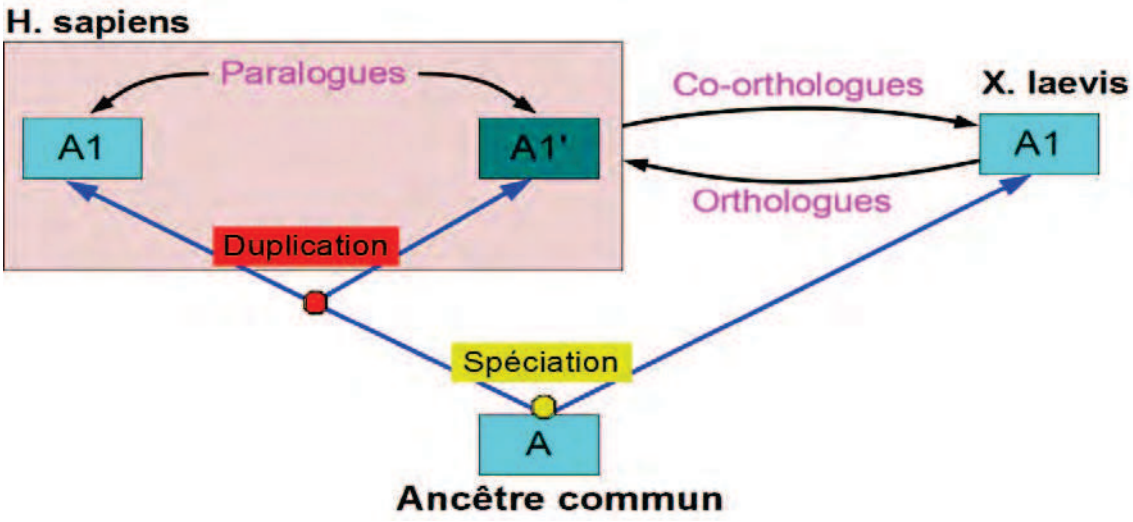


Illustration 9 : relations entre les gènes les plus usités

Ces termes, particulièrement ceux d'orthologie et de paralogie, correspondent à des concepts primordiaux en génomique évolutive. L'orthologie apparaît après des événements de spéciation tandis que la paralogie apparaît après des événements de duplication (Illustration 9). Ces termes permettent de décrire les relations qui existent entre les gènes et les traits évolutifs qui les concernent. Ces relations permettent, à partir de certains gènes annotés, d'inférer de manière fiable la fonction d'autres gènes non annotés en fonction des relations qui les lient. En effet, les gènes orthologues ont tendance à partager des fonctions similaires, ce qui est moins le cas des gènes paralogues (Eisen & Fraser, 2003).

Après duplication, il existe trois sortes de conséquences : la pseudogénisation, la néofonctionnalisation d'un des deux duplicatas, ou une subfonctionnalisation qui entraîne le partage de la fonction initiale entre les deux duplicatas. La définition des relations qui existent entre gènes est donc complexe (Illustration 10) et peut facilement entraîner des confusions. Ainsi, il faut bien discerner entre les termes :

- orthologie et coorthologie qui reflètent des relations entre les gènes d'espèces différentes,
- paralogie qui peut concerner des gènes dans des espèces différentes mais également au sein d'une même espèce,
- inparalogie et outparalogie qui concernent des gènes qu'au sein d'une même espèce.

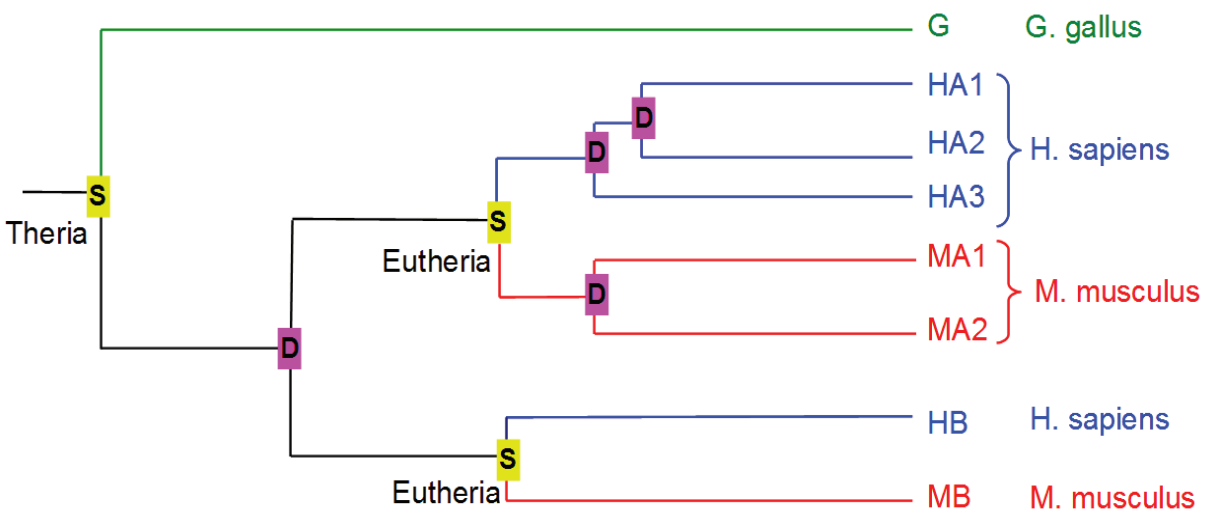


Illustration 10 : Relations complexes entre les gènes

L'arbre illustré ci-dessus représente les relations entre les gènes de trois espèces G. gallus, H. sapiens et M. musculus. Le gène G est orthologue à tous les autres gènes, réciproquement tous les autres gènes sont coorthologues au gène G. Les gènes HB et MB sont orthologues et ceux de HA(1,2,3) et MA(1,2) sont coorthologues entre eux. Les gènes HA(1,2,3) sont inparalogues entre eux, il en est de même pour les gènes MA(1,2). Les gènes HA(1,2,3) et HB sont des outparalogues comme le sont également les gènes MA(1,2) et MB quand on prend en compte seulement M. musculus et H. sapiens. En prenant en compte G. gallus, HA(1,2,3) et HB ainsi que MA(1,2) et MB sont vus comme des inparalogues.

Des difficultés existent dans l'utilisation des termes d'inparalogie et d'outparalogie car les relations peuvent varier en fonction du point de vue choisi, dans la mesure où ces termes sont toujours définis relativement à un événement de spéciation. Il en va de même pour les relations de gènes entre des espèces différentes (orthologie et paralogie) : ce ne sont pas des concepts transitifs. Par exemple, parmi les données de l'Illustration 10, le gène G est entre autres orthologue à HA1 et MB, mais HA1 et MB sont paralogues entre eux.

1.5.8 Le point sur les théories de l'évolution

Depuis la théorie de l'évolution proposée par Darwin, de nombreuses contributions ont été apportées. L'apport de nouvelles connaissances au début du XX^{ème} siècle a conduit à la **théorie synthétique de l'évolution** (néodarwinisme) qui est encore de nos jours la théorie prépondérante acceptée par la communauté scientifique. Les réflexions théoriques et les travaux scientifiques ont conduit à des controverses et des propositions tantôt compatibles avec la théorie néo-darwinienne de l'évolution (théorie neutraliste, théorie des équilibres ponctués et théorie hiérarchique), tantôt diamétralement opposées (théorie du gène égoïste). Faisons le point.

1.5.8.1 **Théorie neutraliste**

Cette théorie formulée par Motoo Kimura a considérablement modifié la conception du processus d'évolution. Pour cet auteur, le rôle central de l'évolution est le hasard et non plus la nécessité qui est la base de la théorie Darwinienne (Kimura, 1968). Cette théorie explique que l'évolution se fait au niveau des gènes soumis constamment à des mutations génétiques qui s'avèrent neutres en regard de la sélection naturelle. Les mutations sont le fruit du hasard et se produisent chez toutes les espèces à un rythme similaire, quel que soit l'environnement. La majorité des mutations étant sélectivement neutres, elles représentent des possibilités génétiques qui peuvent devenir par chance, des réalités morphologiques. Une importante idée subsidiaire à cette théorie est que l'évolution ne conduit pas forcément à la complexification des organismes. Cette théorie ne remet pas en cause le rôle de la sélection naturelle mais la relègue à un second plan. Les modèles mathématiques basés sur cette théorie permettent d'expliquer de nombreux phénomènes observés, mais elle repose cependant sur des bases qui ne sont que partiellement validées. Par exemple, on sait que le taux de mutations n'est pas constant et qu'il peut être modifié dans des conditions de stress ; on sait aussi que la taille des populations est rarement constante. Malgré tout, cette théorie est une contribution intéressante à la modélisation et à la compréhension de l'évolution.

1.5.8.2 **Théorie des équilibres ponctués**

Niles Eldredge et Stephen Jay Gould avancent en 1972 la théorie dite des « équilibres ponctués », selon laquelle l'évolution des espèces est régie par de longues périodes de stabilité et de courtes périodes d'intense activité évolutive (spéciation) (Stephen Jay Gould, 2012). Cette théorie va à l'encontre de l'évolution graduelle prônée par la théorie néodarwinienne mais ne s'oppose pas à la théorie Darwinienne. Elle l'élargit et permet d'expliquer le manque d'individus intermédiaires dans les registres fossiles de l'évolution des espèces. Gould explique que les phénomènes évolutifs sont rapides à l'échelle du temps géologique mais lents et progressifs à l'échelle humaine. Le temps de spéciation des espèces serait suffisamment faible pour se retrouver dans une même strate géologique, expliquant ainsi l'absence de traces dans les archives géologiques (Stephen Jay Gould, 1993).

1.5.8.3 **Théorie hiérarchique**

Récemment Stephen Jay Gould révise et élargit encore plus la théorie darwinienne en présentant une « théorie hiérarchique de la sélection » (S.J. Gould, 2002). La conception darwinienne classique défend une sélection seulement centrée sur l'individu ; la théorie hiérarchique prend en compte de nombreux autres niveaux : les espèces en tant qu'entités individuelles, les organismes, et au plus bas niveau dans la hiérarchie, les gènes. Les gènes sont singuliers car ils ont la propriété d'enregistrer les changements.

1.5.8.4 **Théorie du gène égoïste**

En 1978 Dawkins présente la théorie du gène égoïste qui centre les modalités de l'évolution au niveau des gènes. Il considère que la lutte pour la survie des espèces se fait à un niveau très bas, celui des molécules. En conséquence, Dawkins présente les organismes que nous observons comme des robots aveuglément programmés pour assurer la survie des molécules égoïstes que seraient les gènes (Dawkins, 2003). Il soutient que la sélection des organismes et des populations ne l'emporte jamais sur la sélection des gènes. Cette perspective centrée sur le rôle du code génétique est intéressante mais occulte l'importance des autres processus, telle la sélection naturelle, qui n'opèrent pas directement sur les gènes.

1.5.8.5 **Synthèse**

En dépit des les nombreuses controverses suscitées, la théorie de l'évolution proposée par Darwin il y a 150 ans, riche d'interprétations, s'est étoffée et a donné naissance à de nombreuses théories, dont le néodarwinisme. C'est actuellement la plus partagée par la communauté scientifique. Malgré les hypothèses récentes qui se focalisent sur des aspects précis (les gènes, les espèces, la sélection, l'adaptation, les changements graduels ou soudains), il est encore difficile de décrire avec précision la réalité sous-jacente. Elle semble si complexe que nous en sommes peut-être seulement à l'étape qui nous permet de voir la partie émergée de l'iceberg. L'évolution est source de débat, mais la communauté scientifique s'accorde sur l'enjeu majeur qu'elle représente pour les différents domaines de la biologie. Elle permet de donner un sens aux observations faites depuis la base du vivant, les gènes, jusqu'au biome le plus large représenté par la biosphère. L'étude de l'évolution reste un défi car elle nécessite un large champ d'investigation, et fait intervenir de nombreuses disciplines : la paléontologie, la biologie moléculaire, l'informatique, les mathématiques, l'écologie etc.

2 Le concept de perte de gènes

Durant de nombreuses années le phénomène de pertes de gènes n'a pas suscité l'attention, tant était forte et persistante l'image de la complexification croissante des organismes. L'idée qu'il existe des pertes de gènes germe dans les années 50 : on commence à imaginer que la délétion des gènes pourrait apporter des changements phénotypiques (Race, Sanger, & Selwyn, 1951). Ces changements dans les organismes étudiés semblent anormaux et engendrer des handicaps (Jackson, Macleod, & Krauss, 1959 ; Ose & Bush, 1962 ; Race *et al.*, 1951 ; Sherman & Slonimski, 1964). Il faut attendre les travaux de Ohno (Ohno, 1972) pour comprendre que le processus de perte de gènes n'engendre pas forcément un phénomène handicapant pour l'organisme. Il propose le concept de pseudogénisation, défini comme une désactivation de gènes par accumulation de mutations pouvant mener à terme à la disparition complète des signaux d'origine. Il fait l'hypothèse que la majorité des gènes dupliqués sont pseudogénisés, en raison de la redondance des fonctions et du relâchement sélectif que cela provoque. L'hypothèse est scellée quelques années plus tard par la découverte chez *Xenopus laevis* du premier pseudogène, lors de l'étude de l'ADN codant pour l'ARN 5S, une des molécules qui compose la grande sous-unité du ribosome (Jacq, Miller, & Brownlee, 1977). Par la suite, de nombreux pseudogènes sont découverts dans la majorité des organismes, des bactéries jusqu'aux animaux en passant par les plantes et les champignons (Fsihi *et al.*, 1996 ; Loguercio & Wilkins, 1998 ; Ramos-Onsins & Aguadé, 1998 ; Wirth, Leh-Louis, Potier, Souciet, & Despons, 2005).

Pendant des années, la perte de gènes est conçue comme un phénomène ayant peu d'impact sur l'évolution des espèces, si ce n'est sur la taille des génomes par l'accumulation d'ADN inutile lors de pseudogénisations. En effet, la pseudogénisation pourrait expliquer la quantité d'ADN non codant retrouvée dans les génomes.

Un nouvel engouement apparaît lors de la découverte de pertes, de gènes bien établis au sein des génomes, liées à des changements phénotypiques importants au sein des espèces. L'apparition des séquences complètes de génomes permet, grâce à la génomique comparative, d'observer de nombreuses pertes de gènes dans différentes lignées. Une diminution drastique de la taille de certains génomes est également mise en évidence. Ainsi la perte de gènes devient un événement génétique incontournable qui participe au flux des gènes et à la diversification des espèces. L'étude de la perte de gènes est une « brique » essentielle dans la compréhension des processus évolutifs des espèces.

2.1 Les vestiges dans les génomes

Les génomes sont composés de deux types de séquences, les séquences codantes, composées de gènes (codant pour des transcrits, qui ne sont pas forcément traduits en protéines), et les séquences non codantes. La grande diversité de la taille des génomes est donc définie par le nombre de gènes présents mais aussi par le matériel non codant. Chez les vertébrés, la grande majorité de l'ADN est non codant. Chez l'homme la partie de l'ADN codant est estimée à environ 5% (Lindblad-Toh *et al.*, 2011). Une grande part des éléments non codants peut être imputée à des éléments dits transposables ou **transposons**. Les transposons sont des séquences d'ADN capables de se déplacer dans le génome et de se multiplier de manière autonome, par un mécanisme appelé **transposition**. Presque la moitié du génome humain est composé d'éléments transposables. Le pourcentage peut varier énormément d'une espèce à l'autre. Par exemple, le génome du maïs est composé à 70% d'éléments transposables tandis que le pourcentage chez la drosophile n'est que de 15%. Outre les éléments transposables, l'ADN non codant est constitué de nombreux **pseudogènes**. A titre d'exemple, on estime que le génome humain et celui de la souris contiennent environs 20 000 pseudogènes (Bischof *et al.*, 2006 ; Khelifi *et al.*, 2005 ; Torrents, Suyama, Zdobnov, & Bork, 2003 ; Z. Zhang, Carriero, & Gerstein, 2004 ; Z. Zhang, Harrison, Liu, & Gerstein, 2003).

2.1.1 Les pseudogènes : une définition ambiguë

Les pseudogènes découverts à la fin des années 1970 (Jacq *et al.*, 1977 ; Little, 1982), sont des séquences ADN similaires à des gènes fonctionnels, altérées cependant par la présence de délétions et de codons stops prématurés. Depuis lors, le terme pseudogène a été utilisé pour décrire l'ensemble des séquences génomiques possédant les deux caractéristiques suivantes : la séquence est similaire à un gène fonctionnel ; elle possède des mutations génétiques empêchant la création d'un produit fonctionnel.

La première caractéristique peut être aisément déterminée par l'alignement de la séquence d'un pseudogène avec son paralogue. Le paralogue peut être déterminé au sein de l'espèce où le pseudogène est étudié, ou au sein d'espèces différentes. Cette deuxième solution est obligatoire lorsqu'aucun paralogue fonctionnel n'existe dans l'espèce où l'on étudie un pseudogène.

La seconde caractéristique est plus difficile à définir et à établir. Elle est classiquement décrite comme la présence de mutations entraînant l'incapacité d'une séquence à produire une

protéine ou une structure ARN. Cette définition implique que les pseudogènes sont non fonctionnels. Au début des années 2000, des études rapportent l'existence de pseudogènes fonctionnels, actifs au niveau transcriptionnel (Frith *et al.*, 2006 ; P. Harrison *et al.*, 2002 ; P. M. Harrison, Zheng, Zhang, Carriero, & Gerstein, 2005 ; Yamada *et al.*, 2003 ; Zheng *et al.*, 2005). Ces découvertes sont embarrassantes pour la définition conventionnelle des pseudogènes et nécessite une réévaluation de la définition. Ainsi, un pseudogène peut être défini par la perte de sa fonction d'origine, suite à l'échec de la transcription ou de la traduction, mais aussi à la production d'une protéine qui n'a pas le même répertoire fonctionnel que le gène d'origine (Mighell, Smith, Robinson, & Markham, 2000). De ce fait, un pseudogène n'évoluera pas forcément en suivant la théorie de la neutralité de l'évolution moléculaire comme on l'admettait jusqu'alors.

Trois types de pseudogènes peuvent être décrits selon leur état fonctionnel. Il existe les pseudogènes exaptés qui ont gagné une nouvelle fonction biologique, les pseudogènes « mourants » qui ont encore une activité transcriptionnelle, et les pseudogènes « morts » qui n'ont plus aucun signe de fonctionnalité et qui évoluent selon la règle de la neutralité (Zheng & Gerstein, 2007).

Le statut fonctionnel d'un pseudogène est difficile à évaluer. La description des pseudogènes et leur définition se sont donc naturellement centrées sur les caractéristiques de leur séquence, aux dépens de leur statut fonctionnel. Ainsi les pseudogènes au sens large, possédant ou non un statut fonctionnel, peuvent être répartis en trois groupes distincts selon les caractéristiques de leurs séquences. Ces caractéristiques sont étroitement liées aux types de mutations qui ont donné naissance aux pseudogènes.

2.1.2 Classification des pseudogènes selon la caractéristique des séquences

2.1.2.1 Les pseudogènes non processés

Le terme processé provient du terme anglais décrivant l'épissage. Ainsi, un gène ou pseudogène non processé possède encore ses introns, tandis qu'un gène ou pseudogène processé correspond à une séquence épissée et ne possédant pas d'introns.

Le phénomène de pseudogénisation proposé par Ohno, correspond à une désactivation d'un gène par l'accumulation de mutations délétères. Ce phénomène, qui se produit au cours de la duplication d'un gène fonctionnel, peut entraîner un relâchement de la sélection sur un duplicata du fait de l'existence de fonctions redondantes, et produire des mutations délétères

amenant à la création d'un pseudogène (Lacy & Maniatis, 1980 ; Proudfoot & Maniatis, 1980). Ces pseudogènes décrits par Ohno sont classiquement appelés **pseudogènes non processés** en opposition aux **pseudogènes processés (**Illustration 11**).**

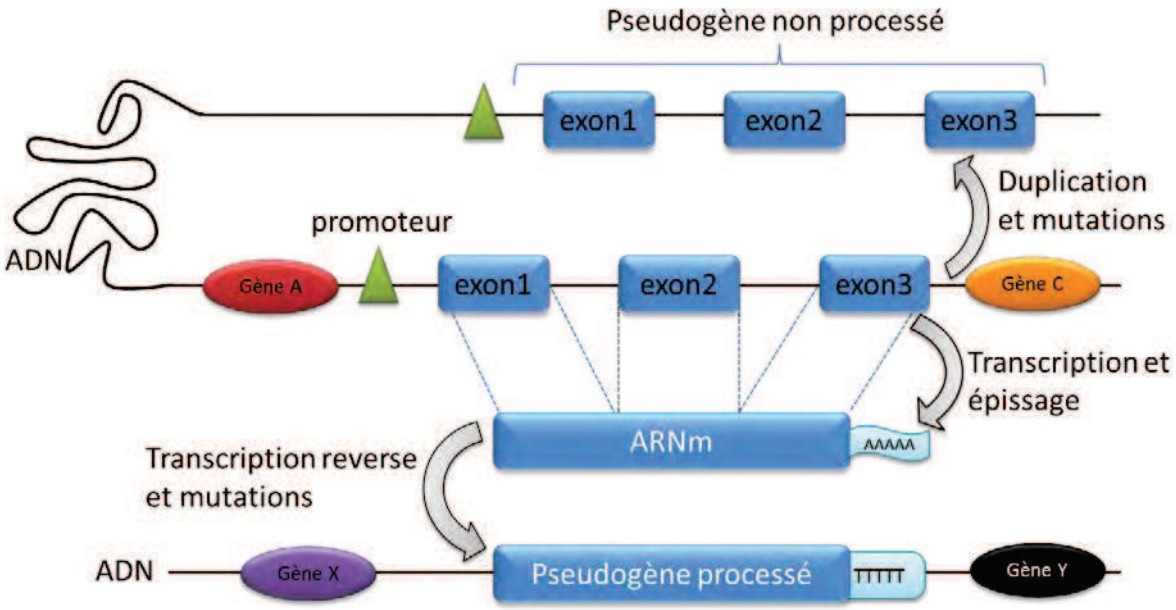


Illustration 11 : Processus d'apparition des pseudogènes processés et non-processés

2.1.2.2 Les pseudogènes processés

Ce sont des séquences qui proviennent de la transcription inverse d'ARN messagers (ARNm) rendue possible par l'utilisation des enzymes (*transcriptase inverse*) de rétrotransposons (M. J. Chen, Shimada, Moulton, Harrison, & Nienhuis, 1982 ; Hollis, Hieter, McBride, Swan, & Leder, 1982). L'ADN codant (ADNc) ainsi créé est réinséré de façon aléatoire dans le génome. Ces pseudogènes sont dits processés car ils ne possèdent pas d'introns, étant donné qu'ils proviennent de la transcription inverse d'ARNm épissé. Les pseudogènes processés sont assimilés à des gènes mort-nés, car dans la majorité des cas, ils ne possèdent pas les éléments nécessaires à leur fonctionnalité, les séquences promotrices par exemple. Dans certains cas, ils peuvent tout de même être fonctionnels (McCarrey & Thomas, 1987). Des exceptions de plus en plus nombreuses montrent que les rétrotranspositions de gènes fonctionnels peuvent amener à la fixation de nouveaux gènes. Le gène *Sult3a1* décrit dans la publication présentant l'outil GLADX, qui est dédié à l'analyse des pertes unitaires de gènes (Article 2), en est un exemple. Le gène *Sult3A1* est décrit comme pseudogène processé chez l'homme (Freimuth, Wiepert, Chute, Wieben, & Weinshilboum, 2004). L'analyse de ce pseudogène par GLADX montre que le gène orthologue d'origine qui devait être non processé semble avoir disparu tandis que le rétrotransposon s'est fixé dans la lignée humaine.

Puis il semble que le rétrotransposon est devenu pseudogène plus tard par accumulation de mutations délétères, et est devenu un pseudogène unitaire chez les Primates.

2.1.2.3 **Les pseudogènes unitaires**

Dans une lignée ou une espèce donnée, les pseudogènes unitaires sont des séquences qui n'ont aucun paralogue fonctionnel. Les pseudogénisations amenant à l'apparition de pseudogènes unitaires (Mitchell & Graur, 2005) concernent des gènes dont la fonction est bien établie. Ces gènes bien établis au sein des génomes d'une lignée, deviennent pseudogènes après des millions d'années d'existence à travers l'apparition de mutations délétères. L'apparition de pseudogènes unitaires est généralement assimilée à la perte des fonctions des gènes d'origine. La pseudogénisation de gènes précédemment établis dans les génomes reflète les conséquences d'un changement de pression de sélection au niveau de ces gènes. C'est en ce sens que l'étude privilégiée de ce type de pseudogènes est essentielle pour obtenir des informations fondamentales afin de mieux appréhender l'histoire évolutive des espèces.

Les pseudogènes unitaires sont traditionnellement décrits comme une sous-famille de pseudogènes non processés car la structure exon-intron du gène ancestral est conservée. Mais pour définir les pseudogènes unitaires, il est erroné de se fier à la structure de leurs séquences. En effet, ces pseudogènes peuvent provenir de gènes qui possèdent ou non des introns. Ainsi, pour différencier ce type de pseudogène il est primordial de connaître l'histoire évolutive du gène. Pour cela il faut passer par la comparaison génomique de différentes espèces afin de connaître les états ancestraux.

2.1.2.4 **Résumé sur l'apparition des pseudogènes**

Les différents types de pseudogènes décrits précédemment sont résumés dans l'Illustration 12 ci-dessous. Un type de pseudogène présent dans l'illustration n'a pas encore été décrit. Il s'agit des pseudogènes provenant de la duplication de pseudogènes déjà existant (Cleary, Schon, & Lingrel, 1981). Comme souligné dans l'illustration, ce phénomène entraîne inévitablement l'apparition de nouveaux pseudogènes qui ont la particularité de provenir de séquences qui n'ont jamais été fonctionnelles. Ces pseudogènes peuvent provenir sans distinction de pseudogènes processés et non processés.

Dans l'illustration on perçoit que parmi les pseudogènes provenant de gènes fonctionnels, seuls les pseudogènes unitaires sont liés à une perte de la fonction du gène d'origine.

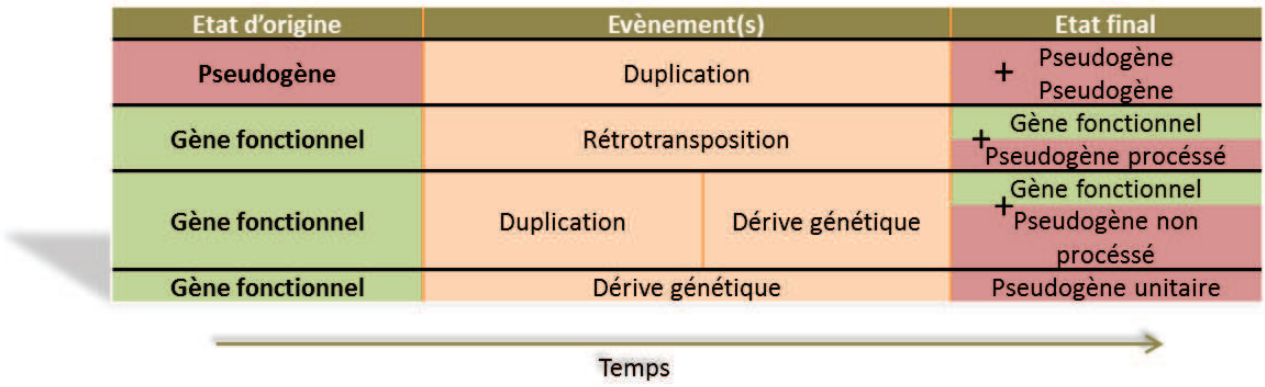


Illustration 12 : Processus d'apparition des pseudogènes par rapport aux séquences d'origines

2.1.3 Le flux des pseudogènes au cours de l'évolution des génomes

Les génomes actuels sont le résultat de milliards d'années d'évolution ; ils ont été façonnés par une multitude d'événements de types divers, plus ou moins nombreux, avec des impacts différents sur les organismes qui les accueillent. L'illustration qui suit met en relief le flux des pseudogènes au sein des génomes au cours de l'évolution (Illustration 13). Elle n'aborde pas les événements entraînant les de pertes de gènes, dans le sens de la disparition du signal d'un gène causée par différents mécanismes mutationnels tels que la délétion et le transfert horizontal. Cette illustration permet donc d'observer les modalités (Rétrotransposition, Duplication, Dérive génétique) d'apparition des différents types de pseudogènes et leur devenir au sein d'un génome. L'épaisseur des flèches reflète l'importance relative de chaque processus. Nous pouvons ainsi observer que les pseudogènes unitaires sont minoritaires parmi les différents types de pseudogènes existant au sein des génomes. Il n'est pas aisé, en conséquence, de les analyser car ils nécessitent une différenciation spécifique parmi de nombreux autres pseudogènes. Les pseudogènes unitaires, selon l'existence d'introns dans les gènes à leurs origines, ressemblent soit à des pseudogènes non processés soit à des pseudogènes processés. La présence de la queue poly A de l'ARNm à l'origine du pseudogène processé peut aider à différencier un pseudogène processé d'un pseudogène unitaire n'ayant pas subi de rétrotransposition. Pour être sûr de ne pas confondre un pseudogène unitaire avec les autres types de pseudogènes, il est nécessaire de connaître l'histoire évolutive qui a donné naissance au pseudogène. En résumé, le schéma ci-dessous (Illustration 13) illustre et représente le fait que les pseudogènes unitaires sont noyés au sein d'une masse d'autres pseudogènes. Seule la connaissance de l'histoire évolutive d'un pseudogène permet de différencier un pseudogène unitaire des autres types de pseudogènes.

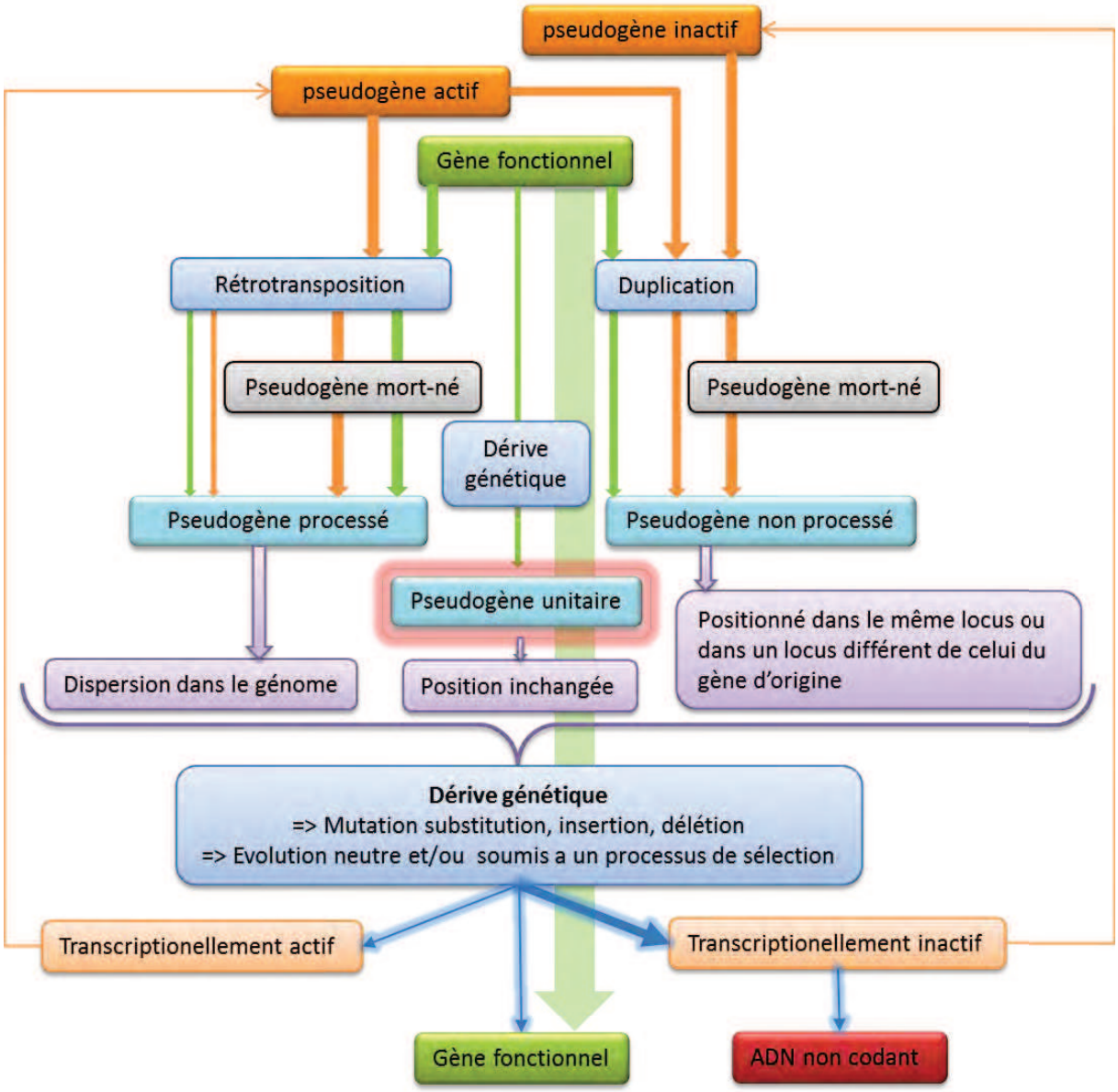


Illustration 13 : Vue centrée sur le flux des éléments non géniques au cours de l'évolution.
La taille des flèches reflète approximativement l'importance de chaque voie.

2.2 La perte de gènes unitaires

On définit par **gène unitaire**, le dernier représentant d'un gène ancestral établi. Ce concept est contraint par le choix du phylum observé. Cette caractéristique est mise en évidence dans l'illustration ci-dessous (Illustration 14). La perte du gène *Hsa1* est une perte de gène unitaire si l'on observe le phylum des Catarrhiniens (*Catarrhini*) (Ctr) mais ce n'en est pas une si l'on observe le phylum des Euthériens (*Eutheria*) (Eth). En effet, dans le phylum des Euthériens un représentant du gène ancestral est présent chez l'Homme (*H. sapiens*) (*Hsa*) avec le gène *Hsa2*.

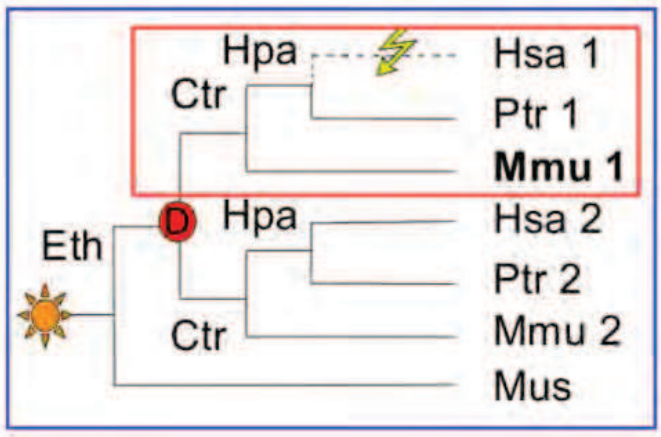


Illustration 14 : Impact du phylum observé sur l'étude de pertes de gènes unitaires

La **perte de gènes unitaires** est donc une perte par **délétion** ou **pseudogénisation** de gènes précédemment établis et n'ayant pas de paralogues fonctionnels dans la lignée concernée. Ces pertes sont donc susceptibles de révéler des pertes de fonctions. Le concept de perte de gènes unitaires se retrouve dans la littérature, sous les termes de **perte de lignées spécifiques** (Aravind, Watanabe, Lipman, & Koonin, 2000 ; Go, Satta, Takenaka, & Takahata, 2005) ou perte **de gènes longtemps établis** (Zhu *et al.*, 2007). C'est dans un souci d'harmonisation avec le terme et le concept de pseudogène unitaire, que je propose et utilise dans ce manuscrit le terme et le concept de pertes de gènes unitaires.

Les pertes de gènes unitaires sont inévitablement liées à la pseudogénisation ou à la délétion, mais ces événements ne conduisent pas forcément à des pertes de gènes unitaires.

Le processus de pseudogénisation causant des pertes de gènes unitaires a été abordé dans le paragraphe précédent (2.1). Les événements de délétion à l'origine de pertes de gènes unitaires peuvent se dérouler au cours de la phase de recombinaison de la méiose, ou encore survenir au cours de la vie d'un organisme, dans les cellules sexuelles. Les délétions mènent à la réduction de la taille des génomes alors que la pseudogénisation laisse les séquences en place. La perte d'un gène unitaire par pseudogénisation peut-être différenciée de la perte d'un gène unitaire par délétion par la présence ou l'absence du signal du gène d'origine. Cela reste possible quand la pseudogénisation n'est pas trop ancienne, car, passé un certain temps, la dérive génétique efface entièrement le signal de la séquence d'origine. A défaut, la différenciation peut se faire grâce à l'étude de la synténie conservée, en observant s'il y a disparition de la séquence entre les deux gènes voisins, ou si d'autres gènes mitoyens ont également disparu.

L'étude des pertes de gènes unitaires est rendue possible par l'apparition de génomes

entièrement séquencés. La génomique comparative permet de reconstruire l'état ancestral et donc de déceler la présence ou l'absence de gènes dans les génomes ancestraux de manière fiable. Cette reconstruction permet de dire s'il y a eu perte de gènes ancestraux préalablement établis au sein des génomes. La perte de gènes unitaires semble un événement assez marginal par rapport aux autres événements auxquels sont soumis les génomes, mais c'est semble-il un phénomène riche d'informations pour comprendre l'évolution des espèces. Les pertes de gènes unitaires qui se fixent au sein des génomes des espèces au cours de l'évolution ne semblent pas engendrer de désavantage évolutif. Si ces pertes étaient désavantageuses, les individus possédant le gène intact seraient les vainqueurs exclusifs de la compétition pour la survie et la reproduction. D'un point de vue évolutif, ces pertes concernent des gènes qui ont eu une fonction importante dans les espèces ancestrales, et dont la fonction est devenue non essentielle dans les espèces modernes. L'hypothèse « *less is more* » affirme que ces pertes peuvent avoir eu un impact important sur l'adaptation des espèces au cours de l'évolution (Olson, 1999). L'investigation des pertes de gènes unitaires a montré que ce phénomène est ubiquiste, et existe aussi bien chez les Eucaryotes que chez les procaryotes (Aravind *et al.*, 2000 ; Costello, Han, & Hahn, 2008 ; Hughes & Friedman, 2005 ; Roelofs & Haastert, 2001 ; Zhu *et al.*, 2007).

La perte de gènes peut avoir un impact impressionnant sur les génomes. Ainsi, il est largement admis qu'elle est particulièrement présente chez les organismes intracellulaires (parasites, mutualistes, endosymbiontes). Le génome des bactéries de cette catégorie, dérive d'ancêtres vivants à l'état libre qui possédaient des génomes plus conséquents (Andersson & Kurland, 1998 ; N. A. Moran & Mira, 2001 ; Sällström & Andersson, 2005). La réduction de la taille des génomes est associée à l'adaptation à leur mode de vie intracellulaire (Darby, Cho, Fuxelius, Westberg, & Andersson, 2007 ; N. Moran & Wernegreen, 2000 ; Zomorodipour & Andersson, 1999). Le même phénomène se retrouve également chez des Eucaryotes intracellulaires (Baumann *et al.*, 1995). La perte massive de gènes ne se cantonne pas uniquement aux organismes intracellulaires : des pertes à grande échelle ont été observées dans les lignées menant aux *Fungi*, aux Nematodes, ou encore chez la Cione (Hughes & Friedman, 2005 ; Koonin *et al.*, 2004). Les pertes de gènes unitaires se retrouvent dans l'ensemble des génomes étudiés, avec des intensités variables d'une lignée à l'autre ; elles participent significativement à la divergence des organismes au cour de l'évolution (Aravind *et al.*, 2000 ; Hughes & Friedman, 2004a ; Wyder, Kriventseva, Schröder, Kadowaki, & Zdobnov, 2007). De nombreuses études mettent en lumière les pertes de gènes unitaires dans

le parcours évolutif à l'origine de l'espèce humaine. Par exemple, les récepteurs olfactifs semblent s'être détériorés très rapidement chez les Primates pouvant refléter des réponses à des changements environnementaux ou comportementaux (Go *et al.*, 2005 ; X. Wang, Thomas, & Zhang, 2004).

Les études montrent que les pertes de gènes unitaires sont d'autant plus importantes que la perte d'un seul gène peut avoir une incidence prépondérante au sein d'un organisme. Par exemple, la perte du gène Gulo chez les Primates, qui code pour une enzyme qui catalyse la dernière étape de la biosynthèse de la vitamine C (Nishikimi *et al.*, 1994), rend impossible la synthèse de la vitamine C. C'est pourquoi les Primates doivent trouver la vitamine C dans leur alimentation.

Les pertes de gènes lignées spécifiques ou autrement dit de gènes unitaires reflètent des phénomènes importants. Elles peuvent être le fruit de changements

- **liés à des changements environnementaux** ayant pour conséquence :
    - un relâchement de la sélection d'un gène qui n'a plus d'utilité

Dans ce cas la séquence du gène n'est plus soumise à aucune pression sélective et évolue sous la neutralité.

    - une perte adaptative d'un gène qui crée un handicap pour l'organisme

Ces pertes sont le fruit de sélections négatives liées à des événements adaptatifs. Dans ce cas, elles apportent un bénéfice aux espèces, alors que le maintien de ces gènes au sein des génomes aurait été défavorable.

- **non liés à des changements environnementaux**
    - par le relâchement de la sélection lié à l'apparition de gènes accomplissant la fonction initiale.
    - par le relâchement de la sélection lié à des modifications (éthologique, physiologique, etc.) de l'organisme qui rend le gène inutile à sa survie.

Dans ces cas la séquence du gène n'est plus soumise à aucune pression sélective et évolue sous la neutralité.

Quelle que soit la raison de la perte de gènes dans une lignée spécifique, on peut faire l'hypothèse qu'un changement important s'est produit ; il est intéressant de le découvrir pour comprendre l'histoire évolutive de la lignée. L'étude des pertes de gènes unitaires permet

donc de comprendre les changements survenus au niveau de l'organisme, de son environnement, et des liens pouvant exister entre ces deux éléments.

L'observation d'évolutions neutres ou adaptatives peut se faire au niveau des allèles qui contiennent des pertes de gènes unitaires, en étudiant la fixation de ces allèles au sein des populations. Lorsque les pertes de gènes unitaires sont le fruit de pseudogénisation, il est difficile d'étudier le type de sélection aboutissant à ces pertes, car les pseudogènes évoluent sous la neutralité (W.-H. Li, Gojobori, & Nei, 1981). Même quand une pression de sélection négative est à l'origine de l'apparition d'un pseudogène unitaire, dès la première mutation délétère du gène, le pseudogène nouvellement formé évolue sous la neutralité. Qu'elle soit liée ou non à une sélection, chaque perte de gènes observée peut donner une information sur les changements survenus dans l'organisme ou dans son environnement. C'est pourquoi l'étude des pertes de gènes unitaires qui ont affecté différentes lignées indépendantes peut mener à définir des facteurs adaptatifs communs : la niche écologique, le comportement, le mode de vie, etc. Convergence évolutive et coévolution adaptative peuvent être détectées au travers des événements de perte de gènes.

Lorsque des espèces proches ont un gène pseudogénisé commun, il est important de savoir si les pseudogénisations sont indépendantes. La détermination de mutations (qui ont conduit à une pseudogénisation) communes à plusieurs espèces, permet de faire cette différence, et de dater avec précision le début de la pseudogénisation. Cela permet également de retracer l'histoire du processus. L'analyse en soi de chaque pseudogénisation est instructive pour la bonne compréhension du processus et la connaissance des événements déclencheurs.

2.2.1 Pertes de gènes unitaires : délétion VS pseudogénisation

Lorsqu'une perte est récente, il est possible de faire l'hypothèse qu'elle soit liée à un événement de délétion si aucune observation de pseudogènes n'est faite.

Différencier une délétion d'une pseudogénisation qui mène à une perte de gènes unitaires est impossible lorsque ces pertes sont anciennes. En effet, avec le temps, et la dérive génétique subie sous évolution neutre, les pseudogènes ne sont plus reconnaissables au sein des génomes. Il est donc difficile de connaître la part des pertes imputables à des délétions et celle des pertes imputables à des pseudogénisations. Même dans les lignées qui impliquent une réduction des génomes, il est impossible, pour toutes les pertes anciennes, de savoir si une pseudogénisation a précédé la délétion.

Les pertes par délétion peuvent être détectées en étudiant la synténie des gènes de la région

concerné dans les génomes investigués. En effet, l'absence des gènes qui encadrent une perte étudiée peut consolider l'hypothèse d'un événement de délétion. Cependant, les investigations menées dans la lignée menant à la drosophile, et dans celle menant à l'homme, montrent que les pertes de gènes unitaires récentes peuvent être imputées presque exclusivement à des pseudogénisations (Costello *et al.*, 2008 ; Z. D. Zhang, Frankish, Hunt, Harrow, & Gerstein, 2010 ; Zhu *et al.*, 2007).

La délétion de séquences codantes ou non codantes peut être vue comme une optimisation des génomes. D'un autre côté, l'accumulation de séquences de pseudogènes peut être également vue comme quelque chose de bénéfique, car elle peut constituer un vivier important favorisant la création de nouveautés génétiques. Ces questions ne sont pas abordées dans cette thèse.

2.2.2 État des connaissances sur l'analyse des pertes de gènes unitaires

L'analyse des pertes de gènes unitaires implique la reconstruction des états ancestraux à partir des données contemporaines. Malgré les trous de séquençages, qui existent dans les génomes entièrement séquencés, et les limites des outils dont nous disposons pour analyser les séquences, les méthodes de génomique comparative semblent donner de bons résultats dans la détection des pertes de gènes unitaires. Malgré tout, l'étude des pertes de gènes unitaires en traitant les pseudogènes unitaires s'avère difficile, car il faut pouvoir distinguer les pseudogènes unitaires des autres types de pseudogènes présents en grand nombre au sein des génomes.

La publication de génomes entièrement séquencés a concouru au développement de méthodes d'analyses et d'outils spécifiques pour mener à bien des études à grande échelle. De nombreuses études à grande échelle de pertes de gènes unitaires ont été réalisées dans différentes lignées (Annexe 1). Il existe deux grands types d'études : celles qui utilisent des approches fondées uniquement sur les protéomes et l'absence de protéines orthologues ; et celles qui se basent exclusivement sur la présence de pseudogènes. Ces études sont complémentaires, mais aucun outil ne permet, actuellement, l'analyse simultanée au niveau protéique et nucléotidique.

2.2.2.1 Détection des pertes par l'identification des protéines/gènes manquants

Ces méthodes prennent en compte uniquement l'absence du gène ou de la protéine orthologue dans les bases de données qu'elles utilisent. Elles utilisent de préférence la similarité des séquences protéiques car les séquences protéiques divergent moins que les séquences

nucléotidiques au cours de l'évolution. En effet, au sein d'un codon, certaines substitutions nucléotidiques ne changent pas l'acide aminé produit. Ces méthodes peuvent détecter des pertes de gènes unitaires très anciennes, mais elles ne permettent pas d'analyser les pseudogènes. Les études utilisant ce type de méthodes sont nombreuses : parmi les 25 études décrites en annexe (Annexe 1), 21 utilisent cette approche.

2.2.2.1.1 Méthodes classique de BLAST

Ces méthodes se basent sur le BLAST des séquences du protéome d'une espèce de référence, sur le protéome d'autres espèces (Illustration 15). Connaissant l'arbre des espèces utilisé, il est possible de définir dans quelles lignées sont survenues des pertes de gènes unitaires. Si le BLAST ne trouve pas de séquence similaire dans les autres espèces, alors on parle de gènes orphelins qui sont spécifiques à l'espèce de référence utilisée.

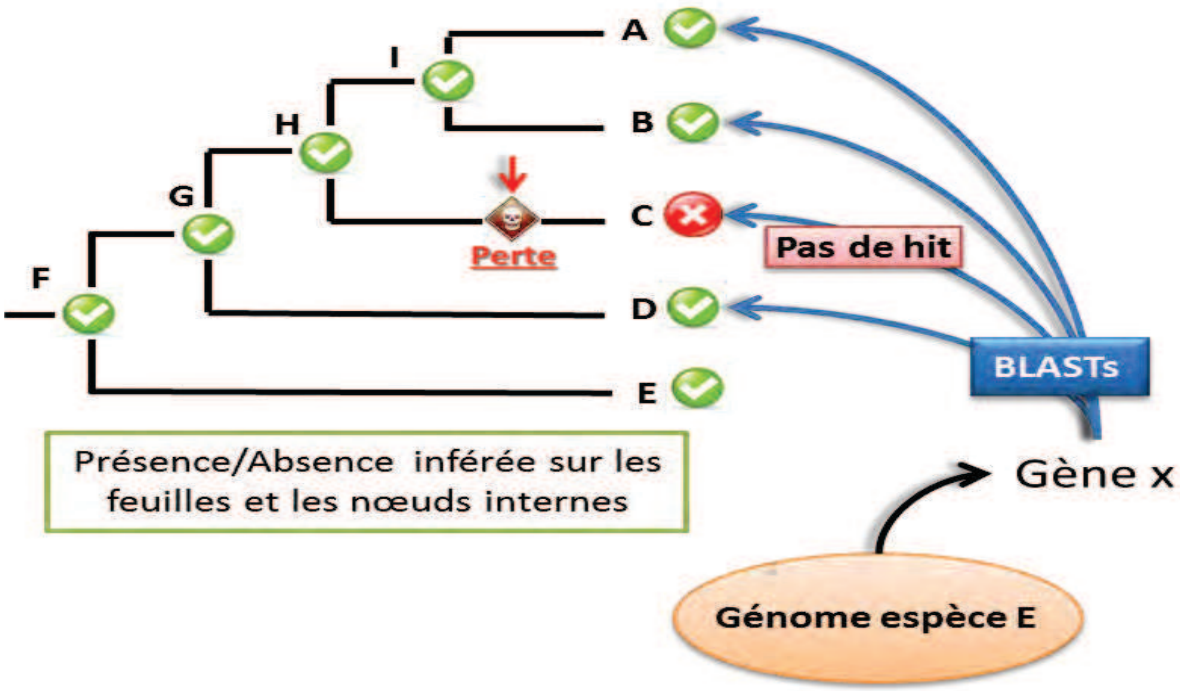


Illustration 15 : Processus d'analyse des pertes de gènes unitaires basé sur le BLAST
A, B, C, D, E sont des espèces contemporaines.

Les outils les plus utilisés pour observer les similarités de séquences protéiques sont le BLAST (Altschul, Gish, Miller, Myers, & Lipman, 1990) et le PSI-BLAST (Altschul *et al.*, 1997) qui sont des outils utilisant des algorithmes beaucoup plus rapides que l'algorithme traditionnel de Smith-Waterman (T. F. Smith & Waterman, 1981). Ce dernier algorithme a été fortement optimisé. Il faut souligner que le terme « BLAST » est entré dans le langage courant et signifie « la recherche de séquences similaires » de manière générale sans

forcément définir le type de d'algorithme utilisé, qui peut être de type BLASTN, BLASTP, PSI-BLAST, PHI-BLAST, etc.

2.2.2.1.2 **Méthodes des groupes d'orthologues (GOs)**

Des méthodes utilisant la similarité de séquences ont été développées afin d'étudier de nombreux génomes simultanément. La technique dite des COGs (R. L. Tatusov, 1997) s'est imposée pour regrouper des séquences de différentes espèces selon leur similarité. Elle permet de créer des clusters de groupes d'orthologues appelés **COGs** (Cluster of Orthologous Groups of Proteins). Elle se base sur la méthode de « best BLAST hit réciproque » qui crée un noyau d'orthologues à partir des séquences les plus proches dans les différentes espèces utilisées. Ce noyau est ensuite complété de manière agrégative par l'ajout d'inparalogues représentés par d'autres hits proches. Cette approche donne naissance à des **groupes d'orthologues** (**GOs**) putatifs.

A partir de ces GOs, les espèces représentées sont inférées sur les feuilles de l'arbre des espèces utilisé. Ensuite, les méthodes de parcimonie permettent d'inférer les états présents ou absents dans les nœuds ancestraux, et d'observer les gains et les pertes de gènes dans les lignées observées (Illustration 16).

De nombreuses méthodes ont été développées pour fabriquer des GOs qui s'inspirent de l'approche des COGs. Ces méthodes sont fréquemment utilisées pour étudier les pertes de gènes unitaires car elles sont rapides et d'utilisation facile. Parmi les 25 études décrites en annexe (Annexe 1), 16 utilisent ce type d'approche. Les études basées sur l'analyse de GOs sont efficaces pour étudier les flux de gènes (gains et pertes) au sein des différentes lignées et deviennent un élément central pour comprendre l'évolution des espèces. Ainsi, lorsqu'un génome est nouvellement séquencé, il devient habituel de l'étudier en utilisant l'approche des COGs, ce qui permet d'obtenir rapidement beaucoup d'informations importantes. Néanmoins, la qualité des GOs créés, sur lesquels on se base pour inférer la présence et l'absence de gènes, est très sensible à la qualité d'annotation des génomes. De plus, la précision de la définition des GOs par ces méthodes est relativement moindre que celle produite par des méthodes basées sur la phylogénie.

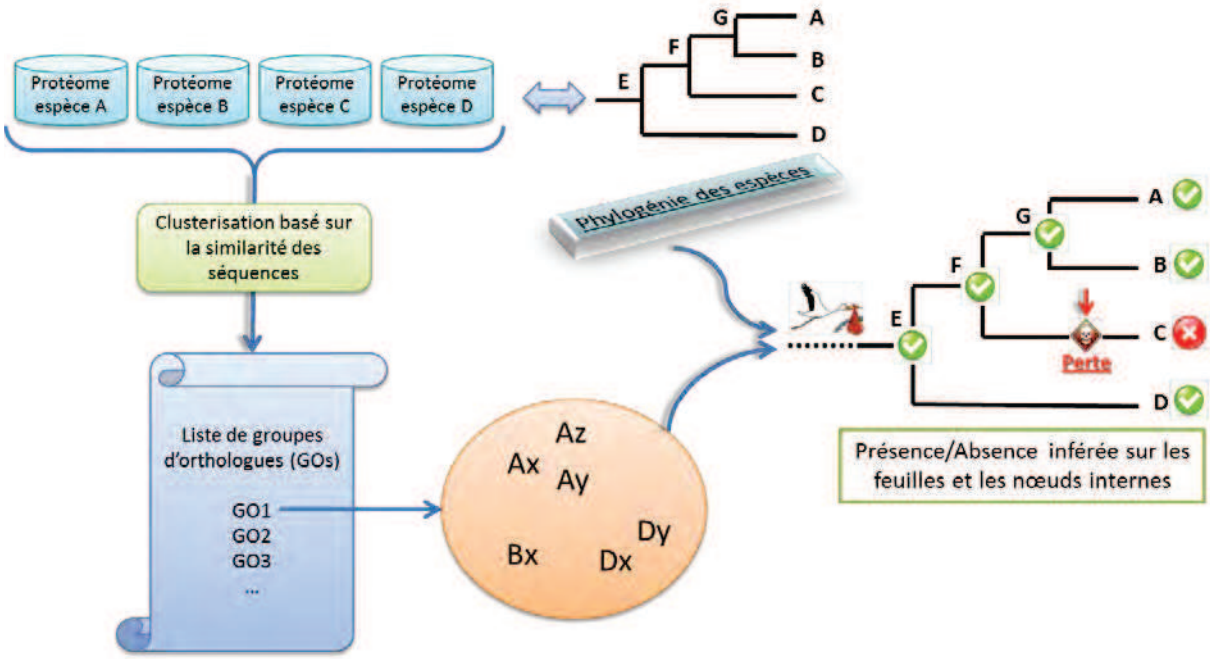


Illustration 16: Processus d'analyse des pertes de gènes unitaires basé sur la création de groupes d'orthologues

2.2.2.1.3 **Méthodes phylogénétiques :**

Les méthodes phylogénétiques reposent sur des définitions fines des relations de parenté entre les séquences. Les relations d'orthologie et de paralogie sont bien décrites et permettent une détection plus robuste des pertes de gènes unitaires. Cependant ces méthodes restent d'usage relativement rare (Blomme *et al.*, 2006 ; Kuraku & Kuratani, 2011) car elles nécessitent la production de phylogénies ainsi que leur analyse. La création et l'analyse de phylogénies restent fastidieuses, difficiles à automatiser, et demandent un temps de calcul largement supérieur aux méthodes des COGs.

Les méthodes basées respectivement sur le BLAST classique, sur la création de clusters de séquences orthologues (COGs) putatives, ou sur la phylogénie, ont toutes leur place dans l'étude des pertes de gènes unitaires ; elles manquent cependant de précision. En effet, dans le cas de pertes récentes elles ne permettent pas de différencier les pertes dues à une délétion de celles dues à une pseudogénisation. Elles ne permettent donc pas l'analyse du processus de pseudogénisation. Ces méthodes se basent sur les annotations protéiques existantes ; elles ne permettent pas de déceler les séquences géniques qui auraient pu être oubliées lors de l'annotation des génomes. Cette caractéristique peut conduire à une sur-prédiction du phénomène de pertes de gènes unitaires.

2.2.2.2 Détection des pertes de gènes unitaires par détection des pseudogènes unitaires

Pour l'étude des pertes de gènes unitaires, il existe d'autres approches basées sur l'analyse des séquences génomiques et la détection des pseudogènes unitaires. Ces approches consistent à rechercher par génomique comparative, un pseudogène dans une espèce A qui correspond à un gène décrit chez une espèce B. Elles sont basées sur l'utilisation des algorithmes de recherche de similarités traditionnelles. Les pseudogènes ainsi trouvés correspondent à des pseudogènes unitaires. L'analyse de l'alignement entre la séquence du pseudogène unitaire et du gène orthologue fonctionnel, permet de connaître les mutations délétères qui affectent le pseudogène. Ces approches utilisées pour étudier les pseudogènes unitaires sont limitées par le nombre d'espèces utilisées (souvent deux espèces) et nécessitent l'utilisation d'espèces relativement proches. De plus, l'identification des pseudogènes unitaires requiert non seulement la recherche de similarité entre un pseudogène et son gène homologue, mais aussi l'attribution d'une relation d'orthologie. Définir l'orthologie en utilisant les scores de similarités peut induire des erreurs. La synténie peut être utilisée pour vérifier l'orthologie des séquences étudiées (Z. D. Zhang *et al.*, 2010 ; Zhu *et al.*, 2007). Mais la synténie peut également induire des erreurs en ratant les coorthologues apparus par duplication dans les espèces étudiées. En effet, la perte d'un orthologue à sa position d'origine n'est pas synonyme d'une perte de gène unitaire, car un autre orthologue, à une position différente, peut toujours subsister dans un génome (Illustration 17). La recherche de pertes de gènes unitaires implique de vérifier l'absence de tous les orthologues.

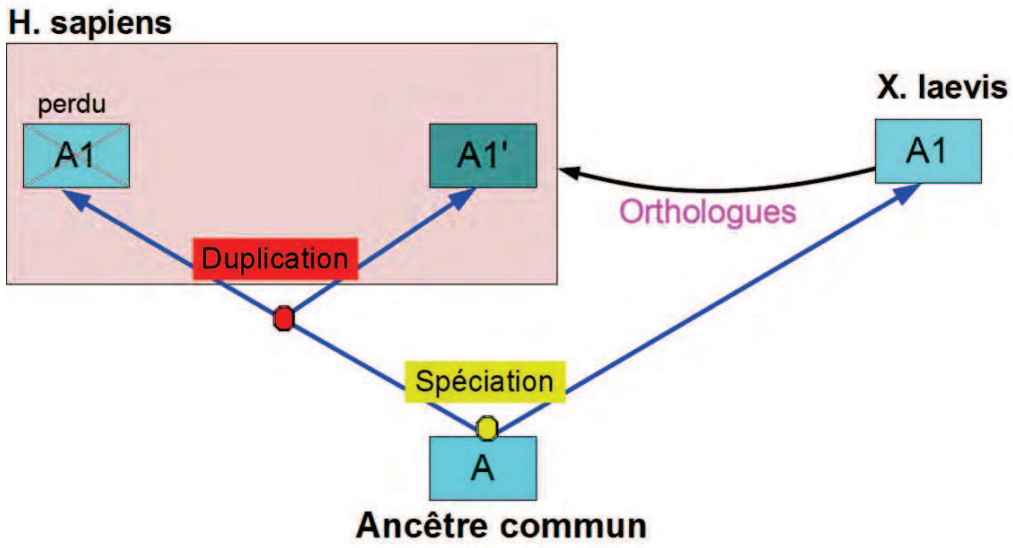


Illustration 17: Persistance d'un orthologue après la perte de l'orthologue d'origine

Les méthodes d'étude des pseudogènes unitaires peuvent être utilisées à la suite d'une analyse qui utilise la création de GOs : l'analyse des GOs est alors utilisée comme filtre. En effet, les pertes détectées par l'approche des GOs sont des cibles privilégiées pour la recherche de pseudogènes unitaires (Costello *et al.*, 2008 ; Schrider, Costello, & Hahn, 2009 ; Z. D. Zhang *et al.*, 2010).

3 Objectifs

Les séquences génomiques contiennent toutes les informations génétiques nécessaires au développement des organismes. Selon Theodosius Dobzhansky « Rien n'a de sens en biologie, si ce n'est à la lumière de l'évolution ». Cet aphorisme se traduit de nos jours par la nécessité de mettre en exergue tous les résultats d'événements génétiques (recombinaison, perte, substitution, duplication, THG, etc.) présents dans les génomes à la lumière de l'évolution.

L'ensemble des informations géniques a été décodé pour un grand nombre d'organismes et a permis d'accéder à un large ensemble de données : génomes, exomes, transcriptomes, protéomes, voire intéractomes. L'analyse à grande échelle de ces données par différentes méthodes, permet d'appréhender les mécanismes impliqués dans l'évolution. Le but de ces analyses est de comprendre les causes et les conséquences des mutations, à tous les niveaux d'organisation des organismes. Les mutations élémentaires s'observent au niveau des gènes et peuvent engendrer des événements en cascades à différents niveaux phénotypiques : transcriptionnel, protéique, biochimique, physiologique, éthologique, etc. Il est donc primordial de comprendre la fonction des gènes, et leur évolution au sein des génomes au travers des différents événements génétiques.

La meilleure approche pour étudier des gènes dans un contexte évolutif est l'approche phylogénétique. La phylogénétique donne tout d'abord une idée robuste des liens de parenté entre des séquences géniques. La connaissance des liens de parenté permet par exemple d'annoter la fonction des gènes par inférence. Des gènes orthologues auront tendance à partager une même fonction, plus que des gènes paralogues (Eisen & Fraser, 2003). Les liens de parenté permettent également de reconstruire des états ancestraux à partir de caractères contemporains, et ainsi de décrire et dater les événements génétiques survenus au cours de l'évolution. L'étude des événements génétiques dans une perspective temporelle permet d'établir des liens sous-jacents entre les mutations, les conséquences phénotypiques sur les

organismes étudiés, et l'environnement. La fixation de mutations peut être liée à des adaptations ou être neutre. Les adaptations peuvent être liées à une optimisation des organismes dans un environnement stable mais peuvent également être une réponse à des modifications environnementales. L'augmentation du nombre des génomes séquencés et des informations qui en découlent permet de faire des analyses corrélatives à grande échelle et de mettre en évidence des phénomènes de convergence, de co-évolution, etc.

Dans le contexte actuel, le facteur limitant de la recherche, c'est l'analyse des données produites de manière massive et non plus sur l'obtention de ces données. Les données disponibles constituent un réservoir phénoménal d'informations, leur exploitation optimale est un vrai défi. Pour suivre les inévitables avancées conceptuelles et techniques de la recherche, il est nécessaire de spécifier l'étude de chaque type d'événement dans un contexte flexible.

Les recherches menées au laboratoire s'inscrivent dans ce contexte où le tri, la sélection de données nombreuses et hétérogènes est nécessaire afin de produire des résultats de qualité. La quantité importante de données biologiques disponibles est une mine d'informations mais l'exploitation directe de ces données par les biologistes est difficile, car les informations pertinentes sont noyées dans une énorme masse de données. Notre laboratoire automatise l'exploitation des données biologique afin d'en extraire les informations pertinentes qui permettent de répondre aux diverses questions biologiques que nous nous posons. Les méthodes automatisées que nous développons, sont flexibles, adaptables, et permettent de fines expertises.

3.1 **Objectifs du laboratoire**

Le laboratoire cherche à analyser l'ensemble des événements génétiques et leurs conséquences au cours de l'évolution. Pour cela, il développe des outils adaptés aux analyses à grande échelle, spécifiques aux différents types d'événements qui se sont produits au cours de l'évolution. Lors de mon arrivée, le laboratoire travaillait sur le projet de recherche EvolHHuPro (*Evolutionary Histories of Human Proteome*) qui consiste à définir l'histoire évolutive du protéome des Chordés (*Chordata*). Ce travail de recherche s'organise autour de la construction du phylome humain qui sert de référence, et de l'analyse de l'ensemble des événements apparus au niveau protéique et nucléotidique au sein des différentes lignées de 13 espèces de Chordés.

Dans ce projet, chaque protéine humaine est utilisée pour produire une phylogénie de 13

Chordés dont l'homme. Des méthodes expertes qui utilisent ces phylogénies sont développées afin d'éclairer les différents types d'événements pouvant être étudiés (duplication, gain, perte, nouvelles architectures en domaines de protéines, THG, etc.). La phylogénie est donc une étape déterminante pour permettre une analyse à la lumière de l'évolution. En fonction des événements étudiés, si les données le permettent, des investigations complémentaires sont effectuées au niveau nucléotidique en plus des investigations au niveau protéique. L'ensemble des informations récoltées permet de déterminer avec précision les événements qui se sont produits au cours de l'évolution. Tous les événements peuvent être corrélés, *a posteriori*, à des changements qui ont eu lieu à des niveaux différents (fonction, localisation, etc.).

Dans ce vaste projet d'analyse de l'évolution du protéome des Chordés, mon sujet de thèse porte sur la conceptualisation d'une méthode permettant l'étude spécifique des événements de pertes de gènes unitaires.

3.2 Objectifs de la thèse

Le sujet de ma thèse s'inscrit dans le projet EvolHHuPro du laboratoire et se focalise sur un type particulier d'événement génétique : la perte de gènes unitaires. La perte de gènes unitaires est un type particulier de perte de gènes au sein des génomes. Ces pertes sont intéressantes à analyser car elles touchent des gènes bien établis. Elles concernent des gènes ancestraux, présents dans de nombreuses lignées descendantes, et qui ont néanmoins disparu dans certaines lignées. L'établissement d'un gène, et donc de sa fonction associée, est un phénomène sélectionné. En conséquence, la perte d'un gène unitaire peut être synonyme de perte de fonction(s). Des exemples montrent que ces pertes peuvent se traduire par la perte de fonctions qui jouent un grand rôle (Nishikimi *et al.*, 1994 ; Stedman *et al.*, 2004 ; Varki, 2001 ; Wu *et al.*, 1989). Il s'agit d'un problème majeur en génomique évolutive car ces pertes peuvent refléter des changements importants au sein des organismes et de leurs environnements. L'analyse des pertes de gènes unitaires est fondamentale, et instructive, pour la compréhension des conditions de sélection, tout comme celle des événements génomiques qui participent à ces pertes. Plus concrètement, les pertes comme les gains de gènes sont des événements de l'évolution qui renseignent sur les fonctions à différents niveaux (moléculaire, cellulaire, physiologique, etc.), et sur ce qui différencie les espèces les unes des autres.

Le phénomène de perte de gènes unitaires est encore mal connu, du fait de la découverte récente du phénomène et de la disponibilité récente de génomes entièrement séquencés. Il est

essentiel de disposer de génomes entièrement séquencés pour pouvoir détecter des pertes avec une bonne fiabilité, et éliminer ainsi la possibilité que l'absence d'un gène ne soit due à une zone non séquencée du génome.

Dans la majorité des publications portant sur les pertes de gènes unitaires les chercheurs utilisent les protéomes et se basent sur la présence ou l'absence de protéines orthologues. Ces études, très instructives sur l'aspect dynamique des génomes (tendances de gains et de pertes de gènes), donnent pourtant des résultats manquant considérablement de précision, car elles utilisent les méthodes de clusters de séquences protéiques similaires pour déterminer l'orthologie des séquences. Elles font aveuglement confiance aux annotations protéiques disponibles dans les bases de données. Il peut pourtant manquer des protéines non prédites, et certaines protéines peuvent être mal prédites dans ces bases de données. De plus, l'utilisation des scores de similarités de séquences pour les annotations d'orthologie n'est pas la méthode la plus robuste. De nombreuses erreurs peuvent survenir, en particulier lors de l'étude de familles multigéniques.

Les études plus fines qui recherchent des traces liées aux événements de pertes (pseudogènes) au sein des génomes, font appel à de nombreuses manipulations manuelles qui les rendent fastidieuses. Elles utilisent des espèces proches et un nombre d'espèces très restreint. Bien souvent, elles se contentent de la comparaison des séquences de deux espèces (Hahn & Lee, 2005 ; X. Wang, Grus, & Zhang, 2006).

Les études qui portent sur le processus de pseudogénisation sont encore moins automatisées que la recherche de pseudogènes au sein des génomes. Ces études nécessitent des analyses évolutives qui se basent sur l'utilisation de concepts et d'outils variés, ainsi que de l'expertise humaine.

Les études basées sur la présence de pseudogènes se font, dans la majorité des cas, à partir de la comparaison de séquences contemporaines deux à deux et non sur une base évolutive qui est nettement plus riche d'enseignements (Levasseur, Pontarotti, Poch, & Thompson, 2008). Les outils qui permettent ces études dans un contexte évolutif sont encore rares, souvent mal adaptés, gourmands en calcul, et d'utilisation peu aisée.

- Le but premier de cette thèse est de **conceptualiser et développer une méthode** originale permettant d'**étudier la perte de gènes unitaires** au cours de l'évolution des génomes.

  Pour ce faire la méthode doit :

- **s'appuyer sur les techniques existantes** au sein du laboratoire et celles de la communauté scientifique qui ont montré leur efficacité.
- être **entièrement automatisée** et pouvoir être utilisée à **grande échelle.**
- **s'inspirer de l'expertise manuelle** actuellement utilisée pour l'étude des pertes de gènes unitaires et la mimer.
- être capable de faire la différence entre les pertes où le **signal des séquences des gènes d'origine est absent**, et celles où **les pseudogènes sont encore présents**.
- être aisément **adaptable à l'analyse d'espèces et de lignées variées.**
- fournir une analyse sur une base **évolutive.**
- pouvoir **annoter les séquences** pour différencier les gènes intacts des pseudogènes.
- **analyser les séquences au niveau protéique et nucléique** en fonction de la divergence des séquences.
- pouvoir **analyser les mutations génétiques** des pseudogènes.
- pouvoir **analyser le processus de pseudogénisation.**
- fournir des **informations classées** de manière logique pour permettre leur réutilisation et interpréter de nouveaux résultats.
- fournir une **lecture facilitée des résultats.**

L'automatisation de la méthode est utile pour éviter les erreurs qui peuvent être induites par de trop nombreuses manipulations manuelles, pour gagner en rapidité, et pour permettre des études à grande échelle. Le choix de mimer l'expertise humaine s'est fait naturellement avec l'apparition de nombreux moyens techniques. L'automatisation de l'expertise humaine permet d'utiliser les méthodes éprouvées par les biologistes, de faire des choix dans le cheminement des études en fonction des résultats récoltés, de prendre en compte la complexité et les différentes composantes du problème à résoudre.

L'étude des pertes de gènes unitaires au cours de l'évolution des génomes, implique de manière concomitante l'**étude de l'apparition des gènes**. Les informations concernant les apparitions et les pertes de gènes permettent de déduire le temps de fixation des gènes au sein des génomes. Cette information est essentielle pour avoir une idée de l'importance d'un gène. De manière intuitive, plus un gène est établi depuis longtemps au sein d'un génome, plus on peut penser qu'il remplit un rôle central. Donc, pour avoir toutes les

cartes en main et répondre au mieux aux questions qui touchent la perte de gènes unitaires, il est nécessaire d'intégrer l'analyse de divers événements (gains, pertes, duplications, mutations) qui caractérisent le parcours évolutif d'une famille de gènes. Toutes ces composantes doivent être associées dans l'outil que je cherche à développer.

- Dans une deuxième phase, j'utilise l'outil que j'ai développé dans le cadre d'une **étude à grande échelle** des pertes de gènes unitaires chez les Eucaryotes. Pour réaliser cet objectif, j'étudie spécifiquement la **lignée humaine** qui est une contribution au projet Evolhhupro portant sur l'analyse du protéome des Chordés. Mon travail vise à rendre exhaustive l'étude de l'évolution du protéome humain. Comme l'étude ne peut pas être initiée à partir du protéome humain, il s'agit de **développer une stratégie** afin de **sélectionner à partir d'autres espèces, les gènes** d'intérêt à étudier dans la lignée humaine. C'est donc à partir de l'identification des gènes présents chez les autres espèces et la reconstruction des états de présence et d'absence dans les ancêtres de la lignée humaine, qu'il est possible de déterminer si un représentant d'un gène ancestral existe chez l'homme. Pour mener à bien cette recherche de perte de gènes unitaires à grande échelle, la stratégie développée associe la méthode basée sur l'absence de protéines codantes grâce à l'étude des GOs, et celle basée sur la phylogénie développée dans le cadre de ma thèse au travers de GLADX. Ce travail doit permettre de vérifier, à l'aide de GLADX, les résultats trouvés par la méthode des GOs, d'approfondir les résultats publiés dans la littérature, de rechercher de nouvelles pertes et de nouveaux pseudogènes, et d'étudier la pseudogénisation.

Chapitre II - Automatisation de l’étude des pertes de gènes unitaires

La détection de perte de gènes unitaires est fondée principalement sur l'annotation des relations des séquences entre elles. La prédiction d'orthologie y joue un rôle central. Bien que l'orthologie soit basée sur des critères phylogénétiques, les méthodes d'annotation les plus utilisées actuellement utilisent la comparaison de séquences deux à deux. Ces méthodes permettent la création de groupes d'orthologues putatifs et, par comparaison de leur contenu avec un arbre d'espèces, permettent de détecter les pertes de gènes unitaires. Les méthodes basées sur les phylogénies sont complexes, difficiles à automatiser et à exploiter, et demandent de longs temps de calcul. Par conséquent, les méthodes phylogénétiques, à quelques exceptions près, ont été très peu utilisées dans les études à grande échelle. Toutefois, elles permettent, comme les méthodes de COGs, de détecter les pertes de gènes unitaires. Les méthodes de COGs ou de phylogénies ont quelques points faibles. Les gènes non annotés sont systématiquement considérés comme perdus, et les pertes ne peuvent jamais être précisées au niveau nucléotidique. D'autres méthodes, peu ou pas du tout automatisées, détectent les pertes de gènes unitaires en se basant sur la présence du « cadavre » des gènes sous forme de pseudogènes. Elles permettent un gain d'information évident sur le phénomène de pseudogénisation. Les pseudogènes ont une durée de vie limitée : la fenêtre de temps pour les observer reste limitée. Des phénomènes récents de pertes par délétions peuvent, en outre, ne pas être détectés.

Pour étudier les pertes de gènes unitaires de manière rigoureuse, j'ai choisi d'utiliser une méthode phylogénétique au lieu d'une méthode basée sur la création de GOs par similarités de séquences. Cette méthode est plus robuste, et le laboratoire a une bonne expérience dans le domaine de l'automatisation de constructions et d'analyses phylogénétiques. Les phylogénies faites à partir de bases de données peuvent mettre en évidence des gènes manquants mais peuvent servir à prouver l'orthologie de séquences non annotées dans les bases de données de gènes et de protéines. Ces séquences orthologues doivent être étudiées au niveau protéique et au niveau génomique afin de déterminer leur nature. De la sorte, il est possible de différencier les gènes putatifs non annotés dans les bases de données utilisées des pseudogènes, et aussi d'analyser les mutations présentes. Une conceptualisation de la démarche doit être faite, afin d'automatiser l'analyse dans un cadre informatique. La méthode développée doit permettre de puiser ses ressources dans les bases de données disponibles, gérer les difficultés causées par une étude à l'échelle des protéomes et des génomes, créer et interpréter des données de niveaux supérieurs. Cet objectif nécessite donc des développements informatiques spécifiques pour répondre au défi de l'exploitation de nombreuses données de natures et de sources différentes. Il nécessite également l'utilisation des dernières avancées de la phylogénétique,

afin d'améliorer les méthodes d'analyse des pertes de gènes unitaires basées sur la phylogénie, et les rendre ainsi beaucoup plus attractives.

1 Matériel

1.1 Les bases de données

Les données nécessaires aux études in-silico utilisées dans cette thèse sont les séquences de type génomique, protéique, EST, etc., et leurs annotations. Toutes ces données sont extraites des bases de données du NCBI, JGI et Ensembl. La base de données Ensembl est particulièrement utilisée car elle contient de riches informations de génomes complets.

1.2 Les outils d'analyses développées au laboratoire

Les recherches menées au laboratoire ont conduit à la création d'outils spécifiques utilisables dans différents contextes.

1.2.1 FIGENIX

FIGENIX est une plateforme bioinformatique, basée sur un système expert qui permet d'automatiser des pipelines biologiques complexes (Gouret *et al.*, 2005 ; Paganini J., 2012). De nombreux pipelines ont été développés en son sein. Le pipeline de création de phylogénie qui a été, par exemple, particulièrement utilisé pour la création du phylome humain dans le cadre du projet EvolHHuPro. Dans le cadre de l'analyse des pertes de gènes unitaires, j'ai utilisé 4 pipelines différents de la plateforme FIGENIX (phylogénie, BLAST, prédiction protéique, reconstruction de séquences ancestrales). Dans le cadre de la modernisation de l'architecture des outils informatiques du laboratoire, FIGENIX a été récemment intégrée au *framework* DAGOBAH (cf. 1.2.3).

1.2.2 Phylopattern

La phylogénie est un élément important utilisé dans les projets de recherche du laboratoire. L'expertise automatisée des phylogénies est donc primordiale. La librairie logicielle PhyloPattern (Gouret, Thompson, & Pontarotti, 2009), qui est une approche à base de règles grammaticales implémentée directement par des règles logiques (non algorithmiques)

permettant la détection de motifs syntaxiques, a été développée au laboratoire afin d'effectuer aisément des analyses, des annotations, et des manipulations d'arbres phylogénétiques. PhyloPattern est présente au sein de DAGOBAH (cf. 1.2.3) et est systématiquement utilisée lors d'expertises phylogénétiques.

1.2.3 DAGOBAH : une infrastructure informatique pluripotente

La volonté du laboratoire d'être capable d'analyser l'ensemble des événements qui caractérisent l'évolution d'un phylum (celui des Chordés dans le projet EvolHHuPro) nécessite d'analyser une quantité très importante d'informations présentes dans les bases de données. Afin de fournir des analyses de qualité pour chaque type d'événement étudié, le système informatique multi-agents appelé DAGOBAH a été développé. Au sein de DAGOBAH chaque agent peut être assimilé à un système expert. Un agent ou un groupe d'agents communiquant entre eux, forment des modules qui peuvent être dédiés à l'analyse d'un type d'événements particuliers. Différents agents peuvent être lancés en parallèle. Les agents peuvent être utilisés par plusieurs modules et servir à des recherches ayant des buts différents. DAGOBAH peut être caractérisé comme un *framework*, un contexte structuré, dans lequel des agents spécifiques peuvent être développés pour répondre à des problématiques ciblées, comme la perte de gènes ou le changement d'architecture en domaines de protéines. Ces agents peuvent communiquer (au travers d'interfaces) avec des bases de données locales ou distantes (NCBI, Ensembl, JGI, etc.). Le système est plus flexible qu'un pipeline. Avant le développement de DAGOBAH, le laboratoire avait créé la plateforme de pipelines appelée FIGENIX permettant la création de pipelines dans un contexte structuré (Gouret *et al.*, 2005). De nombreux pipelines ont été développés permettant d'effectuer des BLAST, des prédictions de gènes, des phylogénies, etc. Ces pipelines produisent des résultats mais sont incapables de les interpréter, à la différence des agents de DAGOBAH qui permettent une expertise poussée grâce aux **règles logiques** qui les composent. Certain agents sont développés spécifiquement pour communiquer avec les pipelines. Ils tirent profit des capacités des pipelines et expertisent leurs résultats.

1.2.3.1 Implémentation des règles logiques

Les agents au sein de DAGOBAH suivent un modèle BDI (*Belief Desire Intention*). Ils possèdent un ensemble de connaissances connues à tout instant (*Belief*), et un ensemble d'objectifs qu'ils cherchent à atteindre (*Desire*). A partir des connaissances dont ils disposent, et des méthodes et actions implémentées, ils cherchent à atteindre leurs objectifs. Pour

formaliser tout cela, la logique du premier ordre est choisie, avec des règles de système expert, en chaînage avant (Gouret, 2009). L'implémentation est faite en langage informatique PROLOG, et son exécution s'appuie sur la libraire GNU PROLOG for JAVA (Plotnikov, 2000).

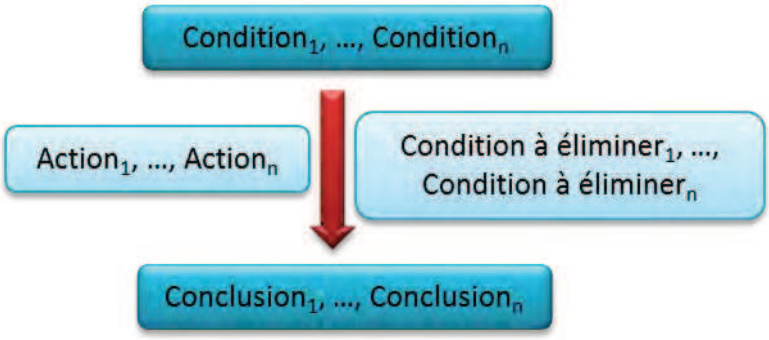


Illustration 18 : Règles de système expert implémentées dans les agents

Lors du fonctionnement d'un agent, le moteur qui infère les règles utilise l'approche suivante : si l'ensemble des conditions est connu, et si au moins une des conclusions n'existe pas, alors les actions sont réalisées et les conditions à éliminer sont détruites, et les nouvelles conclusions obtenues sont de nouveaux faits utilisables par l'agent (Illustration 18). Chaque agent est défini par un ensemble de règles de ce type, et dispose d'une base de connaissances qui évolue de manière dynamique à chaque règle appliquée. Ces règles permettent de formaliser l'expertise humaine.

1.2.3.2 **Les types d'événements analysés par DAGOBAH**

DAGOBAH est développé pour appréhender l'ensemble des événements qui caractérisent l'évolution et permettre l'exploitation des données qui sont de plus en plus nombreuses. L'intégration dans DAGOBAH des concepts de la biologie évolutive et d'une expertise, proche de ce qui peut se faire manuellement, permet d'étudier un large panel d'événements. La force de DAGOBAH réside dans ses capacités à automatiser des expertises humaines, et sa malléabilité permet le développement de modules adaptés à la résolution de nombreuses questions. Il répond de manière efficace aux études à grande échelle. Sa structure permet de factoriser le travail à effectuer et d'évoluer pour répondre à des questions de plus en plus fines en fonction de l'expertise implémentée. Actuellement les agents de DAGOBAH sont capables d'étudier un grand nombre d'événements :

- la synténie

- les changements d'architecture en domaines de protéines (pertes et gains de domaines, échanges homologues et non homologues)

- le transfert horizontal de gènes

- les duplications de gènes

- la perte de gènes et la pseudogénisation

DAGOBAH est aussi utilisé pour d'autres objectifs, tels que la prédiction de caractères (structurels et fonctionnels). La prédiction de caractères ancestraux et contemporains se fait en utilisant les arbres phylogénétiques et des méthodes de maximum de parcimonie (MP), ou de maximum de vraisemblance (ML). Ces méthodes, qui utilisent les caractères connus au niveau des feuilles d'arbres phylogénétiques, permettent de déterminer dans les nœuds internes les caractères ancestraux et d'utiliser les caractères ancestraux pour remonter l'information sur les feuilles non annotées.

Des agents de DAGOBAH permettent de confronter les événements détectés entre eux et essaient de mettre en exergue des processus de coconvergences et de coévolutions.

1.2.3.3 Sauvegarde et utilisation des données produites : la base de données ontologiques

L'étude à grande échelle de tous les événements qui caractérisent l'évolution conduit à manipuler et à créer de nombreuses données hétérogènes qui peuvent être utilisées manuellement et par les agents de DAGOBAH. Dans DAGOBAH les données entrantes peuvent venir de bases de données relationnelles classiques (Ensembl, NCBI, etc.), de fichiers fournis (par exemple des fichiers FASTA) et de bases de données ontologiques associées à DAGOBAH. Les données nécessaires à une seule étude par un agent de DAGOBAH peuvent avoir des origines différentes. Les données produites par la plateforme de pipeline FIGENIX sont stockées dans des bases de données relationnelles et peuvent également être utilisées comme données d'entrée pour effectuer des analyses.

Une ontologie spécifique a été développée au laboratoire, pour fournir une formalisation des concepts biologiques et évolutifs. La base de données ontologique associée à cette ontologie est utilisée par DAGOBAH. Elle permet de stocker les résultats produits et de décrire les relations complexes qui peuvent exister entre les éléments qu'elle contient. Par exemple il est possible d'associer à un événement génétique son impact au niveau protéique. Cette ontologie contient une description fine des données, des événements et des processus biologiques autant

qu'informatiques, sur lesquels se base DAGOBAH pour mener à bien les expertises demandées. L'organisation structurée de termes et de concepts dans l'ontologie permet de donner un sens aux données dans le domaine de la biologie et plus particulièrement en biologie évolutive. Elle est formalisée par une description logique (DL) qui permet de formaliser des relations entre les classes avec des formules logiques (Illustration 19).

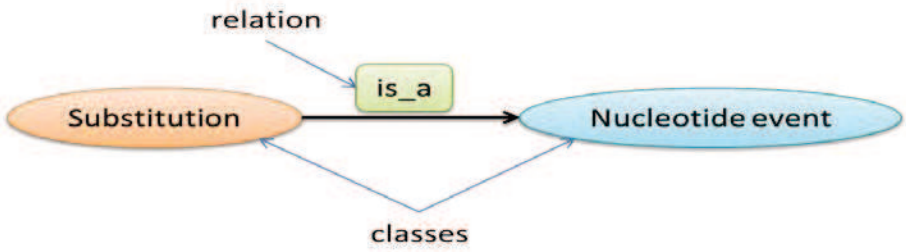


Illustration 19 : Exemple de formalisation ontologique DL

L'ensemble des éléments stockés dans l'ontologie permet de :

- Reprendre une étude après un arrêt de DAGOBAH sans refaire le travail déjà fait
- Pouvoir interroger les données présentes, par le biologiste ou par le système expert
- Utiliser les connaissances déjà acquises pour déterminer d'autres événements, et confronter ces données pour déterminer de la coconvergence d'événements.

L'utilisation de phylogénies permet de dater les événements. La datation des événements se fait sous la forme d'une période définie par deux espèces ou ancêtres entre lesquels les événements ont pu survenir. Les événements et les caractères déduits peuvent être annotés dans un arbre phylogénétique, et c'est cet arbre qui est sauvegardé dans l'ontologie. Les datations peuvent être utilisées dans des analyses corrélatives pour mettre en évidence des cooccurrences d'événements qui peuvent être de types différents.

1.2.3.4 **Synthèse**

Les études entreprises au laboratoire se basent sur 3 éléments essentiels :

- Les bases de données qui sont diverses et variées, locales ou distantes (génomes, protéomes, données fonctionnelles, etc.)
- La plateforme de pipelines qui permet l'automatisation de tâches particulières (prédiction, BLAST, phylogénie, etc.)
- DAGOBAH qui contient les connaissances, les règles, permettant l'expertise. Il joue le rôle de chef d'orchestre.

L'étude des événements qui ont conduit le processus évolutif est possible par automatisation et expertise avec DAGOBAH. Le système multi-agents de cet outil permet de produire des agents robustes dédiés à différentes questions et met à profit, par factorisation, les données produites. Ainsi, au sein de DAGOBAH, une phylogénie peut être utilisée aussi bien pour détecter des pertes, que pour définir des groupes d'orthologues, observer les duplications ou analyser l'architecture en domaines des protéines.

L'approche et le contexte particulier fournis par DAGOBAH, la description de l'ontologie qui l'anime et celle des différents événements étudiés en son sein, sont développés dans l'article 1.

Article 1 - Integration of Evolutionary Biology Concepts for Functional Annotation and Automation of Complex Research in Evolution: The Multi-Agent Software System DAGOBAH

(Springer & Pontarotti Pierre (Ed) 2011)

Chapter 5
Integration of Evolutionary Biology Concepts for Functional Annotation and Automation of Complex Research in Evolution: The Multi-Agent Software System DAGOBAH

Philippe Gouret, Julien Paganini, Jacques Dainat, Dorra Louati, Elodie Darbo, Pierre Pontarotti, and Anthony Levasseur

Abstract Various strategies have been proposed for predicting protein function. They are derived from the classical homology-based approaches and emerging alternative approaches taking into account gene history in the framework of phylogenetic comparative methods. The growing numbers of available genome sequences and data require bioinformatics tools, in which methodological approaches are set according to the biological issues to be addressed. Much effort has already been devoted to integrating evolutionary biology into bioinformatics tools; e.g., homology-based functional annotation has been successfully integrated in a pipeline-assisted method. In addition, new concepts based on correlation of evolutionary events are emerging. For example, two independent events (e.g., systematic loss of specific genes) that happen repetitively can therefore be functionally linked. However, correlated gene profiles, also called "contextual annotation," makes use of different bioinformatics resources based on multi-agent development. In this chapter, we describe evolutionary concepts and bioinformatics approaches proposed for future functional inference.

P. Gouret • J. Paganini • J. Dainat • E. Darbo • P. Pontarotti
UMR6632, Evolutionary Biology and Modeling, Université de Provence, 3 place Victor Hugo, 13331 Marseille, France
e-mail: philippe.gouret@univ-provence.fr

D. Louati
UMR6632, Evolutionary Biology and Modeling, Université de Provence, 3 place Victor Hugo, 13331 Marseille, France

(LAMSIN-IRD) ENIT, Ecole Nationale d'Ingénieurs de Tunis BP 37, Le Belvèdere 1002-Tunis, Tunisia

A. Levasseur
INRA, UMR1163 de Biotechnologie des Champignons Filamenteux, IFR86-BAIM, Universités de Provence et de la Méditerranée, ESIL, 163 avenue de Luminy, CP 925, 13288 Marseille Cedex 09, France

Universités Aix-Marseille 1 et 2, UMR1163, 163 avenue de Luminy, CP925, 13288 Marseille Cedex 09, France

P. Pontarotti (ed.), *Evolutionary Biology – Concepts, Biodiversity, Macroevolution and Genome Evolution*, DOI 10.1007/978-3-642-20763-1_5,

5.1 Functional Annotation Strategies: Current and Future Approaches

5.1.1 Homology-Based Functional Annotation

Eisen was the first to conceptually rationalize phylogenetic methods to improve the accuracy of functional predictions. In 1998, he proposed a phylogenetic prediction of gene function and compared it to similarity-based functional prediction methods (Eisen 1998). In this work, all known functions on a phylogenetic tree were overlaid. The prediction task could then be split into two steps. In the first step, the tree could be used to decipher orthology and paralogy relationships. Most of the reports based on evolutionary biology methods used ortholog information to transfer functional annotation (see Gouret et al. 2005 and Danchin et al. 2007). Functional assignment could be performed for uncharacterized proteins only if the function of an ortholog was known (and if a similar function was evidenced for all characterized orthologs). Ideally, functional inference should be carried out for experimentally validated orthologs. Bibliographic analysis indicates that orthologs are more likely to keep a similar function than paralogs (e.g., Collette et al. 2003). Theoretically, after duplication, one of the copies is lost, or both duplicates undergo subfunctionalization, or one of the duplicates evolves toward a new function (Force et al. 1999). However, Studer has challenged this assumption, as orthologs and paralogs could have comparable mechanisms of divergence (Studer and Robinson-Rechavi 2009). Different and more complex fates of duplicates could also be evidenced (for a review, see Levasseur and Pontarotti 2011).

In the second step, parsimony reconstruction or alternative reconstructive propagation methods could be used to assign functions of uncharacterized genes by identifying the evolutionary scenario that requires the fewest functional changes over time. Inference of ancestral state on phylogenetic tree requires that character mapping be accurate. Uncertainty about trees and mapping is therefore counterbalanced by introducing Bayesian statistical methods, taking into account this inherent error parameter (Ronquist 2004).

To the best of our knowledge, the first report using both approaches was integrated in the work of Engelhardt et al. (2005). The authors constructed a model of molecular function evolution to infer function in a phylogenetic tree. The model takes into account evidence of varying quality and computes a posterior probability for every possible molecular function for each protein in the phylogeny. Different hypotheses were included in the strategy, i.e., each molecular function may evolve from any other function, and a protein's function may evolve more rapidly after duplication events than after speciation events (Engelhardt et al. 2005). Branch length and duplication are integrated in the methodological approach. In brief, methods may be summarized as propagating functional information from leaves to the root of the phylogeny and then propagating back out to the leaves of the phylogeny, based on the probabilistic model of function evolution.

Homology-based functional annotation is summarized in Fig. 5.1.

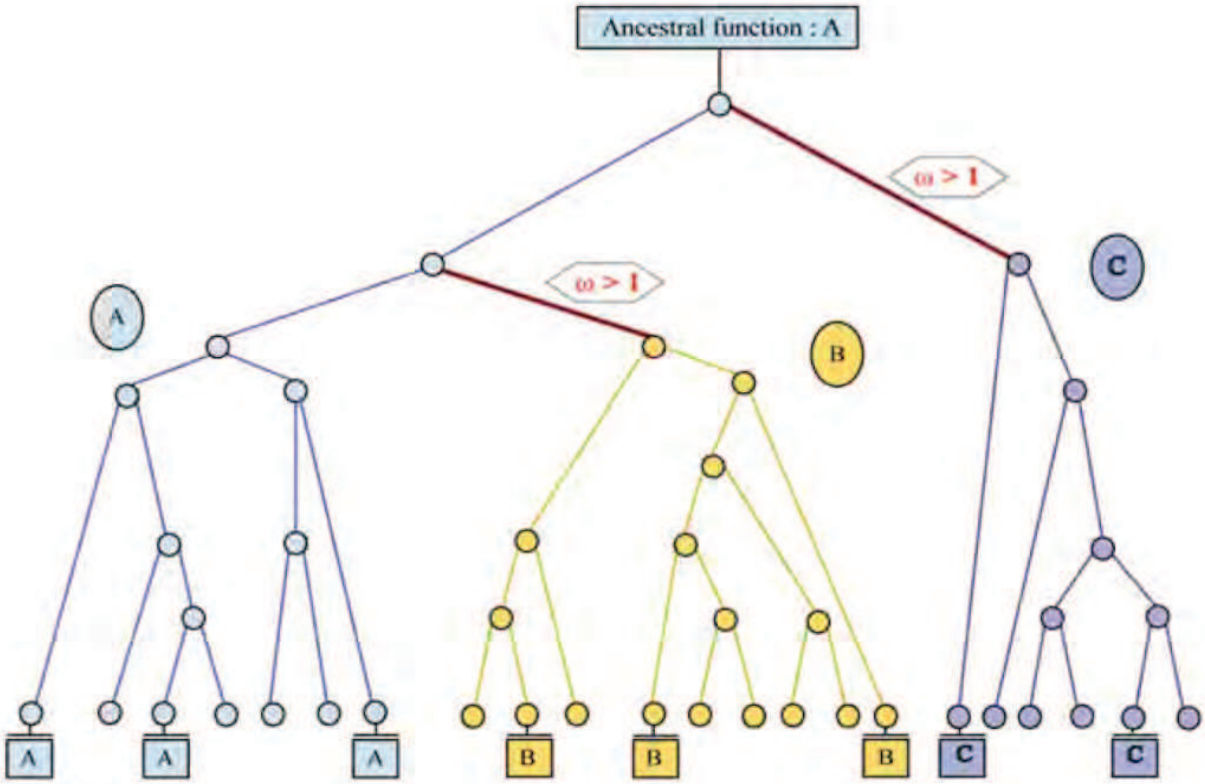


Fig. 5.1 Homology-based functional annotation. Functionally annotated leaves are labeled, respectively, as function A (*blue*), B (*yellow*), and C (*dark blue*). Putative function of non-annotated leaves is inferred after ancestral reconstruction based on propagation of functional information from leaves to the root of the phylogeny. Red branches: evolutionary and functional shift (using $\omega = dN/dS > 1$, i.e., Darwinian selection). (Adapted from Levasseur and Pontarotti 2008)

5.1.2 *Strengthening Functional Annotation: Integration of Correlative Approaches*

Functional prediction using "contextual information" is tricky because of (i) technical difficulty in detecting occurrence profiling and (ii) statistical methods required to correlate and infer function accurately. Co-occurrence and correlated gene profiles could result from phylogenetic inheritance among closely related species. Alternatively, co-occurrence could also result from individual adaptive functions, for instance when genes appear or are lost independently in several distinct lineages (Barker and Pagel 2005). Thus the probability of functional linkage between genes is proportional to the number of multiple independent phylogenetic events. A simplified example of co-occurrence and functional links is depicted in Fig. 5.2. Unlike the overall counting of presence or absence of genes, phylogenetic methods enable us to investigate ancestral states and decipher independent multiple evolutionary events.

Different methods for occurrence profiling have already been proposed, mainly on the basis of the parsimony principle and maximum likelihood (ML).

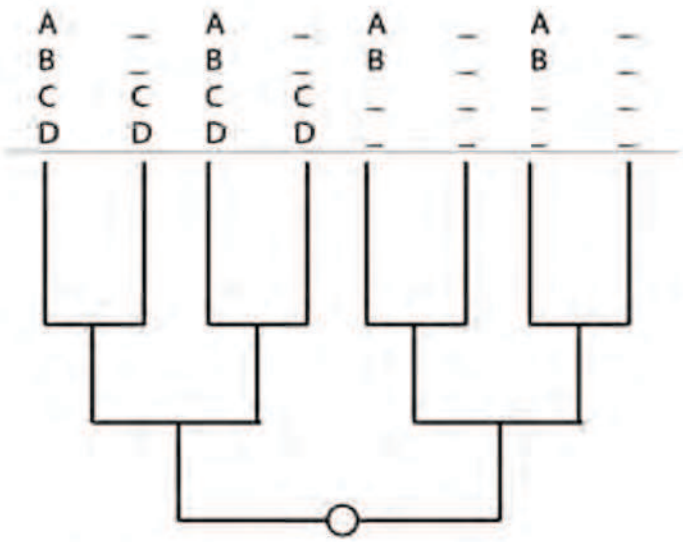


Fig. 5.2 Co-occurrence and functional link. Example of the need for comparative phylogenetic methods. Presence/absence of genes (A, B, C, D) is reported on the leaves of the phylogenetic tree. Here, multiple independent phylogenetic events of gain/loss of gene pairs (i.e., four independent events for genes A and B) are opposed to the apparent correlation arising from shared inheritance of gene pairs loss (resulting from one ancient event for genes C and D). The different steps can be summarized as follows: (i) detection of event: A is lost, (ii) convergence detection: A is lost several times, (iii) co-convergence detection: A and B are lost together several times. Subsequently, statistical tests are carried out. The function of non-annotated genes could be deduced from the correlated annotated genes

As described in the work of Barker and Pagel (2005) and Barker et al. (2007), a common pattern of presence and absence across a range of distinct genomes could be integrated as a method for detecting functionally linked proteins. Thus correlated gains and losses of genes on a phylogenetic tree of species could improve the detection of functionally linked pairs of proteins, compared with the original across-species methods from Pellegrini et al. (1999). Several phylogenetic methods were compared in their work to evaluate the accuracy of their method. Methods were based on either Dollo parsimony (Farris 1977) or ML, including a general model, but also using a constrained model in which the rate of gain of genes is not estimated from the data, but set at a low value. The fixed value of the ML should model gene content evolution better, by preventing the modeling of multiple gains of the same gene in different parts of the phylogeny. In the parsimony case, the reconstructed ancestral states could be very uncertain and parsimony could be applied when rates of changes are rather low. Note that parsimony intervals are proposed to account for the uncertainty of the parsimony methods. For instance, Zhou et al. proposed a dynamic programming algorithm to calculate such parsimony intervals. The best 100 suboptimal ancestral states were determined, and the authors compared the number of correlated events, while allowing for the degree of suboptimality of the reconstructions (Zhou et al. 2006). By contrast, ML accounts for the branch length and uncertainty of topology in the tree, and the estimate of the likelihood values is an independent parameter (i.e., corresponding to all ancestral state possibilities). The authors conclude that all the phylogenetic methods except

unconstrained ML achieved higher specificity than the across-species approach (ML model being capable of greater accuracy and sensitivity than a Dollo parsimony-based approach) (Barker et al. 2007).

5.1.3 *Toward Reliable Global Functional Annotation: The Need for Bioinformatics*

Bioinformatics has unlocked vast amounts of genomic data and developed software applications based on increasingly powerful mathematical algorithms – which themselves produce large volumes of results –, but the amounts of data involved simply cannot be interpreted with any real depth using statistical correlations. We therefore need to develop smart software systems able to support researchers in their efforts, which means systems automatically handling the major routine component of their *in silico* research protocols, and helping analysts interpret the huge volumes of results generated. Such smart software systems could ease the most burdensome part of the workload, leaving researchers to channel their energy into the "sharp end" of their research.

In early 2002, evolutionary biologists were handling vast quantities of biological data made available through the Internet, and running an array of software tools based on probabilistic algorithms working on these data or on data derived from other mathematical tools. The models associated with these tools were all task-specific – sequence similarity, gene prediction, phylogenetic tree-building, and so on. However, they never integrated a large number of concepts employed in biological knowledge and reasoning into a single, integrative software solution. Hence individually, they were unable to answer complex questions posed by biologists or to verify their hypotheses. Consequently, we had to automatically chain mathematical computations through what bioinformaticians call pipelines.

According to the functional annotation strategies described above, homology and correlative approaches were integrated into specific bioinformatics platforms.

A bioinformatics strategy designed for homology-based functional annotation was first implemented by creating FIGENIX (Gouret et al. 2005). FIGENIX is a Java (java.sun.com) platform that automates simple pipeline schemes, such as basic phylogenetic tree-building from a protein sequence by (i) similarity searching against protein databases, (ii) simple filtering, (iii) alignment, and (iv) tree computation. Mathematical tool chaining, through this first version of FIGENIX or any of the pipeline systems available at the time, was unable to completely automate a process: this meant that biologists still had to intervene between computation phases to verify, correct, and synthesize data output from the mathematical tools and guide the workflow to the relevant part of the pipeline. The only way to resolve this automation issue was to introduce an expert system (with Prolog language; Warren et al. 1977) into FIGENIX to model a part of biologists' knowledge and thus act as a human scientist as and when necessary. By introducing specific logical rules in the expert system, a pipeline was created and was dedicated to gene

predictions *via* an approach combining ab initio predictions and homology through a lab method. Tested against a known benchmark, the pipeline clearly proved successful. A complex phylogeny pipeline with 50 steps and a lot of expertise modeling was designed. The first version was stabilized in late 2003, and has since enabled the laboratory and its collaborators to produce thousands of phylogenic trees from protein queries. These trees form the basis of our evolutionary research. This pipeline, along with others, was intensively used on laboratory projects, generating several published papers (Danchin et al. 2004, 2006, 2007; Paillisson et al. 2007; Levasseur et al. 2006, 2010). It continued to undergo improvements and enhancements, with upgrades including automatic detection of orthologs in the final process-synthesized tree by online recovery of functional data associated with these orthologs (GO (Ashburner et al. 2000), MGI (www.informatics.jax.org), NCBI (www.ncbi.nlm.nih.gov)), and EST integration (Balandraud et al. 2005). Part of the software developed, called PhyloPattern, emerged as a crucial independent component (Gouret et al. 2009). The aim of this tool was to reproduce human reading of phylogenetic trees, i.e., phylogenetic tree annotation and pattern recognition. Inside the phylogeny pipeline, this tool is used to detect incongruence or isolate specific subtrees, from which biases are then corrected. PhyloPattern now makes it possible to detect events in the history of species, genes, or any other characteristic (from domain to function and further), as well as highlighting artifacts in the phylogenetic trees. We are continuing to improve PhyloPattern as a free open-source JAVA/Prolog API.

5.2 From Pipelines to Multi-Agent Strategies

In 2005, it became clear that the "pipeline approach," even with the controlling expertise introduced, remained limited to computation processes. In addition, functional annotation using the correlative approaches strategy required flexible and more sophisticated data processing architecture. Computation processes are essential, but are not really able to resolve complex tasks of interest to the laboratory, such as automatically highlighting genetic events in the human genome and detecting convergences and co-convergences among these events. Any solution to these issues needs to be driven by expertise through parallel and more "intelligent" processes than the rigid, deterministic pipelines. We also note that the "pipeline approach" does not extend to establishing an explicitly described semantic universe that would allow accurate meta descriptions of data. It thus remains impossible to raise the abstraction level of software tasks, and interfacing them with other software systems is not natural.

Integration of correlated gene profiles for functional annotation requires a three-step process: (i) specific detection of all evolutionary events, (ii) correlation using phylogenetic comparative methods leading to a compelling statistical results, and (iii) deducing the function of non-annotated genes from the correlated annotated genes.

5.3 Technical System Specifications

Accordingly, a new software system was conceived and is able to implement complete automation of actual full research via bottom-up (from biological data) strategies specified by the laboratory, rather than "just" complex computation workflows. We opted for the following research strategy: (i) working from known or computed features to find evidence for generating new hypotheses, (ii) attempting to verify hypotheses to transform them into features, (iii) correlating verified features to deduce new features, and so on. A set of characteristic specifications was drawn up:

- The treatments had to be flexible, modular, and parallelized.
- The strategies for identifying and verifying the facts had to be led by expertise.
- Communication with external software systems (online databases, web services) should systematically gather the relevant results produced by these platforms, such as Ensembl (Hubbard et al. 2009), NCBI, String (Szklarczyk et al. 2011), and ArrayExpress (Parkinson et al. 2011).
- The results had to be placed in an accurately described semantic universe that was not redundant but interfaced with data from external systems.
- Some modules had to work together and communicate directly, while others, such as modules for intelligent correlations of events, had to work in stand-alone mode directly on the mass of results produced by the full set of modules.
- The modules had also be able to work at different times.
- The system had to be resistant to failure; as such, very costly computational treatments should have to be run only once.

5.4 Technical State of the Art

The field of biology now has a number of software tools, approaches, standards, and publications that could be recycled for our needs. The type of system targeted here required establishing an integrated data model, placed between structured biological data (e.g., genomic databases) or unstructured data (publications) located inside or outside the laboratory, and the research strategies desired by laboratory researchers. Software systems clearly have to work with large-scale data banks, but what is most important now is to work with different kinds of data, many of which are not a direct representation of biological objects but are more abstract concepts.

We could therefore rule out relational database management systems, which are not powerful enough or flexible enough to describe semantics in biology. Some recently developed software tools such as the alignment expert system ALEXSYS (Aniba et al. 2009) are based on the UIMA framework (http://sourceforge.net/projects/uima-framework/), which offers a powerful architecture and is well-suited to the introduction of a virtual model on unstructured data, i.e., building meta-information from artifacts such as scientific publications (also see DiscoveryLink

(Hass et al. 2001) or BioMOBY (Wilkinson and Links 2002)). We are more focused on trying to directly model actual genomics or evolutionary concepts. Also, the UIMA approach is only "object-oriented," and we believe that this kind of modeling architecture is not rich enough to integrate the complexity of biological paradigms, especially compared with approaches based on mathematical first-order logic ontology techniques such as Description Logic (http://dl.kr.org/). The W3C-standardized OWL language (http://www.w3.org/TR/owl-ref/) is an XML representation of DL. Initially applied to the semantic web, it is fast becoming a standard for ontology modeling. In DL, relations between classes are not limited to aggregation or inheritance links but can be formalized with logical formulae. However, we note that DL does not integrate concepts of inductive, temporal, or fuzzy logic, which in the long term could direct the natural extension of our systems.

Biology now has many ontologies (e.g., NCI Cancer Ontology: http://www.mindswap.org/2003/CancerOntology/). Some are defined in OWL but to our knowledge, none computationally exploit the descriptive capacity of description logic (DL). This situation is surely set to change. We note the existence of relational ontology (Smith et al. 2005), placed between "object" modeling and DL modeling, which attempts to standardize relations in biological ontologies. This point will be revisited below. There appears to be a continuing dichotomy between the activity of defining ontologies, considered as vocabularies by many biologists, and the establishment of DL-based software and databases within and between laboratories or institutes. We believe that this dichotomy is an error, as it has very adverse repercussions, such as poor software systems and bad interoperability.

As stated above, to fully automate *in silico* research strategies, the type of system we are targeting has to be less rigid and deterministic than pipelines. A natural candidate solution would be multi-agent systems. In bioinformatics, these systems are used essentially to model and simulate biological networks (reactive agents), although they are also used to parallelize mathematical computations through agents with very fine granularity. They are rarely employed for building integrative applications where "smart" agents work with biological information. Nevertheless, like the FIPA institute (http://www.fipa.org/), we are convinced that this kind of architecture built from cognitive agents (with large granularity) communicating inside an ontological semantic universe can be applied to bioinformatics automation. The JADE software framework (http://jade.tilab.com/) is a Java implementation of FIPA specifications. At our lab, we used JADE to develop a first prototype multi-agent system named CASSIOPE (Rascol et al. 2009), dedicated to highlighting conserved synteny.

Recently, eHive emerged from EBI as a new workflow system (Severin et al. 2010). It is built as a multi-agent "blackboard" architecture. Here, the blackboard, i.e., the communication area between agents, is reduced to chaining rules between agents. Thus the tasks produced by the system are driven by predefined functional relations between agents and not by the autonomous interpretation, by agents, of the data resulting from other agents' work. The Ehive blackboard database has a rigid structure with no data modeling. Also, agents' source code is written with the Perl

language, which albeit very widely used in bioinformatics remains very poor in expertise and knowledge modeling.

As stated earlier, we are seeking to deploy expertise-driven research strategies, which means that all agents need to be built with expert-system architectures. Rule engines do exist – one example is Jess (www.jessrules.com) – but it would be preferable to write our own engine in Prolog language to reap the benefit of tools we developed previously, especially PhyloPattern. After years of hands-on experience, we can confirm that the Prolog language is very well-suited to bioinformatics. Its benefits for the target system include: (i) a natural capacity to generate all the solutions for a question, (ii) easy and native manipulations of lists and tree structures, which are intensively used in bioinformatics data, (iii) development of expert systems in backward- and/or forward-chaining mode (verification and/or production of facts), (iv) formalisms (e.g., ontological relations) representable directly in the language's syntax, (v) brevity and simplicity of knowledge descriptions, and (vi) interpreted language that strengthens the experimental aspect of certain developments.

5.5 System Architecture

Our system was called DAGOBAH. It is shaped as a multi-agent software (see Fig. 5.3), with a voluntarily hybrid model summing of a model called "Belief Desire Intention" with a model called "Blackboard" (Ferber 1995). The BDI model is suitable for cognitive agents with high granularity and therefore high "intelligence." In the BDI model, agents have a plan formed for our purposes by logical rules. This highly flexible rule system is used by each agent to implement a specific strategy, but can also be used as a traditional expert system to produce high-level facts deduced from simpler facts. For example, an agent capable of sifting through actions to detect several equally probable genetic events from a phylogenetic tree will be able to retain only one event, through a set of logical rules associated with a set of criteria.

The semantics for one rule is defined as follows:

- $Action_1 \ldots Action_k$
  $ConditionFact_1 \ldots ConditionFact_n \rightarrow ConclusionFact_1 \ldots ConclusionFact_m$
  $ToBeRemovedFact_1 \ldots ToBeRemovedFact_z$

The meaning is "if all condition facts (n) are known by the agent ($\subset$ Belief) and if at least one of the conclusion facts (m) is not present and if the agent is capable of achieving all actions (k) ($\subset$ Intention) successfully, then all conclusions (m) ($\subset$ Desire) are considered truthful, and all indicated facts (z) are removed from the agent's knowledge."

Here is an example rule, used in the DAGOBAH agent dedicated to searching for domain architecture events. We suppose that for a specific protein with the domain architecture A-B-C, DAGOBAH detects an event that produced the B-C part of the architecture by analyzing the phylogenetic tree of domain B, and we suppose

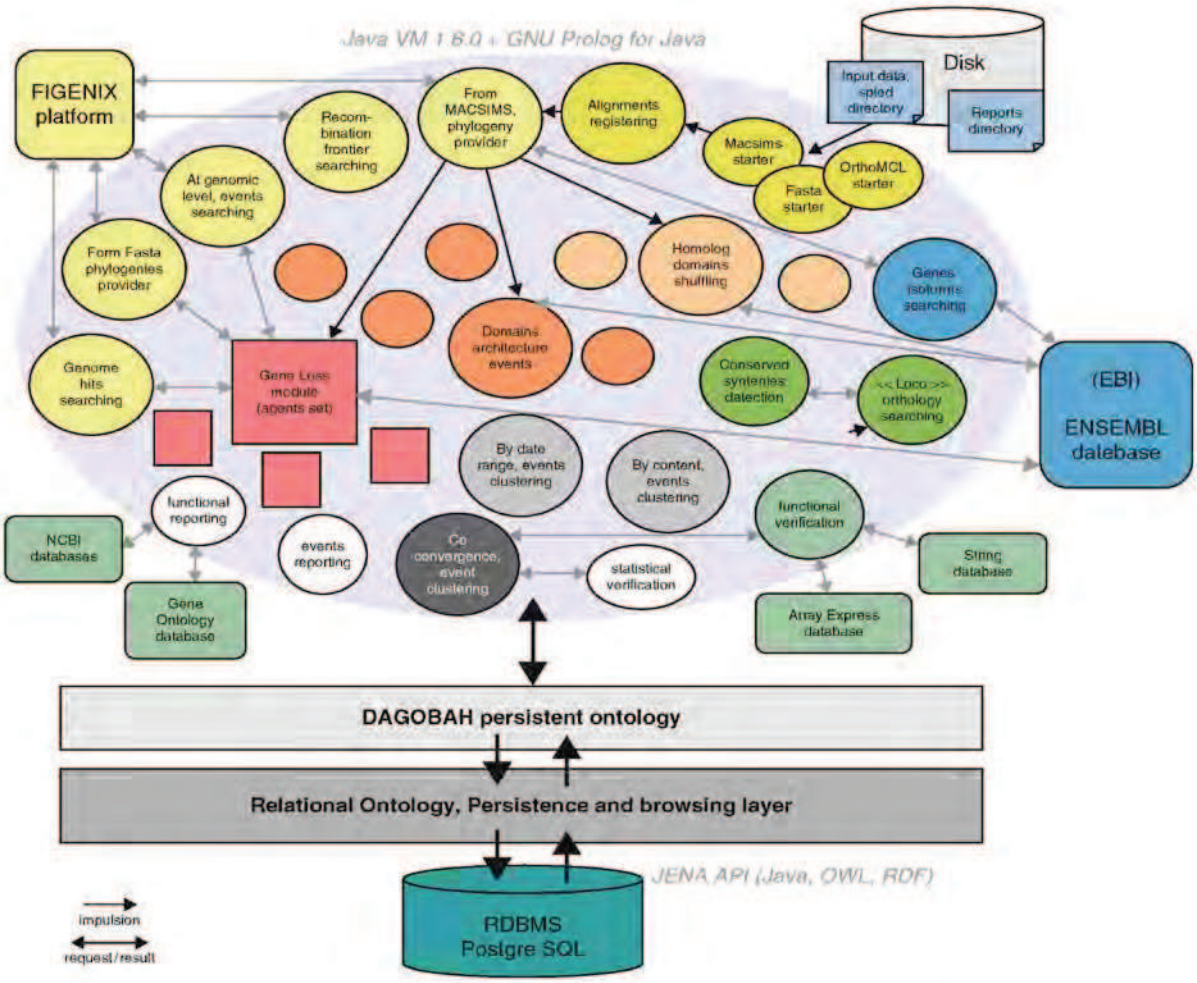


Fig. 5.3 DAGOBAH multi-agents system architecture. All agents (*disks*) or modules (*squares*) (set of agents) that compose DAGOBAH are contained in the large blue ovoid. Around it are displayed the external software systems interacting with the agents by the network. At the bottom of the scheme is shown the ontological database, containing the biological results produced and shared by the agents

that DAGOBAH hesitates between identifying the event as a shuffling or a gain. A simple rule, if it is applicable, allows DAGOBAH to definitely assert there is a gain (see Fig. 5.4):

- *verify_similarity_of_signal_between(P1, P2, [B, C])*
  event_found_under_ancestral_node(N),
  apomorphic_chosen_protein(P1, [A, B, C]), → *gain_event_found(N, P1, [C])*
  plesiomorphic_chosen_protein(P2, [A, B])
  event_found_under_ancestral_node(N)

The "Blackboard" model introduces an area of information shared by agents, i.e., any important result produced by an agent is placed on the blackboard. The blackboard architectural model chosen in DAGOBAH is defined as a persistent ontology (an ontological database) representing the semantic universe in which the agents work. These results are used by other agents, unless they are forced to

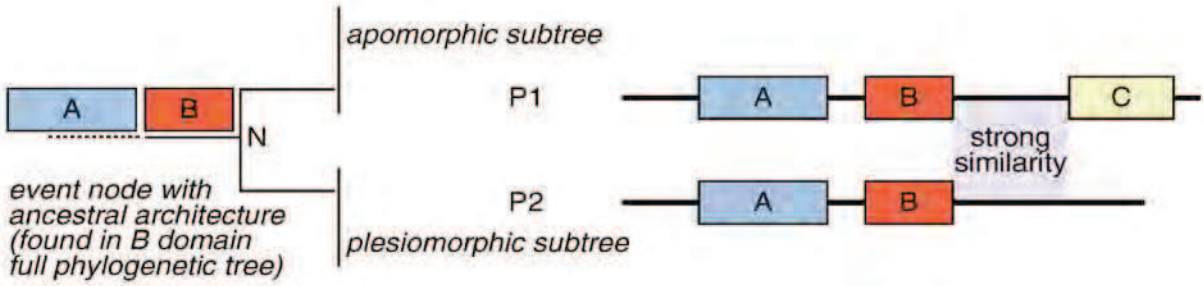


Fig. 5.4 A virtual example for a domains architecture event. Here again event is confirmed because the genomic signal between domains B and C on the apomorphic sequence is strongly conserved after domain B on the plesiomorphic sequence

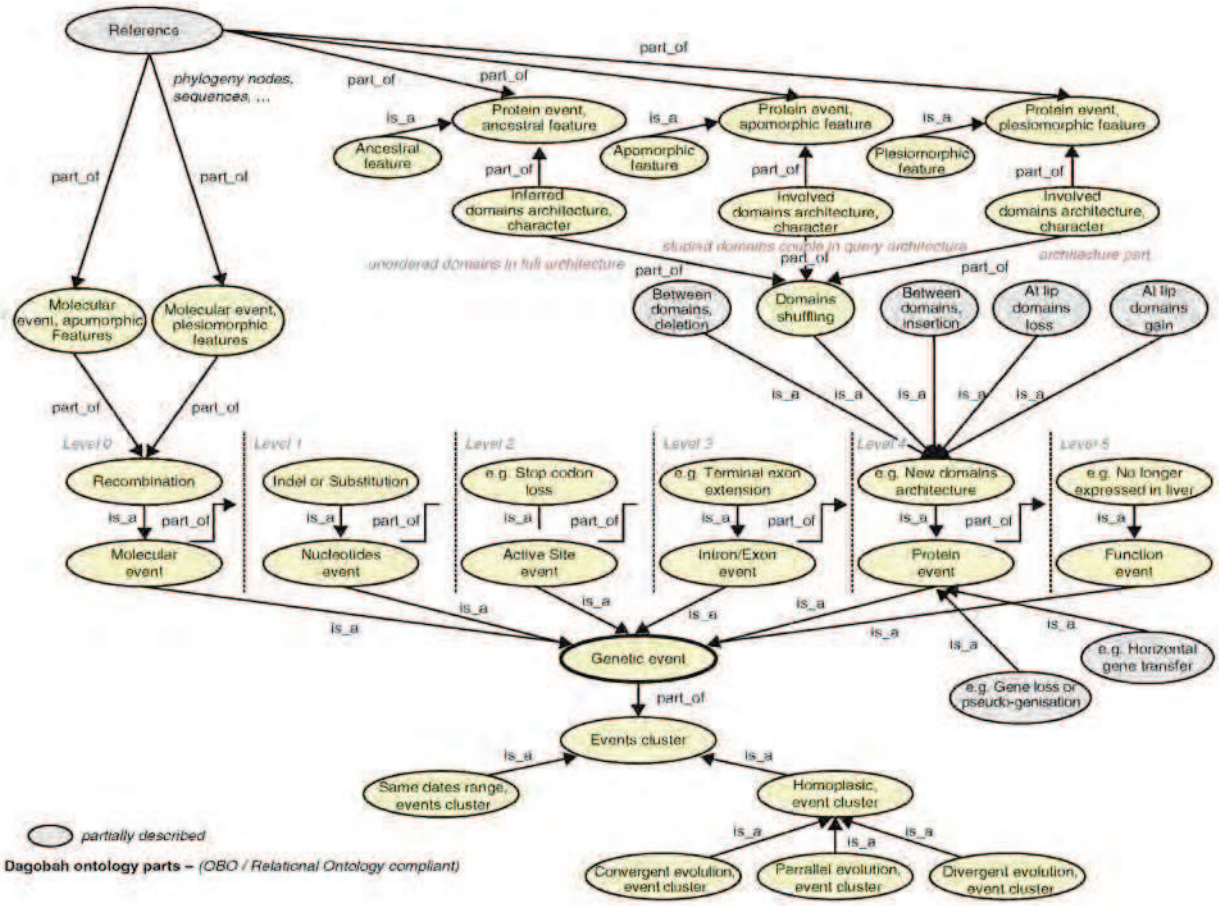


Fig. 5.5 The core of DAGOBAH ontology. Some genetic event classes laid out by their reading level are presented. As an example, we give all the classes participating in a nonhomologous domain shuffling event, induced by a recombination event. Clustering classes are also displayed with their inheritance relationships

explicitly and systematically exchange them. Figure 5.5 illustrates the main parts of the DAGOBAH ontology. Genetic event classes are grouped by reading level. For example, a recombination event can be described at a "protein" level if we are talking about domains involved in recombination, but also at a "molecular" level if we are talking about the position of the recombination on a chromosomal region. Ancestral, apomorphic, and plesiomorphic features associated with an event are

always explicitly expressed. This model is also particularly well-suited to studying automatic correlations of genetic events, and is able to correlate several events detected by DAGOBAH and temporally localized between speciation event pairs. For example, DAGOBAH may find that two genes A and B are lost twice "together" for two different lineages, which could prove very interesting in a functional perspective. In this case, if the "function" of gene A is known and the "function" of B is not, we can assume that the B gene may be involved in the "same" function as A. "By Dates" event clusters and homoplasic event clusters are the sources of a co-convergent event clustering process in DAGOBAH. For example, a "convergent evolution event cluster" is produced for events that have the same apomorphic feature objects.

The DAGOBAH ontological database must not have redundancy vs. external databases (like Ensembl; Hubbard et al. 2009). Consequently, we only model, by classes and relations, those concepts associated with specific laboratory research themes, and references were kept only to biological data or results held in external databases. The current DAGOBAH ontology adopts the Relational Ontology standard, although in the future we will probably abandon this standard so as to fully exploit the capabilities of Description Logic.

5.6 DAGOBAH Functionalities and Summarized Strategies

As described in Fig. 5.2, the strategies used in DAGOBAH can be conceptually subdivided into these different steps: (i) detection of evolutionary events, i.e., gain or loss of genes, shuffling, etc. (ii) detection of convergence between one or more gene pairs, (iii) detection of co-convergence between linked genes, (iv) search for functionally annotated gene and infer the function of correlated non-annotated gene. These four steps can be considered as forming the core of the phylogenetic comparative methods.

5.7 Detection of Events (New Architecture Appearance)

The current DAGOBAH version offers a broad panel of functions, ranging from automatic detection of genetic events to homologous domain shuffling, nonhomologous domain shuffling, insertion, deletion, gain and loss, plus gene losses and pseudogenization, and on to horizontal gene transfer and duplications (compilation on gene and species trees). A simplified summary of DAGOBAH's general strategy for event detection is:

1. Use "domain-annotated" protein alignments built from a query protein to outsource phylogeny trees building (domain trees and protein trees) to the FIGENIX platform.
2. Automatically read these trees with PhyloPattern to highlight possible events.

3. Seek to verify and clarify the putative events at a genomic level.

For new protein domain architecture events, actual examples of putative events in trees are given in the PhyloPattern publication. For this kind of event, a dedicated DAGOBAH agent studies each consecutive domain pair in the query protein architecture to investigate whether the association is the result of an event. Ideally, it finds an event's phylogenetic pattern (see Fig. 5.4) on each domain phylogenetic tree, which strengthens the event hypothesis.

The full confirmation of the event is achieved at genomic level by searching for an alignment break position between two DNA segments – one associated with the most representative apomorphic sequence and the other associated with the most representative plesiomorphic sequence. DNA segments are extracted between the domains involved (see Fig. 5.6). The most representative apomorphic sequence is chosen as the one nearest the parent node (the agent uses neighbor joining for branch lengths), while the most representative plesiomorphic sequence is chosen as the one whose domain architecture is closest to the ancestral node architecture (Dollo, Sankoff, and Mirkin parsimony algorithms (Sankoff 1975; Farris 1977; Mirkin et al. 2003) are integrated into PhyloPattern and used by the agent to infer ancestral domain architectures). If several plesiomorphic sequences share the same architecture comparison "score," the agent chooses a sequence from the nearest species in the species tree.

Gene losses and pseudogenization are studied by a set of agents in DAGOBAH, which form a module named GeneLoss. It starts the study by searching for missing species in the biggest ortholog group of the query protein tree. Each species is then studied by independent agents.

Describing the strategy in schematic terms, agents set out to determine whether the species is really missing, whether a new gene should be annotated, or whether there are some mutations or indels that can explain a pseudogenization process.

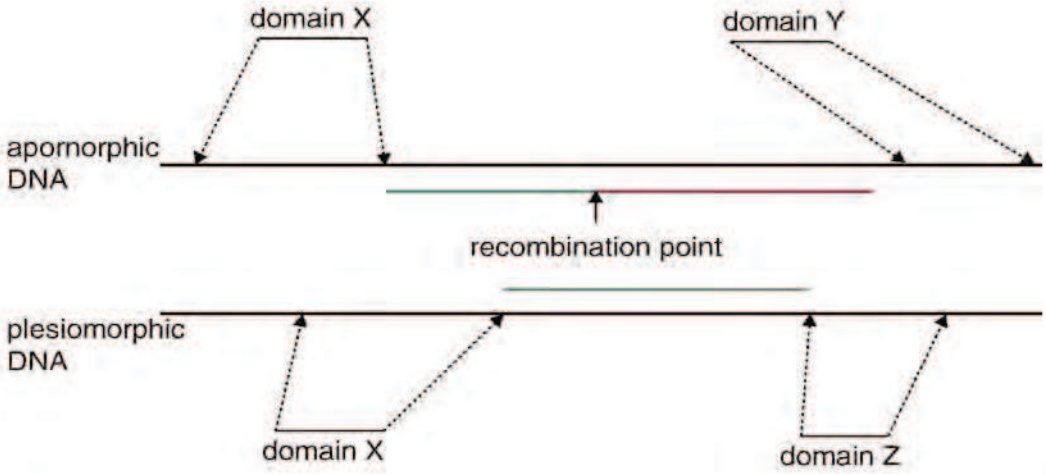


Fig. 5.6 Summary of the verification of a domain new architecture event at a genomic level. The DNA segments between domains on the apomorphic and the plesiomorphic sequences are intelligently extracted from chromosomes or scaffolds; they are then aligned and the recombination point is searched for as an alignment break

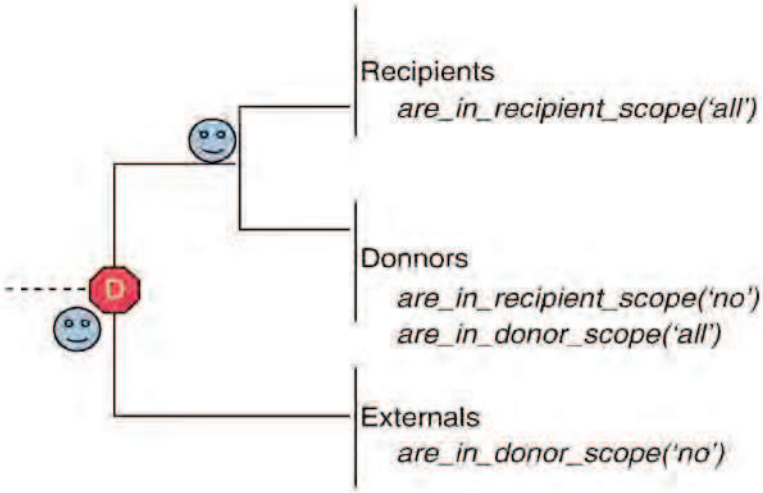


Fig. 5.7 A pattern to detect horizontal gene transfers from a phylogenetic gene tree. This means a duplication node, because the subtree does not have to match the species tree. The "donor" subtree must contain only species of a specific scope, and not from the "recipient" scope and *vice versa*

Full complex GeneLoss module strategy and results will be published separately at a later date.

Horizontal gene transfer events are detected from the query protein tree. A recipient species scope and a donor species scope are defined so as to orient the search. The dedicated agent uses PhyloPattern to annotate each internal node of the tree with two tags: are_in_recipient_scope_species and are_in_donor_scope_species, which can take three values: "no" if no species of a subtree falls in a scope, "some" if some species of a subtree fall in a scope, or "all" if all the species of a subtree fall in a scope. Then, *via* PhyloPattern, the agent applies a specific phylogenetic pattern (see Fig. 5.7) that directly gives the branch with potential HGT events.

The expert idea behind this pattern is to search the gene tree to find recipient species closer to donor species than other species that are normally placed between the recipient and donor species in the species tree.

5.8 Convergence and Co-Convergence Detection

Another important function in DAGOBAH is event convergence and co-convergence detection as conceptually described in the correlative approaches described above. Convergence identification is easy to obtain from the DAGOBAH ontological database, as a dedicated agent groups events into homoplasic convergent clusters. For example, two events are in the same convergent cluster if they have the same apomorphic character. The definition of an apomorphic character can easily be user-defined as a Prolog "ontological" pattern. The clustering mechanism is independent of the pattern definition. Co-convergence detection is a more complex task. It starts by homoplasic clustering, after which an agent produces date range clustering. Inside DAGOBAH, events are dated with tuples:

[TaxidSpeciationBefore, NumberOfDuplicationsBefore, NumberOfDuplicationsAfter, TaxidSpeciationAfter]

This tuple is determined by taking the nearest speciation event (SBE) before the event (E) on its parent branch. NumberOfDuplicationsBefore equals the number of duplication events on the branch between SBE and E. TaxidSpeciationBefore is the common parent taxid of all species in the SBE subtree. The same approach is then reapplied for the next speciation event. Date range clustering is also "user-defined" through date range patterns. Two events whose dates fit the same date pattern are pooled in the same date range cluster.

Co-convergence clusters are built with a hierarchical clustering method. A minimum co-convergent cluster is formed by four events: Eh1, Eh2, Eh1′, Eh2′. Eh1 and Eh1′ have to be in the same homoplasic cluster, while Eh2 and Eh2′ have to be in another homoplasic cluster. Eh1 and Eh2 have to be in the same date range cluster, while Eh1′ and Eh2′ have to be in another date range cluster.

We can model this basic cluster as a square:

--- Eh1, Eh2,

--- Eh1′, Eh2′

The clusters can be rectangular, if they come from more date clusters than homoplasic clusters (shape 1) or the opposite (shape 2). The hierarchical clustering method enables us to build the biggest possible clusters, and implies the definition of a distance method between two clusters. Our distance method favors clusters with shape 1 rather than shape 2.

Once the biggest clusters are determined, the agents seek to verify them, both statistically, *via* the Pagel method (Pagel 1994), and functionally, using the String database (Szklarczyk et al. 2011) to see whether proteins associated with events in the same homoplasic cluster belong to the same protein interactions network, and using the ArrayExpress database (Parkinson et al. 2011) to see whether proteins associated with events in the same homoplasic cluster concern the same expression experiments.

In conclusion, DAGOBAH is designed to exploit the modern functional annotation strategies and specially the evolutionary-based biology concepts. In addition, it could be addressed to various general biological questions such as searches of conserved synteny regions from a given region associated to a species to another target species.

All public results produced by DAGOBAH are openly available on the IODA Web site (http://ioda.univ-provence.fr/).

Acknowledgments This research was supported by the contract MIE (Maladies Infectieuses Emergentes-Programme Interdisciplinaire, CNRS) and ANR EvolHHuPro (ANR-07-BLAN-0054-01).

References

Aniba MR, Siguenza S, Friedrich A, Plewniak F, Poch O, Marchler-Bauer A, Thompson JD (2009) Knowledge-based expert systems and a proof-of-concept case study for multiple sequence alignment construction and analysis. Brief Bioinform 10:11–23

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G (2000) Gene ontology: tool for the unification of biology. The gene ontology consortium. Nat Genet 25:25–29

Balandraud N, Gouret P, Danchin EG, Blanc M, Zinn D, Roudier J, Pontarotti P (2005) A rigorous method for multigenic families' functional annotation: the peptidyl arginine deiminase (PADs) proteins family example. BMC Genomics 6:153

Barker D, Pagel M (2005) Predicting functional gene links from phylogenetic-statistical analyses of whole genomes. PLoS Comput Biol 1:e3

Barker D, Meade A, 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

Collette Y, Gilles A, Pontarotti P, Olive D (2003) A co-evolution perspective of the TNFSF and TNFRSF families in the immune system. Trends Immunol 24:387–394

Danchin E, Vitiello V, Vienne A, Richard O, Gouret P, McDermott MF, Pontarotti P (2004) The major histocompatibility complex origin. Immunol Rev 198:216–232

Danchin EG, Gouret P, Pontarotti P (2006) Eleven ancestral gene families lost in mammals and vertebrates while otherwise universally conserved in animals. BMC Evol Biol 6:5

Danchin EG, Levasseur A, Rascol VL, Gouret P, Pontarotti P (2007) The use of evolutionary biology concepts for genome annotation. J Exp Zool B Mol Dev Evol 308:26–36

Eisen JA (1998) Phylogenomics: improving functional predictions for uncharacterized genes by evolutionary analysis. Genome Res 8:163–167

Engelhardt BE, Jordan MI, Muratore KE, Brenner SE (2005) Protein molecular function prediction by Bayesian phylogenomics. PLoS Comput Biol 1:e45

Farris JS (1977) Phylogenetic analysis under Dollo's law. Syst Zool 26:77–88

Ferber J (1995) Les systèmes multi-agents. InterEdition, Paris

Force A, Lynch M, Pickett FB, Amores A, Yan YL, Postlethwait J (1999) Preservation of duplicate genes by complementary, degenerative mutations. Genetics 151:1531–1545

Gouret P, Vitiello V, Balandraud N, Gilles A, Pontarotti P, Danchin EG (2005) FIGENIX: intelligent automation of genomic annotation: expertise integration in a new software platform. BMC Bioinform 6:198

Gouret P, Thompson JD, Pontarotti P (2009) PhyloPattern: regular expressions to identify complex patterns in phylogenetic trees. BMC Bioinform 19 10:298

Haas LM, Schwarz, Kodali P, Kotlar E, Rice JE, Swope WC (2001) DiscoveryLink: A system for integrated access to life sciences data sources. IBMSJ 40:489–511.

Hubbard TJ, Aken BL, Ayling S, Ballester B, Beal K, Bragin E, Brent S, Chen Y, Clapham P, Clarke L, Coates G, Fairley S, Fitzgerald S, Fernandez-Banet J, Gordon L, Graf S, Haider S, Hammond M, Holland R, Howe K, Jenkinson A, Johnson N, Kahari A, Keefe D, Keenan S, Kinsella R, Kokocinski F, Kulesha E, Lawson D, Longden I, Megy K, Meidl P, Overduin B, Parker A, Pritchard B, Rios D, Schuster M, Slater G, Smedley D, Spooner W, Spudich G, Trevanion S, Vilella A, Vogel J, White S, Wilder S, Zadissa A, Birney E, Cunningham F, Curwen V, Durbin R, Fernandez-Suarez XM, Herrero J, Kasprzyk A, Proctor G, Smith J, Searle S, Flicek P (2009) Ensembl. Nucleic Acids Res 37:D690–D697

Levasseur A, Pontarotti P (2008) An overview of evolutionary biology concepts for functional annotation: advances and challenges. In: Pontarotti P (ed) Evolutionary biology from concept to application. Springer, Berlin, pp 209–215

Levasseur A, Pontarotti P (2011) The role of duplications in the evolution of genomes highlights the need for evolutionary-based approaches in comparative genomics. Biol Direct 6:11

Levasseur A, Gouret P, Lesage-Meessen L, Asther M, Asther M, Record E, Pontarotti P (2006) Tracking the connection between evolutionary and functional shifts using the fungal lipase/feruloyl esterase a family. BMC Evol Biol 6:92

Levasseur A, Saloheimo M, Navarro D, Andberg M, Pontarotti P, Kruus K, Record E (2010) Exploring laccase-like multicopper oxidase genes from the ascomycete trichoderma reesei: a functional, phylogenetic and evolutionary study. BMC Biochem 11:32

Mirkin BG, Fenner TI, Galperin MY, Koonin EV (2003) Algorithms for computing parsimonious evolutionary scenarios for genome evolution, the last universal common ancestor and dominance of horizontal gene transfer in the evolution of prokaryotes. BMC Evol Biol 3:2

Pagel M (1994) Detecting correlated evolution on phylogenies: a general method for the comparative analysis of discrete characters. Proc R Soc Lond B 255:37–45

Paillisson A, Levasseur A, Gouret P, Callebaut I, Bontoux M, Pontarotti P, Monget P (2007) Bromodomain testis-specific protein is expressed in mouse oocyte and evolves faster than its ubiquitously expressed paralogs BRD2, -3, and -4. Genomics 89:215–223

Parkinson H, Sarkans U, Kolesnikov N, Abeygunawardena N, Burdett T, Dylag M, Emam I, Farne A, Hastings E, Holloway E, Kurbatova N, Lukk M, Malone J, Mani R, Pilicheva E, Rustici G, Sharma A, Williams E, Adamusiak T, Brandizi M, Sklyar N, Brazma A (2011) ArrayExpress update–an archive of microarray and high-throughput sequencing-based functional genomics experiments. Nucleic Acids Res 39:D1002–D1004

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 USA 96:4285–4288

Rascol VL, Levasseur A, Chabrol O, Grusea S, Gouret P, Danchin EG, Pontarotti P (2009) CASSIOPE: an expert system for conserved regions searches. BMC Bioinform 10:284

Ronquist F (2004) Bayesian inference of character evolution. Trends Ecol Evol 19:475–481

Sankoff D (1975) Minimal mutation trees of sequences. SIAM J Appl Math 28:35–42

Severin J, Beal K, Vilella AJ, Fitzgerald S, Schuster M, Gordon L, Ureta-Vidal A, Flicek P, Herrero J (2010) eHive: an artificial intelligence workflow system for genomic analysis. BMC Bioinform 11:240

Smith B, Ceusters W, Klagges B, Köhler J, Kumar A, Lomax J, Mungall C, Neuhaus F, Rector AL, Rosse C (2005) Relations in biomedical ontologies. Genome Biol 6:R46

Studer RA, Robinson-Rechavi M (2009) How confident can we be that orthologs are similar, but paralogs differ? Trends Genet 25:210–216

Szklarczyk D, Franceschini A, Kuhn M, Simonovic M, Roth A, Minguez P, Doerks T, Stark M, Muller J, Bork P, Jensen LJ, von Mering C (2011) The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. Nucleic Acids Res 39: D561–D568

Warren DH, Pereira LM, Pereira F (1977) Prolog - the language and its implementation compared with Lisp. Proceedings of the 1977 symposium on artificial intelligence and programming languages

Wilkinson MD, Links M (2002) BioMOBY: an open source biological web services proposal. Brief Bioinform 3:331–341

Zhou Y, Wang R, Li L, Xia XF, Sun Z (2006) Inferring functional linkages between proteins from evolutionary scenarios. J Mol Biol 359:1150–1159

2 Développement de GLADX : Module de DAGOBAH dédié à l'analyse des pertes de gènes unitaires

Pour améliorer les connaissances concernant les pertes de gènes unitaires, il est nécessaire de créer un outil performant, automatisé, utilisant une annotation de qualité de l'orthologie et capable d'identifier les pseudogènes unitaires parmi les pertes détectées. A l'aide des outils disponibles, une première version d'un outil totalement automatisé, capable de supporter des études à l'échelle des génomes, a été développée par Phillipe Gouret (Gouret, 2009). J'ai participé aux spécifications de l'outil et réalisé la vérification biologique des résultats produits. A partir des tests et des limites identifiées sur cette version, nous avons vu qu'il fallait aller plus loin. J'ai donc travaillé à mettre au point mon propre outil afin de pouvoir mener des analyses de qualité. L'algorithme de l'outil informatique servant à l'étude de la perte de gènes unitaires a été implémenté au sein de DAGOBAH, le *framework* spécifique du laboratoire. DAGOBAH permet de réutiliser et de partager du code informatique, ainsi que d'intégrer le projet « perte de gènes » dans un large contexte. DAGOBAH permet de formaliser les expertises humaines sous forme de règles (Chapitre II, 1.2.3.1), d'intégrer et d'agréger des agents créés pour répondre à des questions très diverses. C'est l'élément le plus important pour les travaux de recherches du laboratoire. L'outil dédié à l'analyse des pertes de gènes unitaires existe donc sous la forme d'un module de DAGOBAH, appelé ***G**ene **L**oss **A**nalyzer **D**agobah e**X**tension* (GLADX).

Le processus implémenté dans le module GLADX utilise des données présentes dans les bases de données, produit de nouvelles données à partir de ces bases à l'aide de nombreux outils, et assure l'expertise des données initiales aussi bien que celles nouvellement obtenues (Illustration 20). GLADX développé au sein de DAGOBAH contient toutes les règles nécessaires pour déterminer à chaque étape le déroulement à suivre, pour traiter l'ensemble des données nécessaires et pour interpréter le résultat final. Le module GLADX s'appuie sur des outils spécifiques développés par le groupe de bio-informaticiens du laboratoire. Il utilise, par exemple des pipelines de la plateforme FIGENIX, dont celui qui produit des phylogénies et permet également de mettre en évidence les relations d'homologies et d'orthologies de manière automatique. GLADX utilise différentes interfaces :

1) Des interfaces pour communiquer avec les bases de données extérieures comme Ensembl, NCBI et JGI afin de pouvoir les interroger et utiliser au mieux les informations qu'elles contiennent. Ces interfaces permettent d'interroger les bases de

données, afin de connaître les gènes présents à des positions particulières d'un chromosome, d'extraire la séquence d'une position donnée, de connaître les différentes protéines produites par un gène, etc.

2) Une interface pour communiquer avec la base de données ontologique du laboratoire, pour enregistrer et récupérer les données produites.

3) Une interface capable de communiquer avec la plateforme FIGENIX et d'utiliser les pipelines développés au laboratoire qu'elle contient, ainsi que les données produites.

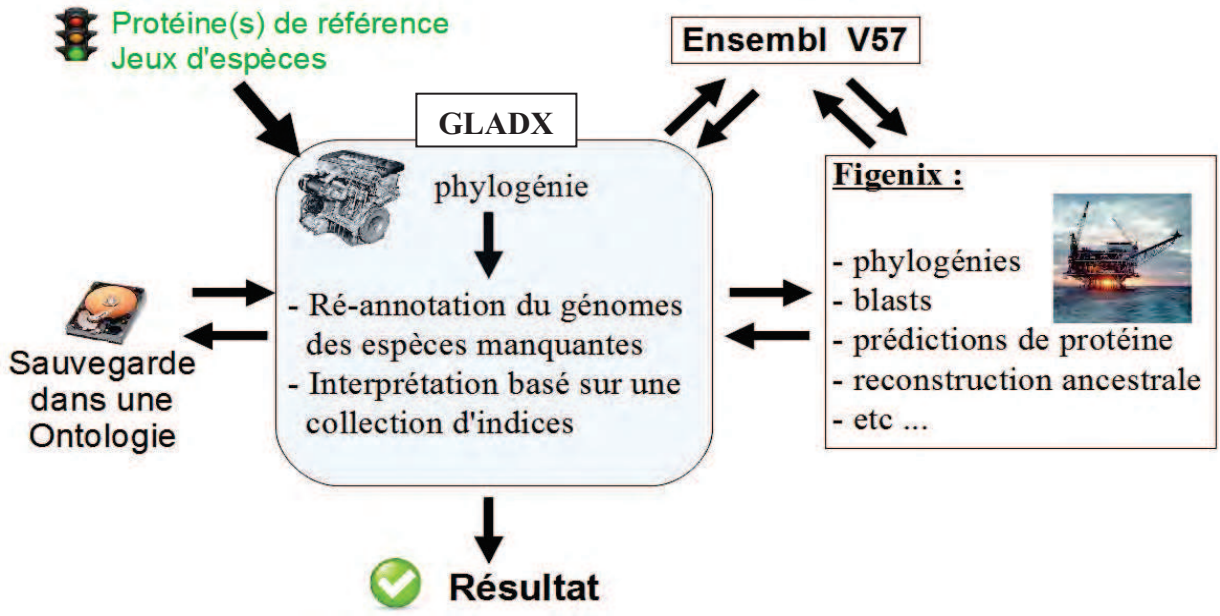


Illustration 20 : Aperçu général du comportement de GLADX développé au sein du framework DAGOBAH

2.1 Axes majeurs du développement de GLADX

GLADX est développé autour d'axes majeurs qui correspondent aux objectifs fixés :

- Utilisation de données à grande échelle et modularité.
- Qualité d'annotation des relations entre les séquences grâce à la phylogénie.
- Capacité à détecter les pseudogènes.
- Analyses au niveau protéique et nucléotidique pour différencier les gènes non annotés des pseudogènes.
- Vision évolutive par reconstruction des états ancestraux et déduction des macro-événements (pertes, pseudogénisations) et micro-événements (mutations génétiques).
- Sauvegarde et description des données dans une ontologie.

Ces différents axes sont développés dans les paragraphes suivants, et éclairent certains points non abordés dans l'article 2.

2.1.1 Étude à grande échelle et modularité

Le déroulement d'une étude par GLADX commence par la prise en main d'un fichier FASTA qui contient une protéine à laquelle on s'intéresse. Ce fichier FASTA doit être déposé dans un répertoire dédié qui est perpétuellement observé par GLADX pour détecter les fichiers FASTA non encore analysés. L'analyse lancée par GLADX porte sur la famille de gènes correspondant à la protéine d'intérêt contenue dans le fichier FASTA. Dans un premier temps GLADX crée une phylogénie qui, ensuite, est utilisée pour rechercher l'absence d'orthologues parmi les espèces utilisées. Ce travail s'effectue selon des paramètres définis qui permettent notamment de choisir les espèces étudiées ainsi que le phylum qui contient la séquence donnée en entrée. L'approche est séquentielle, elle permet d'étudier un fichier FASTA après l'autre, ce qui correspond à une étude gène par gène. Pour répondre à des études plus massives, comme l'étude de différents gènes d'une famille ou des gènes de familles différentes, il suffit de donner à GLADX l'ensemble des fichiers FASTA correspondant aux protéines d'intérêts.

Dans l'étude appliquée aux pertes de gènes unitaires dans la lignée humaine développée dans le Chapitre III de cette thèse, j'ai lancé des études sur plus de 6 000 gènes touchant 26 génomes. L'étude a nécessité quelques mois de calculs. Le temps de calcul de ce type d'étude à grande échelle peut être optimisé en répartissant le travail sur plusieurs machines ou en utilisant des machines performantes.

Les études lancées avec GLADX peuvent utiliser de nombreux génomes selon les choix du biologiste et les génomes disponibles dans les bases de données. Plus le nombre de génomes utilisés est grand, plus les études gagnent en précision pour l'annotation temporelle des événements. La figure ci-dessous (Illustration 21) en montre un exemple. Lors d'une première étude (à gauche) chez *M. musculus*, on détecte la perte d'un gène unitaire. Si cette perte n'est pas associée à un pseudogène, le gène a été perdu à une période comprise entre 0 et 100 millions d'années avant le présent. Dans la deuxième étude (à droite), l'ajout de l'espèce *R. norvegicus* permet de déduire que la perte du gène dans la lignée de *M. musculus* s'est déroulée lors des derniers 40 millions d'années de l'évolution. Cette description triviale montre l'importance du choix des espèces utilisées pour améliorer les connaissances des événements apparus dans des phyla d'intérêts.

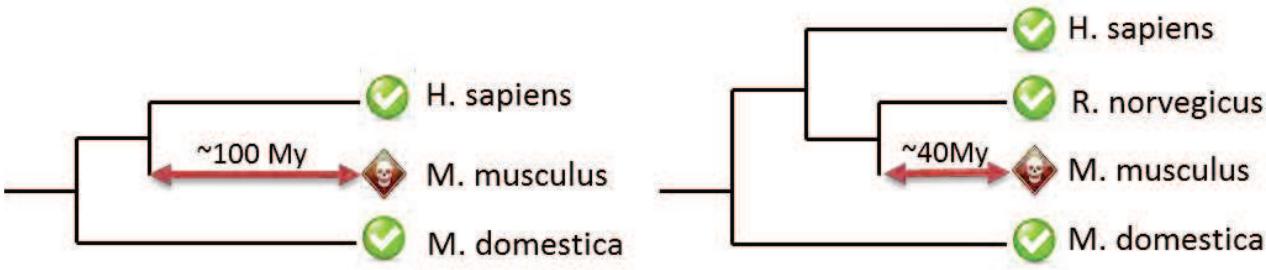


Illustration 21 : Gain de précision sur la datation d'événements grâce à l'ajout d'espèces

2.1.2 La phylogénie, un atout majeur pour la détection des pertes de gènes unitaires

La détection des relations entre les gènes de génomes différents et en particulier la prédiction d'orthologie est un enjeu majeur qu'il est nécessaire d'automatiser. Nous avons vu, dans le Chapitre I (2.2.2), que la grande majorité des stratégies automatisées de détection de pertes de gènes unitaires, utilise des méthodes basées sur la comparaison de séquences. Ces méthodes, pour recenser l'orthologie, s'éloignent du concept même qui la définit. Des analyses récentes montrent que les méthodes s'appuyant sur la phylogénie sont moins sujettes à caution que celles qui utilisent la comparaison de séquences. Les phylogénies permettent d'être au plus près de la définition d'origine, mais l'utilisation de ces méthodes a souffert pendant longtemps des lourdeurs informatiques (complexité des programmes, mémoire nécessaire, temps de calcul). Aujourd'hui, ces méthodes basées sur la phylogénie peuvent être utilisées à grande échelle grâce à l'augmentation de la vitesse des ordinateurs et à l'amélioration des algorithmes, sans oublier le développement d'outils qui permettent une analyse fine des phylogénies obtenues (Gouret *et al.*, 2009). Profitant de l'expérience du laboratoire dans l'automatisation de phylogénies de grande qualité, j'ai utilisé dans le cadre de cette thèse le pipeline de phylogénie présent sur la plateforme FIGENIX développé au sein du laboratoire (Gouret *et al.*, 2005 ; Paganini J., 2012). Ce pipeline de phylogénie qui utilise une protéine de référence en entrée, produit un alignement à partir des séquences les plus similaires obtenues par blast chez les espèces sélectionnées dans les bases de données choisies, et construit une phylogénie de qualité en utilisant de nombreux filtres. Une publication détaillant ce pipeline est disponible (Gouret *et al.*, 2005).

Le processus implémenté dans GLADX utilise la phylogénie à deux niveaux. Une première analyse phylogénétique est faite à partir de la protéine donnée en entrée de GLADX et permet de détecter les pertes putatives de gènes unitaires. Le travail d'annotation des génomes est fastidieux et minutieux. Malgré le travail énorme accumulé, de nombreuses annotations sont

absentes des bases de données. Pour faire face au déficit d'annotations et détecter les éventuels pseudogènes, qui, en revanche, ne sont pas présents dans les bases de données protéiques utilisées, il faut procéder à la recherche de séquences orthologues. Cette recherche permet également de pallier aux séquences éventuellement ratées par la phylogénie initiale. La recherche d'orthologues est effectuée en deux étapes. La première étape recherche des séquences similaires par TBLASN, et la seconde utilise la phylogénie pour définir le lien de parenté des séquences retrouvées. L'utilisation des phylogénies pour définir toutes les relations entre les séquences utilisées permet une analyse fine et apporte énormément d'informations supplémentaires, que ne peut fournir une méthode basée sur la comparaison de séquences.

Pour la recherche d'orthologues, notre approche (blast + phylogénie) est préférée à la synténie pour ne pas rater l'identification des coorthologues apparus par duplication dans les espèces étudiées, et qui ont des positions différentes dans les génomes (Illustration 17). En effet, la définition d'une perte de gènes unitaires ne se définit pas seulement par la perte de l'orthologue à sa position d'origine mais aussi par le fait qu'aucun orthologue fonctionnel ne doit subsister.

2.1.3 Détecter la pseudogénisation lorsque c'est encore possible

On sait que la majorité des pertes de gènes sont dues à un processus de pseudogénisation (Schrider *et al.*, 2009). Le processus de pseudogénisation permet d'observer la séquence durant un certain temps, avant qu'elle ne soit plus reconnaissable au sein d'un génome. En effet, les pseudogènes ont une évolution neutre (W.-H. Li *et al.*, 1981) et accumulent de nombreuses mutations au cours du temps. L'Illustration 22 schématise l'évolution d'un pseudogène, apparu par une première mutation délétère au sein d'un gène fonctionnel, soumis à la dérive génétique. Cette dérive génétique produit au bout d'un certain temps, une séquence dont on ne peut plus reconnaître l'origine.

Nous avons vu dans l'état des connaissances du Chapitre I (2.2.2), qu'aucun outil automatisé ne permet actuellement d'étudier indifféremment les pertes de gènes unitaires sans signal de la séquence d'origine, et celles où des pseudogènes sont présents dans le génome.

L'outil GLADX possède la faculté d'éviter les difficultés rencontrées dans l'utilisation des méthodes existantes et permet de caractériser les pertes de gènes unitaires quand aucun signal de la séquence d'origine n'est présent, aussi bien que celles où un pseudogène existe. GLADX effectue une recherche systématique des pseudogènes au sein des génomes étudiés et

permet d'observer les mutations qui caractérisent les pseudogénisations. GLADX peut ainsi différencier parmi les pertes récentes celles qui seraient le fruit d'une délétion. De plus, lorsque les pseudogènes de plusieurs espèces d'une même lignée sont observés, il est possible de déterminer s'il existe une mutation commune à l'origine de la pseudogénisation du gène dans cette lignée. (A noter que la recherche de pseudogènes peut être désactivée)

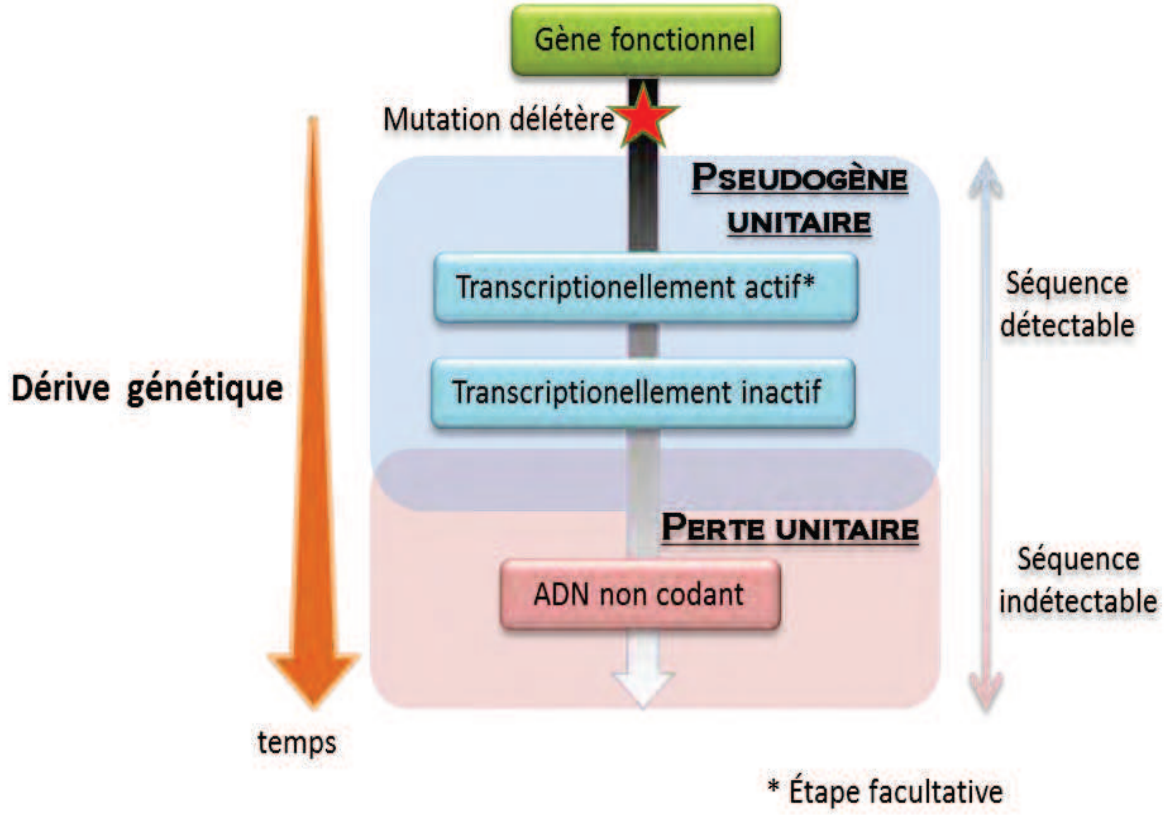


Illustration 22 : Différents stades de la perte d'un gène fonctionnel par pseudogénisation

2.1.4 Annotation des séquences pour détecter les pseudogènes et les gènes intacts

La première étape de GLADX consiste à détecter les pertes de gènes unitaires à travers la lecture d'une phylogénie construite. La deuxième étape consiste à vérifier, pour toutes les pertes détectées, qu'aucun signal orthologue n'est présent dans le génome de l'espèce concernée. Pour rechercher les séquences orthologues manquantes, une recherche de séquences similaires est faite par TBLASTN. Les séquences retrouvées sont ensuite traitées séquentiellement par le pipeline de phylogénie. Les orthologues d'une séquence de TBLASTN présente dans une phylogénie sont alors comparés avec le groupe d'orthologues présents détecté dans la phylogénie au départ d'une étude GLADX. Lorsque des orthologues communs sont trouvés, on en déduit que la séquence de TBLASTN est orthologue. Une étape

d'annotation est ensuite nécessaire. Cette séquence correspond soit à un gène déjà annoté et raté par les premières étapes de GLADX, soit à un gène qui n'a pas encore été annoté dans la base de données utilisée, soit à un pseudogène. Pour connaitre leur statut, ces séquences sont analysées en premier lieu au niveau protéique et en deuxième lieu, lorsque le signal de la séquence semble assez bon, au niveau nucléotidique.

2.1.4.1 **Analyse au niveau protéique**

L'analyse au niveau protéique consiste d'abord à prédire grâce à un pipeline dédié, et au sein d'une séquence nucléique comportant une séquence de TBLASTN identifiée comme orthologue, la protéine la plus proche d'une des protéines du groupe d'orthologues de départ. Quand toutes les prédictions protéiques dans les séquences trouvées orthologues sont terminées, l'étape suivante consiste à vérifier que les protéines prédites n'ont pas tout simplement été ratées par la phylogénie de départ. Pour ce faire, et pour chaque protéine prédite, on compare la séquence avec les annotations présentes dans la base de données utilisée. On relève la position de la prédiction sur le chromosome de l'espèce concerné, ainsi que son sens, et on teste sa similarité avec les protéines qui sont présentes à la même position et dans le même sens dans le génome étudié. Dans la majorité des cas, aucune protéine équivalente voire aucune protéine n'est trouvée annotée. Ensuite, les prédictions correspondant à des séquences non annotées passent par une étape qui permet de contrôler la qualité des séquences nucléotidiques sous-jacentes : la reconstruction de séquences ancestrales réclame, pour réussir, des séquences nucléiques peu divergentes. Ainsi, grâce à la séquence protéique trouvée orthologue par TBLASTN et de la protéine que l'on vient de prédire, GLADX analyse la qualité de la séquence nucléotidique sous-jacente, et détermine si une étude au niveau nucléotidique est possible. Lorsque l'étude reste au niveau protéique, GLADX effectue une comparaison avec les autres protéines orthologues connues, en fonction de leur affiliation phylogénétique et de leur temps de divergence. Ces séquences sont alors reconnues comme des pseudogènes ou gènes putatifs. Lorsqu'à l'issue de ce processus GLADX conclut à un pseudogène, il s'agit probablement d'une pseudogénisation avancée (Illustration 23).

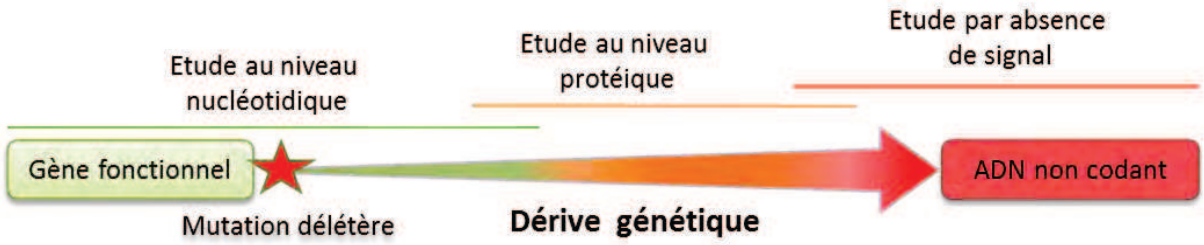


Illustration 23 : Niveau d'observation utilisé par GLADX en fonction de l'avancée de la pseudogénisation
Lorsque la pseudogénisation est récente l'étude se fait au niveau nucléotidique, alors que lorsque la pseudogénisation est avancée elle se fait au niveau protéique, et si elle est fortement avancée aucune séquence ne peut être étudiée.

2.1.4.2 Analyse au niveau nucléotidique

Lorsque le signal de séquences protéiques indique que peu de changements sont apparus au niveau nucléotidique, GLADX effectue une étude des séquences nucléotidiques. Dans le cadre d'un pseudogène cela signifie que le processus de pseudogénisation est encore récent. L'analyse au niveau nucléotidique permet d'observer plus finement les modifications génétiques qui se sont produites. Elle permet également de différencier, parmi les séquences orthologues retrouvées, celles qui correspondent à des gènes fonctionnels et celles qui correspondent à des pseudogènes. L'utilisation d'une étape de reconstruction de séquences ancestrales est implémentée pour permettre un gain de précision et d'information dans l'observation des mutations génétiques.

2.1.4.2.1 La reconstruction de séquences ancestrales

Traditionnellement, les études sur les pseudogènes unitaires effectuent toujours une analyse au niveau nucléotidique pour observer les mutations. Pour ces études, on a recours à un alignement de séquences contemporaines contenant au minimum un gène connu et une séquence d'intérêt potentiellement pseudogène. Pour observer les mutations sur la séquence d'intérêt, les séquences exoniques du gène connu sont comparées avec les séquences correspondantes de la séquence d'intérêt. Le premier problème soulevé avec cette approche est que la comparaison de deux séquences contemporaines ne reflète pas forcément la réalité des événements. En effet, depuis un ancêtre commun, ces deux séquences ont toutes deux divergé de manière indépendante. A une même position, il peut y avoir eu une mutation dans chacune des séquences observées. Le deuxième problème se situe au niveau des insertions et délétions qui ne peuvent pas être différenciées dans cette approche. Car entre deux séquences contemporaines il est impossible de savoir si un gap correspond à une délétion dans une des

séquences ou à une insertion dans l'autre. Dans notre approche, j'ai automatisé l'utilisation de l'outil **Ortheus** permettant, à partir d'un jeu de séquences, de reconstruire les séquences ancestrales correspondantes (Paten *et al.*, 2008). GLADX observe alors les mutations apparues entre une séquence ancestrale et une autre séquence ancestrale, ou une séquence contemporaine avec sa séquence ancestrale. Cela permet de répondre aux deux problèmes soulevés : être le plus proche possible de la réalité des événements observés (un état ancestral par rapport à un état contemporain), et en conséquence pouvoir, par exemple, différencier les insertions des délétions, ou définir les types de substitutions. Cette approche permet donc d'observer les mutations apparues depuis les différents ancêtres reconstruits jusqu'à la séquence contemporaine. Elle permet une vision évolutive du processus génétique caractérisant chaque pseudogénisation. Il est donc possible de mettre en évidence la première mutation ayant entrainée la pseudogénisation.

2.1.4.2.2 **Le scanner**

Le scanner est un agent du module GLADX dédié à l'analyse des mutations apparues entre une séquence et sa séquence ancestrale. GLADX recherche toutes les mutations pouvant avoir un impact pseudogénisant, à savoir les mutations touchant le codon initiateur (apparition, disparition), le codon final (apparition, disparition), les insertions et les délétions, les modifications des sites d'épissage, l'apparition ou la disparition de codons stops dans le cadre de lecture, et enfin l'apparition et la disparition d'exon(s) (Illustration 24). L'ensemble des scans qui comparent les séquences contemporaines et leurs séquences ancestrales et les différentes séquences ancestrales entre elles, permet d'observer l'apparition des mutations au cours du temps qui caractérisent une pseudogénisation.

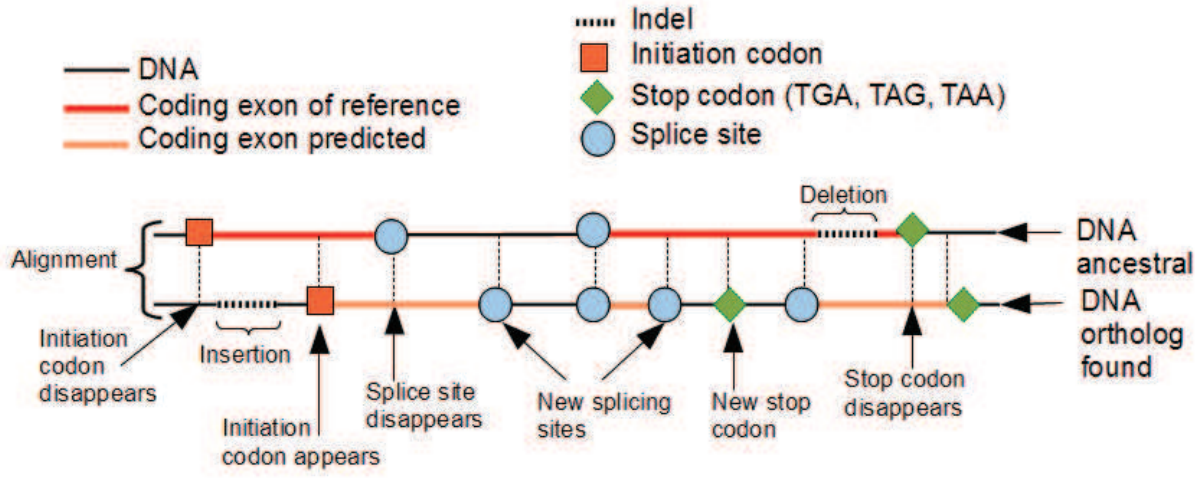


Illustration 24 : Ensemble des mutations observées par scan entre une séquence contemporaine et une séquence ancestrale

2.1.5 **Une vision évolutive**

Lorsque toutes les espèces incluses dans l'étude ont été analysées, GLADX effectue une synthèse des résultats obtenus. Pour avoir une meilleure vision des résultats, ceux-ci sont affichés sur un arbre des espèces. Les mutations trouvées au niveau nucléotidique sont directement affichées sur les branches de l'arbre qui les concerne. Ces informations ont déjà une dimension évolutive et permettent de voir temporellement et étape par étape les mutations apparues au cours de l'évolution. La reconstruction de séquences ancestrales permet cela. Tous les caractères (présent, perdue, pseudogène) qui définissent l'état du gène chez chaque espèce étudiée sont également affichés sur l'arbre des espèces.

A partir des caractères définis sur les feuilles de l'arbre des espèces il est possible de déduire la période relative d'apparition des événements (apparition du gène, perte, pseudogénisation) à l'origine de ces caractères. Pour cela, j'ai implémenté au sein de GLADX l'algorithme de parcimonie de Sankoff (Sankoff, 1975 ; Sankoff & Rousseau, 1975) qui permet de reconstruire les états ancestraux. J'ai préféré l'algorithme de Sankoff à celui de Mirkin (Mirkin *et al.*, 2003) parce que l'algorithme de Mirkin peut donner plusieurs apparitions d'un gène : j'ai posé le postulat que l'apparition d'un gène ne peut avoir lieu qu'une seule fois. En choisissant l'algorithme de Sankoff, je ne prends pas en compte les transferts horizontaux de gènes (THGs). En effet, comme ces THGs sont peu représentés au sein des génomes des métazoaires, j'ai fait l'hypothèse que leur impact sur l'étude de perte de gènes unitaires resterait marginal. Mais les transferts horizontaux de gènes ne sont pas pour autant oubliés, car ils représentent des événements particuliers et très riches d'enseignements pour appréhender l'évolution. Un module dédié à l'étude des THGs existe au sein de DAGOBAH. Il peut être utilisé pour signaler l'existence d'un tel événement afin d'en tenir compte si nécessaire. J'ai également préféré l'algorithme de parcimonie de Sankoff à celui de Dollo (Farris, 1977 ; Lequesne, 1972) parce qu'il offre la possibilité de modifier le poids des événements. Ainsi j'ai mis un poids qui permet une seule apparition d'un gène, et rend impossible une pseudogénisation après une perte totale du gène dans une lignée. La définition des caractères ancestraux grâce à la parcimonie de Sankoff permet de déduire dans quelles branches le gène est apparu, et dans quelles branches se sont produits des événements de pseudogénisation et de pertes unitaires de gènes. Ces événements sont également affichés sur l'arbre d'espèces qui représente la synthèse des résultats obtenus.

2.1.6 Sauvegarde des données dans une base de données ontologique

L'ensemble du travail effectué et des données produites par GLADX est enregistré dans une ontologie. Cette ontologie est commune à tous les modules implémentés dans DAGOBAH et représente l'ensemble des informations produites au laboratoire (analyse de changements d'architecture en domaines de protéines, études de transfert horizontaux de gènes, etc.). Les données ontologiques permettent de décrire les événements observés, ainsi que de factoriser les classes décrites et les données. Ainsi, les données produites lors d'une étude (phylogénies, mutations, etc.) peuvent éventuellement être utilisées par d'autres modules.

Partant de l'ontologie déjà présente, j'ai développé la partie ontologique spécifique à l'étude de perte de gènes unitaires, qui est nécessaire à la description des données et à leur mise en relation (Annexe 2, Annexe 3, Annexe 5).

2.2 Effet des subtilités du concept d'orthologie sur l'étude des pertes de gènes unitaires

Les pertes de gènes unitaires sont définies par le fait qu'aucun gène orthologue n'est présent dans une espèce donnée. Définir les liens d'orthologie et, par conséquent, de paralogie, est un élément central dans l'identification des pertes de gènes unitaires. Les orthologues sont détectés par la lecture des relations de spéciation et duplication d'un gène de référence par rapport aux autres gènes (W. Fitch, 1970). C'est à partir de l'analyse d'arbres phylogénétiques où les événements de duplications et de spéciations des gènes sont annotés que l'on peut connaître la présence ou l'absence d'orthologues à un gène de référence donné et ainsi savoir s'il y a eu pertes de gènes unitaires.

Dans le développement de GLADX j'ai fait le choix de concentrer chaque étude sur la famille spécifique des gènes pris en référence. Ces références correspondent aux séquences données en entrées dans les fichiers FASTA. Le choix de l'espèce dont le gène est utilisé comme référence a un impact sur les orthologues détectés et par conséquent sur les pertes de gènes. Pour illustrer l'importance du choix de l'espèce et du gène de référence, prenons l'exemple décrit par l'Illustration 25. On constate que quand le gène B de *G. gallus* (GB) est pris comme référence, les orthologues et les pertes détectés ne sont pas les mêmes que quand on se réfère au gène de *X. laevis* (X). Dans le premier cas (référence GB) on trouve l'orthologue X et la perte du gène chez *M. musculus* et *H. sapiens*, dans le second (référence X) on trouve les orthologues GA, GB, HA et une perte du gène chez *M. musculus*.

Dans l'étude de pertes de gènes unitaires, la lecture « traditionnelle » d'un arbre phylogénétique qui consiste à rechercher des orthologues à une séquence de référence peut entraîner des erreurs. Ce constat vient du fait que **l'orthologie n'est pas un concept transitif**. En effet, pour vérifier l'absence d'un orthologue dans le génome d'une espèce, GLADX recherche, par une étape de TBLASTN, des séquences similaires qu'il teste ensuite par une nouvelle phylogénie pour vérifier leur orthologie. Une séquence de TBLASTN est vue comme orthologue, donc non perdue, si elle est orthologue à au moins un des gènes orthologues pris au départ comme référence.

Dans l'illustration 25, lorsque le gène X est pris comme référence, on détecte la perte chez *M. musculus* seulement, et la recherche de séquences orthologues par TBLASTN dans le génome de *M. musculus* ne donnera aucun résultat. Ce gène manquant qui est détecté correspond bien à une perte de gène unitaire.

Dans le cas où GB est pris comme référence, les orthologues présents, définis au départ, sont GB et X. Ainsi l'homme et la souris semblent avoir perdu le gène. Lorsque GLADX recherche par TBLASTN un signal similaire au gène X dans le génome de l'homme, il risque de trouver le gène HA. Dans la phylogénie qui suit, faite à partir de la séquence correspondant à HA, HA sera trouvé orthologue au gène X. Ainsi, en comparant le groupe d'orthologues de départ (GB et X) et le groupe d'orthologues de la séquence de TBLASTN (HA et X), GLADX conclut que HA est orthologue à GB, et donc que le gène n'est pas perdu chez l'homme. Cette conclusion peut s'apparenter à un artefact car le gène HA est en fait paralogue à GB. Ceci provient du fait que l'on omet la duplication apparue après la spéciation des Tétrapodes (*Tetrapoda*). En résumé, dans ce cas, la perte chez *M. musculus* est détectée mais celle chez *H. sapiens* est manquée.

En conséquence, cette approche risque de manquer des pertes de gènes unitaires ou des pseudogènes unitaires. L'omission de la duplication dans l'ancêtre des Tétrapodes entraîne d'autres questionnements dans la définition des caractéristiques de la famille du gène étudié. Ainsi, en utilisant GB comme référence, la perte de l'orthologue chez *M. musculus* est détectée mais on pense que le gène perdu est établi depuis l'ancêtre des Tétrapodes puisqu'il existe un orthologue chez *X. laevis*. En fait, le gène perdu chez *M. musculus* est issu d'une duplication apparue chez les Amniotes. Avec une subfonctionnalisation du duplicata B, cela entraîne l'émergence d'une nouvelle famille ou sous-famille. L'origine du gène remonte au moins à l'ancêtre des Tétrapodes mais la famille étudiée est spécifique aux Amniotes. Donc si l'on souhaite faire de l'annotation fine pour voir précisément le type de fonction perdue chez

M. musculus, il faut utiliser uniquement le gène GB qui est plus proche de la fonction perdue. Si la fonction du gène GB n'existe pas, le gène X peut être utilisé, mais il faut savoir que cette fonction peut avoir grandement varié. Connaître l'apparition des gènes est important pour bien comprendre la dynamique des gènes au sein des génomes, et pour comprendre depuis quand le gène est établi. En effet, l'idée que l'importance des gènes est en relation avec leur temps de fixation est un des aspects mis en avant lors de l'étude des pertes de gènes unitaires.

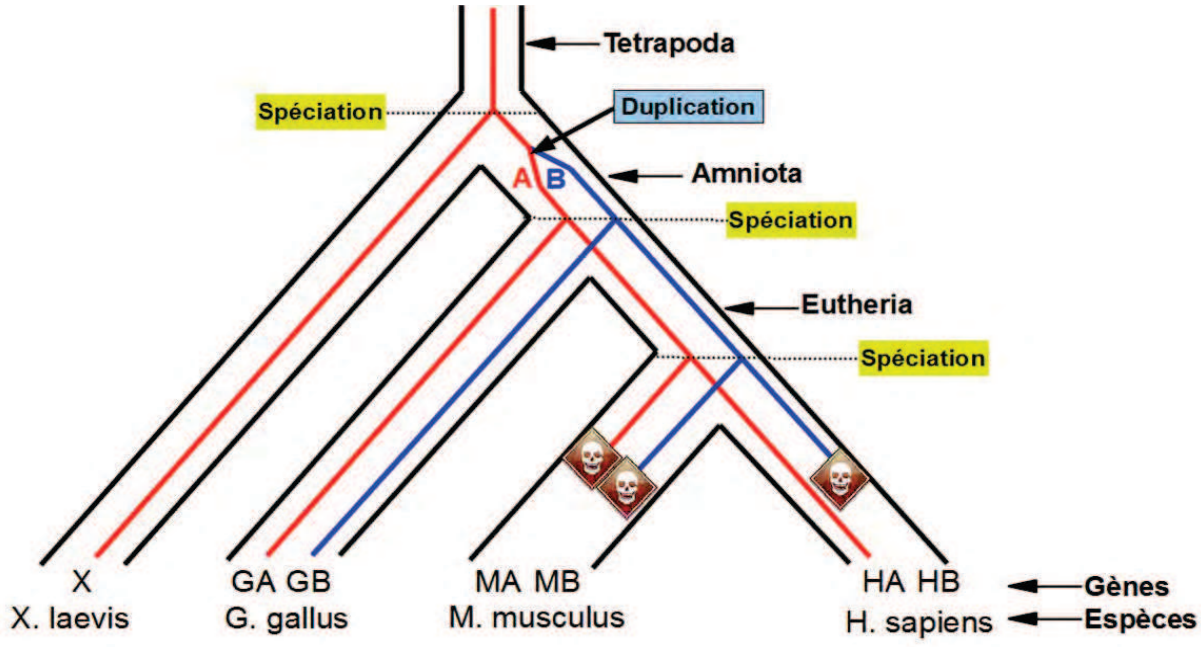


Illustration 25 : Exemple théorique de pertes de gènes unitaires dans la lignée des Tétrapodes
Dans cet exemple, le gène a subi une duplication chez l'ancêtre des Amniotes, le duplicata B du gène ancestral se retrouve néofonctionnalisé, acquérant ainsi une nouvelle fonction.

Une autre procédure était donc nécessaire pour gagner en précision et ne pas répéter les erreurs causées par les approches citées précédemment. Pour cela j'ai emprunté la vision utilisée par les bactériologistes pour étudier les orthologues. Ces spécialistes, considèrent que les orthologues sont des gènes originaires d'un gène ancestral dans le dernier ancêtre commun des espèces étudiées (Koonin, 2005). Pour parler d'orthologie il faut utiliser uniquement des gènes contemporains, c'est-à-dire sur les feuilles d'un arbre phylogénétique. C'est pour cela que les gènes issus d'un gène ancestral forment ce qu'on appelle un **groupe d'orthologues (GO)**. Dans un GO donné, tous les gènes sont orthologues aux gènes de l'espèce la plus extérieure, qui est liée aux autres espèces par un événement de spéciation ou un nœud de spéciation dans un arbre phylogénétique. Dans ce type d'approche, le paramètre principal d'une étude est l'ancêtre choisi. Il permet de définir le phylum étudié.

Selon l'approche classique expliquée précédemment, GLADX définit les orthologues par

rapport à un gène de référence qui est une feuille d'un arbre phylogénétique. Dans la nouvelle approche, GLADX détermine un groupe d'orthologues à partir d'un nœud de spéciation ancestral. L'ensemble des gènes du phylum ainsi défini forme un GO. L'outil GLADX étudie la famille contenant le gène de référence donné en entrée, le phylum étudié est donc celui qui contient le gène de référence. Ainsi, les espèces issues de l'ancêtre défini en entrée, et qui ne sont pas représentées dans un arbre, sont des espèces ayant, suppose-t-on, perdu le gène qui était présent chez leurs ancêtres. Ces pertes sont dites « lignée-spécifiques ». En d'autres termes, si l'ancêtre référent est celui des vertébrés, les pertes détectées dans les espèces du phylum des vertébrés seront des pertes « vertébrés-spécifiques ». Cela signifie que le gène perdu chez ces espèces de vertébrés était toutefois présent chez l'ancêtre des vertébrés. GLADX est donc programmé pour étudier des pertes « lignée-spécifiques ». Pour étudier toutes les pertes lignée-spécifiques dans la lignée menant au gène de référence, GLADX a été doté d'un agent particulier. Cet agent est doté d'une expertise qui lui permet de manière itérative d'avancer ancêtre par ancêtre pour étudier les pertes de gènes unitaires dans des phyla de plus en plus restreints.

Etudions maintenant le comportement de GLADX avec cette nouvelle approche dans l'exemple ci-dessus (Illustration 25) où le gène GB est donné en entrée et l'étude initiée à partir de l'ancêtre des vertébrés. Puisque l'ancêtre des vertébrés n'est pas présent, l'étude commence à l'ancêtre des Tétrapodes. Dans ce cas, les descendants de l'ancêtre du gène GB chez les Tétrapodes se retrouvent chez toutes les espèces sauf chez *M. musculus*. Une recherche du représentant du gène chez *M. musculus* par GLADX ne trouve aucun gène orthologue fonctionnel dans le génome de cette espèce. La perte de gène unitaire détectée chez *M. musculus* est ainsi vérifiée. Le gène perdu correspond à un gène fonctionnel qui existait au moins depuis l'ancêtre des Tétrapodes. Toutes les pertes « tétrapodes-spécifique » ayant été étudiées, l'outil poursuit alors sa recherche en se plaçant au niveau de l'ancêtre des Amniotes. Dans ce cas, l'orthologue du gène GB est absent chez *H. sapiens* et *M. musculus*. Comme la perte unitaire a déjà été étudiée chez *M. musculus*, GLADX vérifie uniquement la perte chez *H. sapiens*. Le résultat final obtenu est une perte de gène unitaire chez *M. musculus* et *H. sapiens* (MB et HB). La perte chez *M. musculus* est « Tétrapodes-spécifique » tandis que celle chez *H. sapiens* est « Amniotes-spécifique ». Dans ce cas nous avons l'information que le gène B est apparu par duplication et qu'il est également absent de *M. musculus*.

L'approche « traditionnelle » induit des artefacts lorsqu'un orthologue Y d'un gène orthologue au gène de référence est récupéré, et qu'en réalité cet orthologue Y est paralogue au gène de

référence. La nouvelle approche permet d'éviter ces artefacts d'orthologies.

GLADX utilise les informations récoltées par l'une ou l'autre des deux méthodes présentées ci-dessus pour en déduire la date d'apparition de la famille du gène étudié ainsi que les dates des pertes. De ces deux dates peut être déduit le temps de fixation des gènes au sein des génomes, au cours de l'évolution. L'apparition d'un gène se fait par lecture de la phylogénie et la date de sa perte se fait en utilisant un algorithme de parcimonie. Ainsi, dans l'exemple précédent (Illustration 25), la perte du gène MA est vue comme spécifique à *M. musculus* tandis que la perte du duplicata B chez *M. musculus* (MB) et *H. sapiens* (HB) est vue comme une perte antérieure au dernier ancêtre commun des Euthériens (*Eutheria*). Aucun moyen ne permet de savoir s'il s'agit de deux pertes indépendantes, sauf à rajouter des espèces qui pourraient nous renseigner. Par exemple, si on ajoute à l'étude l'espèce *R. norvegicus,* phylogénétiquement plus proche de *M. musculus* que les autres espèces, et qui possède le gène B, on peut en déduire que les pertes chez *M. musculus* et *H. sapiens* sont indépendantes. Les études effectuées avec GLADX utilisent par défaut la deuxième approche (dites mode lignée) qui permet de ne pas passer à côté de pertes de gènes unitaires et d'avoir des renseignements plus fins sur les caractéristiques de la famille de gènes étudiée (apparition, fixation, etc).

2.3 Limites de la méthode implémentée dans GLADX

2.3.1 La phylogénie de départ

Création de la phylogénie : Les études réalisées par GLADX reposent sur des analyses phylogénétiques. Les étapes de création des phylogénies sont donc cruciales. En effet, la qualité des séquences, un signal non congruent, ou le nombre de séquences peut empêcher le pipeline de réaliser une phylogénie correcte. Il arrive qu'on ne puisse commencer une étude à cause de l'impossibilité de réaliser la phylogénie de départ.

Réarrangement de la phylogénie : Dans GLADX, le pipeline de phylogénie utilise, lors d'une des étapes finales, l'algorithme NOTUNG (K. Chen, Durand, & Farach-Colton, 2000). A partir de l'arbre des espèces fourni, l'algorithme réarrange la phylogénie de manière cohérente et parcimonieuse, en minimisant le nombre de duplications. Il replace ainsi les gènes mal placés en touchant aux branches qui ont des nœuds peu soutenus. De cette manière, NOTUNG permet de limiter le nombre des faux positifs de pertes de gènes unitaires, dus au mauvais placement des gènes dans la phylogénie. L'impact de cet algorithme n'a pas été

évalué. Il pourrait créer des changements dans la phylogénie, ce qui a pour conséquence la présence de gènes orthologues faux positifs. Des pertes pourraient ainsi ne pas être prises en compte.

Lorsque l'orthologue détecté est absent à la lecture de la première phylogénie, il faut ré-annoter des séquences pour pouvoir récupérer des orthologues laissés de côté par la première phylogénie, annoter de nouveaux gènes, détecter des pseudogènes, ou confirmer l'absence de séquences orthologues. Deux étapes importantes sont alors utilisées. La première étape se base sur la recherche de séquences similaires par TBLASTN et la seconde concerne la construction de phylogénies pour connaître les liens de parentés des séquences retrouvées.

2.3.2 La recherche d'orthologues par TBLASTN

Lors de la recherche d'un signal par TBLASTN dans un génome où une perte a été au premier abord supposée, la valeur arbitraire fixée par défaut à 75 000 nucléotides, définit la distance maximum entre deux « *High Scoring Pair* » (HSP) liés. Les HSP sont censés correspondre à du signal exonique. Les HSP regroupés forment des séquences (hits) de BLAST. Ainsi, on considère que deux HSP distants de plus de 75 000 nucléotides n'appartiennent pas au même hit. Cela signifie que deux exons séparés par un intron d'une taille allant jusqu'à 75 000 nucléotides, se retrouvent dans les HSP d'un même hit. Cette valeur de 75 000 nucléotides a été choisie afin de couvrir la grande majorité des introns existants lors de la recherche de hits. En effet, la taille moyenne des introns chez *A. thaliana*, *D. melanogaster*, *M. musculus* et *H. sapiens* dans les régions codantes est respectivement de 158, 818, 2874 et 3479 bases. La médiane de la taille des introns est quant à elle de 98, 68, 1095 et 1334 bases (Hong, Scofield, & Lynch, 2006). Des gènes composés d'introns supérieurs à 75 000 bases semblent extrêmement rares. La valeur choisie par défaut n'est pas supérieure afin de limiter la création de hits « chimères » qui seraient composés de HSP provenant de paralogues proches. Si un gène comporte un intron d'une taille supérieure, le gène sera découpé en différents hits de BLAST. Une partie du signal manque alors, pour la suite de l'analyse.

2.3.3 Les phylogénies de hits de TBLASTN

Nous avons vu précédemment que de nombreux éléments peuvent empêcher le bon déroulement d'une phylogénie. A cette étape, c'est souvent la qualité du signal de la séquence récupérée par blast qui est problématique. En effet il n'est pas rare de récupérer des séquences trop courtes pour effectuer une phylogénie.

Lorsqu'une phylogénie de hit de TBLASTN fonctionne, il est également possible que

l'utilisation de NOTUNG entraîne une sur-interprétation des relations d'orthologies.

2.3.4 Les prédictions protéiques

Lorsqu'une séquence nucléotidique orthologue est retrouvée, une prédiction de protéines est effectuée avec le pipeline nommé *SlidingGenePredixForSpecificTarget*. Ce pipeline accepte des séquences sans limite de taille, mais la combinatoire augmente de manière exponentielle avec la longueur de la séquence. Pour optimiser le temps de calcul du pipeline, les prédictions se font dans des fenêtres glissantes définies à 25 000 nucléotides. Les morceaux de prédictions faites sur des fenêtres successives sont raccrochés entre eux ; lorsque l'espace est supérieur à la taille d'une fenêtre, ils ne peuvent pas être réunis au sein d'une même prédiction. La valeur assignée de 25 000 nucléotides est largement supérieure à la taille moyenne des introns présents dans les régions codantes (Hong *et al.*, 2006), ce qui permet de couvrir la majorité des cas. Mais quand un intron a une taille supérieure à 25 000 nucléotides, le début ou la fin d'un long gène peut manquer dans la prédiction, ce qui entraîne le raccourcissement de la protéine prédite. La réduction de la protéine n'engendre pas de problème particulier dans le cas d'une étude au niveau protéique. Quelle que soit la taille de la prédiction, quand le score de similarité avec la partie correspondante de la protéine orthologue est congruent avec le temps de divergence des séquences, alors le gène est défini comme intact par GLADX. Dans le cas contraire GLADX conclut que la séquence est un pseudogène et qu'elle ne permet pas de prédire une protéine correcte. Une protéine trop courte correspond, au niveau nucléotidique, au manque d'un ou plusieurs exons. GLADX est conçu pour que ce cas de figure n'engendre pas de problème au niveau nucléotidique. Le manque d'exons d'un gène n'entraîne pas la détermination d'un pseudogène par GLADX. Toutefois, dans le cas où les exons manquants sont ceux du début de la séquence, leur absence peut entraîner un décalage de phase lors de la recherche de mutations dans le reste de la séquence étudiée. Un décalage de phase fait apparaître de nombreux codons stops, et induit ainsi la détection d'un pseudogène au lieu d'un gène intact. Pour éviter ces décalages de phase, le scanner de GLADX chargé de la recherche des mutations commence sa recherche au premier exon trouvé sur la séquence étudiée ; les exons définis sur la séquence étudiée proviennent de l'inférence des exons d'un gène connu. Le scanner élimine si nécessaire les premiers nucléotides pour commencer dans la même phase que celle observée sur le gène de référence.

2.3.5 La recherche de mutations par le scanner

La recherche de mutations au niveau nucléotidique par le scanner est sensible. C'est une étape importante, car elle assure l'annotation d'un gène intact lorsqu'aucun codon stop n'est observé, et l'annotation d'un pseudogène, s'il en existe. Des codons stop erronés peuvent être dus à l'outil de reconstruction de séquences ancestrales, à des erreurs de séquençages des séquences utilisées (un nucléotide en plus ou en moins), ou à des erreurs dans l'alignement des séquences scannées. Ce dernier point est approfondi dans l'article 2.

Les problèmes liés à la taille des introns peuvent être évités en identifiant en amont du lancement de GLADX, la longueur des gènes ; on évite ainsi l'étude de gènes dépassant un seuil fixé. Pour analyser les « longs » gènes, il est possible de modifier les paramètres de GLADX ou utiliser des méthodes manuelles. La modification des paramètres de GLADX doit être temporaire, car les valeurs par défaut optimisent les temps de calculs.

2.4 Résultats du développement de GLADX

Les résultats du développement de GLADX ont fait l'objet d'un article (Article 2). Dans une première partie, il intègre les différents axes énoncés ci-dessus, en approfondissant étape par étape le déroulement et le comportement des processus programmés dans GLADX. Il donne des informations sur les paramètres permettant de contrôler le comportement de GLADX par l'utilisateur. La manipulation de ces paramètres est décrite dans les annexes de l'article 2. Dans une deuxième partie, cette publication montre la capacité de GLADX à répondre clairement et avec précision à des événements de pseudogénisation. Pour s'assurer de l'efficacité de GLADX j'ai testé des cas de pseudogénisation bien décrits dans la littérature. Ces tests retrouvent de façon automatisée les mêmes résultats que ceux précédemment décrits et obtenus manuellement, et, dans la majorité des cas, GLADX parvient à gagner en précision et découvre de nouveaux événements. Le nombre important d'espèces que l'on peut étudier simultanément avec GLADX contribue grandement à apporter de nouveaux résultats. GLADX est un outil innovant. Malgré les difficultés engendrées par l'automatisation de l'expertise et l'utilisation d'outils informatiques variés, il répond pleinement aux objectifs fixés. GLADX illustre la possibilité de développer au sein du *framework* de DAGOBAH des outils efficaces pour répondre à des questions complexes par l'automatisation informatique.

Article 2 - GLADX: An automated approach to analyze the lineage-specific orthologous gene loss and pseudogenisation in Metazoans

(*PLoS ONE*, 2012)

OPEN ACCESS Freely available online

PLoS one

GLADX: An Automated Approach to Analyze the Lineage-Specific Loss and Pseudogenization of Genes

Jacques Dainat*, Julien Paganini, Pierre Pontarotti, Philippe Gouret

Aix-Marseille Université Laboratoire d'Analyse, Topologogie, Probabilités (LATP) UMR–CNRS 7353 équipe Evolution Biologique & Modélisation, Marseille, France

Abstract

A well-established ancestral gene can usually be found, in one or multiple copies, in different descendant species. Sometimes during the course of evolution, all the representatives of a well-established ancestral gene disappear in specific lineages; such gene losses may occur in the genome by deletion of a DNA fragment or by pseudogenization. The loss of an entire gene family in a given lineage may reflect an important phenomenon, and could be due either to adaptation, or to a relaxation of selection that leads to neutral evolution. Therefore, the lineage-specific gene loss analyses are important to improve the understanding of the evolutionary history of genes and genomes. In order to perform this kind of study from the increasing number of complete genome sequences available, we developed a unique new software module called GLADX in the DAGOBAH framework, based on a comparative genomic approach. The software is able to automatically detect, for all the species of a phylum, the presence/absence of a representative of a well-established ancestral gene, and by systematic steps of re-annotation, confirm losses, detect and analyze pseudogenes and find novel genes. The approach is based on the use of highly reliable gene phylogenies, of protein predictions and on the analysis of genomic mutations. All the evidence associated to evolutionary approach provides accurate information for building an overall view of the evolution of a given gene in a selected phylum. The reliability of GLADX has been successfully tested on a benchmark analysis of 14 reported cases. It is the first tool that is able to fully automatically study the lineage-specific losses and pseudogenizations. GLADX is available at http://ioda.univ-provence.fr/lodaSite/gladx/.

Citation: Dainat J, Paganini J, Pontarotti P, Gouret P (2012) GLADX: An Automated Approach to Analyze the Lineage-Specific Loss and Pseudogenization of Genes. PLoS ONE 7(6): e38792. doi:10.1371/journal.pone.0038792

Editor: Sergios-Orestis Kolokotronis, Barnard College, Columbia University, United States of America

Received March 26, 2012; **Accepted** May 10, 2012; **Published** June 18, 2012

Funding: JD was funded by a PhD fellowship from the French Ministry of Research. This work was supported by the French National Research Agency [EvolHHupro: ANR-07-BLAN-0054]. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: jacques.dainat@gmail.com

Introduction

Essential genes, such as housekeeping genes or genes involved in interaction networks, remain stable during evolution due to their central biological role and tend to evolve under purifying selection [1–3]. The more the gene is important, the more it tends to be universally conserved. Unlike the gene losses due to functional redundancy after gene duplication [4], the lineage-specific losses of well-established genes may reflect significant changes [5–7]. Two mechanisms might describe the losses of well-established genes: i) losses that are not linked to environmental shifts, but to the presence of other genes in the genome that can fulfill the original functions, and ii) losses linked to environmental shifts, and that can be either produced by genetic drift with no selection (i.e. they encode functions that are no longer useful), or by adaptive negative selection (i.e. the maintenance of functions that generate handicaps). The counterintuitive concept that gene losses may be an important driver of evolutionary change via adaptive changes was named the "less is more" hypothesis [8]. The lineage-specific gene losses of well-established genes can be due to deletion events or to pseudogenizations. This kind of pseudogenes is called unitary pseudogenes [9]. After a certain time it is not possible to differentiate between the two cases. Indeed, once that a gene was deactivated by a deleterious mutation it becomes a pseudogene which evolves free from selective constraints and undergoes a progressive erosion of the signal by accumulation of numerous further mutations until the footprint of the original sequence becomes unrecognizable in the genome among the non-coding signal. When pseudogenization is not too old, mutations of this kind are still observable.

An extensive orthology/paralogy assessment is necessary to identify gene losses between different species. Recent analyses show that phylogeny-based methods are generally more reliable than similarity-based approaches. Phylogeny-based methods to detect relationships between sequences use reconciliation of species and gene trees to infer speciations and duplications; and to visualize the loss events. These methods have been studied since the 1970s [10–14]. Gene disappearances leading to extinction of functions have been identified in specific gene families and allowed the discovery of unitary pseudogenes [5–7,15]. Whole genomes sequencing has been made technically possible to study by comparative analyses of lineage-specific losses, bringing to light this major evolutionary process [16–18]. The increasing availability of complete genome sequences makes possible the investigation of these losses at a large scale, find co-elimination of functionally-connected groups of genes [17], and thus consider co-losses in different lineages. Many comparative analysis methods were developed to study the lineage-specific losses. The most commonly used method is the creation of orthologous groups by reciprocal blasts and inference of presence and absence of orthologs on a phylogenetic tree [19–23]. Detection of these losses can also be

performed by reading phylogenies [24]. In addition, other methods are specific to analyze the unitary pseudogenes [15,25–27] using the conserved synteny of neighboring genes. Totally automated methods to analyze lineage-specific losses and pseudogenizations are still lacking yet.

The aim of this study being the automation of the lineage-specific losses analyses, we developed a dedicated module that is part of the multi-agent system DAGOBAH [28] that we named **G**ene **L**oss **A**nalyzer **D**AGOBAH e**X**tension (GLADX). Each GLADX step was inspired by human expertise and engineered to closely mimic its characteristics. From a given sequence as input GLADX performs a gene phylogeny based on protein alignment of selected species-set, and by a tree reading method, detects the putative lineage-specific losses of the gene family. For all the candidate species to a lineage-specific loss, the module performs a comprehensive study to confirm losses and search pseudogenes. By a re-annotation systematic method of orthologous sequences recovered in genomes, GLADX is able to find and differentiate pseudogenes and intact but un-annotated genes present in databases and that would have been missed during previous rounds of genome annotation. The distinction between novel genes and pseudogenes is first performed at protein level by comparing the protein predictions, and complemented at nucleotide level when quality of sequences allow it. GLADX offers deeper insights on the pseudogenization, thanks to step of ancestral sequences reconstruction and to the analysis of mutations that occurred during evolution. GLADX offers the possibility to launch simultaneously several studies. For each sequence given as input it automatically resolves in the selected species-set all the events that occurred during the course of evolution of gene family. The evolutionary aspect is given based on gene phylogenies, ancestral sequence reconstruction, and a parsimony algorithm to locate the detected events. All events and traits found by GLADX are summarized and pinpointed on a user friendly species tree. The used innovative approach combines the quality resolution of phylogeny-based homology relations, a search at protein and nucleotide level, an evolutionary view of events, and total automation thereby substantially improving the set of tools which were available yet.

Methods

The DAGOBAH framework [28] in which GLADX is implemented uses the Prolog and Java languages. DAGOBAH framework is a set of agents running in parallel, sharing persistent results and that can communicate between each other and with external software platforms such as Ensembl, NCBI and FIGENIX [29–32]. An agent is like standalone software but belongs to an applicative context. DAGOBAH is designed to automatically predict and localize phylogenetically all the genetic events that occurred during the evolutionary history of genes. Within DAGOBAH, the GLADX module features 13 agents, some of which are not specific to GLADX and can be re-used in other contexts (i.e. gene phylogeny-building, ancestral sequence reconstruction, genes prediction). GLADX is not a standalone tool and depends to the established DAGOBAH framework.

The main purpose of GLADX is to automatically detect the lineage-specific loss, pseudogenization or presence of orthologous genes from a protein sequence in FASTA format given as input. In order to perform a reliable study, GLADX needs to use a database containing the complete proteomes and/or genomes of the desired set of species. The choice of species used by GLADX during studies needs to be specified (Text S1, A). A binary species tree containing these species needs also to be defined in GLADX. This binary species tree and the branch lengths can be easily changed by users. It is used at tree reconciliation step in gene phylogeny pipelines to deduce duplication events, and at different GLADX steps to define the relatedness of species. It is also used at the end of studies to perform an annotated species tree on which are summarized all the found events. It is expected that a change of this species tree can modify the detected losses and the placement of the reconstructed evolutionary events. (A view of the species tree implemented in the downloadable GLADX version is available, Text S3). As studies are performed, new data obtained after each important step is saved in a Report file and in an ontological database (Figure S1). The ontological records make it possible to restart at the last step performed so as to continue a study after an accidental stop. They allow also storing important data that may subsequently prove useful to a biologist or computational biologist.

Currently the GLADX version available for download is configured to study 22 *Chordates* species from Ensembl V57 and allows studying the pseudogenization (correspond to analyses in complete mode). The pseudogenization analyses require the genomes and proteomes sequences of studied species in GLADX. In this mode, a maximum of 51 *Chordates* from Ensembl version 48 to 58 can be analyzed. When studies are launched without research of pseudogenes (corresponds to simple mode), only proteomes are necessary. In this case the proteomes sequences can come from any database, and the number of species used is not limited. The addition of species requires completing the species tree used by GLADX.

To study lineage-specific losses, we developed an approach that detects an orthologs group stemming from a selected common ancestral species. It makes possible to determine the ancestor from which a gene is established and to find among species stemming from this ancestor, those having no representative of the ancestral gene. GLADX considers the orthologous group defined by the sub-tree which has a speciation node on the selected ancestor and containing the input reference gene. By default, if a speciation node in the defined ancestor does not exist, the next speciation node in the leaves direction is used. In a first example (Figure 1, the blue frame), the ancestor considered is the LCA of *Eutheria*. All genes present in the tree form an *Eutheria* orthologous group, because they are co-orthologs to the *Mus* gene. Despite possible loss of genes (here a human gene), no lineage-specific loss of the *Eutheria* ancestral gene will be detected because one representative is present as counterpart in each species of the set. In the second case, the LCA of *Catarrhini* was selected as ancestor. The gene should be present in *Homo*, *Pan* and *Macaca*. There are two sub-trees, but only one containing the reference gene used as input to build the tree will be analyzed. Here, this reference sequence is *Mmu1*. In the found orthologous group, the human gene is absent from the gene phylogeny. This is a *Homo* lineage-specific gene loss of a gene established since the LCA of *Catarrhini* (Figure 1, the red frame). A GLADX agent can be activated to systematically scan all the nodes along the lineage starting from the selected ancestor and leading to the used reference. Each node corresponding to the establishment of a new sub-family (a speciation node after a duplication event) is studied to find the lineage-specific losses of selected species (Text S1, G). This agent is available for all studies performed by GLADX, whatever the number of species used. By default, this option is not activated to allow choosing the kind of lineage specific losses searched. As example, to find *Vertebrate* specific losses (gene that was present in the *Vertebrates* ancestor, and subsequently lost in species of the phylum), the ancestor determined by the user should be those of *Vertebrates*. Several parameters can change how GLADX behaves and can modulate its execution (Text S1). By default lineage-specific losses

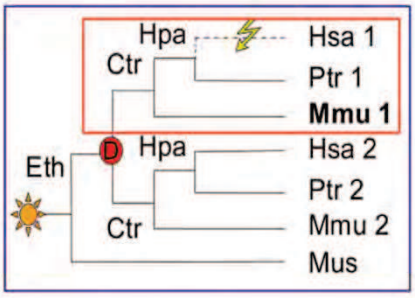


Figure 1. Approaches for detecting orthology and loss events: example of gene trees. Gene appearance detected by the phylogeny is depicted by the yellow star and the red circle represents gene duplication event. In bold is the gene used as input to build the tree. Based on the orthologous group from the Eth ancestor (blue frame) no lineage-specific loss is evidenced because each species has an ortholog to Mus. Based on the Ctr ancestor orthologous group, there are two orthologous groups, but only one group has an lineage-specific loss (red frame). Note that all abbreviations concerning species name and their ancestors are provided in Figure 6.
doi:10.1371/journal.pone.0038792.g001

of genes established since the LCA of *Eutelostomi* are analyzed (Text S1, B). The method implemented in GLADX is developed below in further detail through the different main steps.

The Gene Phylogeny as Starting Point

A FASTA format protein sequence is given as input. The first step builds a gene phylogeny based on protein alignment of species list. The gene phylogeny is built using an automated gene phylogeny pipeline available on the FIGENIX platform [32] connected to DAGOBAH. FIGENIX gives gene phylogenies with the speciation and duplication events annotated using the Forester detection algorithm [14] that compares a consensus tree obtained with the species tree defined in GLADX.

Detection of Species that have no Orthologs

From the gene phylogenetic tree, the PhyloPattern API [33] integrated to DAGOBAH, is used by GLADX to search patterns and so to automatically detect the largest orthologs group containing the initial query sequence, according to the user's choice of phylum to be studied. Cross-comparing the species-set present in the group of orthologous sequences detected against the list of the origin studied species-set, makes it possible to identify species which possess no orthologous gene. These species are candidates for lineage-specific gene loss.

A **simple** mode exists for GLADX, and avoids the verification of putative losses detected at this step. It jumps directly at final steps to display the presence and absence of an orthologous gene on leaves of the final phylogenetic view. Via a Dollo-like parsimony method, GLADX infers the loss events on the lineages in which they occurred. This mode is better suited to study old losses. Indeed, even if the loss comes from a pseudogenization event, after a long evolutionary time it is not necessary to search an ancient pseudogenized sequence whose traces should have erased under neutral evolution [34]. This mode optimizes the computation time and may be used with any protein sequence database.

With **complete** mode (by default), each putative lineage-specific loss is confirmed by a deeper analysis that is developed subsequently. This deeper analysis allows to find pseudogenes and novel genes that will be analyzed both at the protein and nucleotide level. Currently, GLADX was implemented to use the proteomics and genomics databases from Ensembl for studies

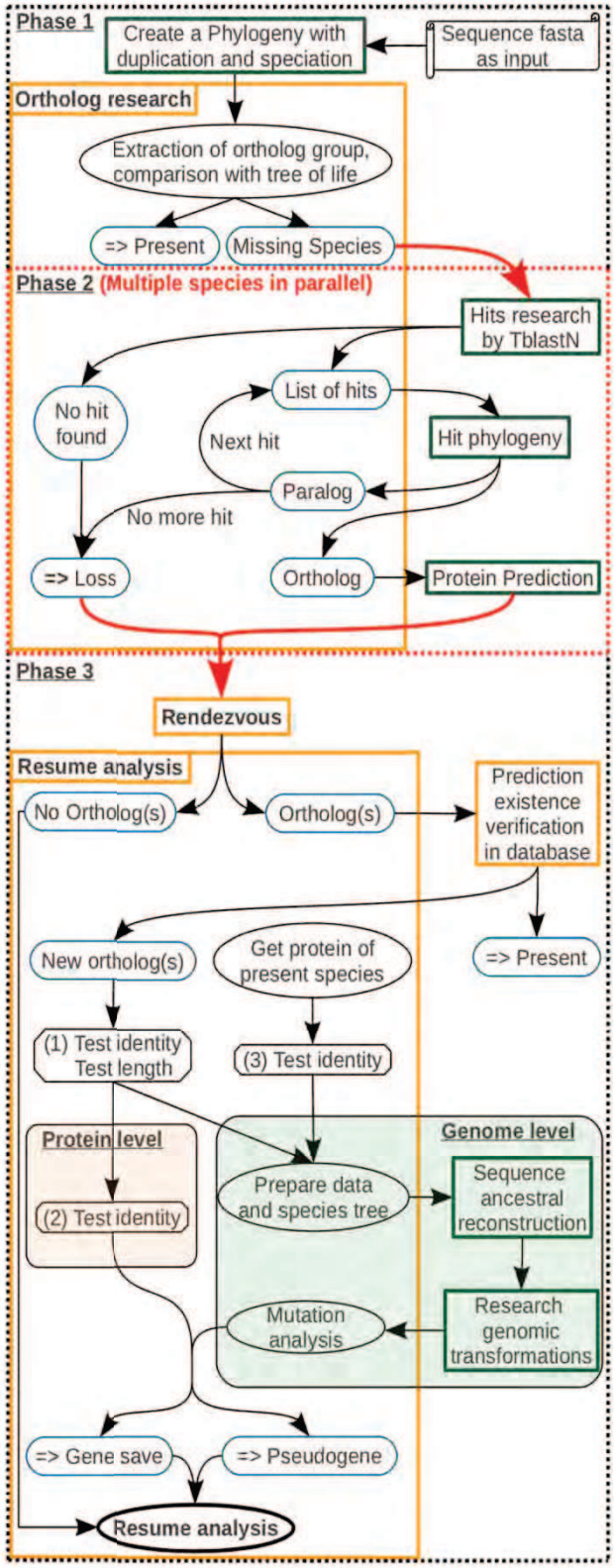


Figure 2. Method for identifying lineage-specific gene losses and pseudogenes. Parchment illustrates the necessary data for starting a study. Rectangles correspond to agents with their descriptions. The agents launching pipelines are in the green frame. Blue round-edge rectangles highlight essential results. Those with horizontal double arrows are final conclusions. Ellipses are a state or an action. Octagons consist in tests or analyses necessary to define the follow-up on the study. Arrows show the study pathway. Phase 1: Detection of species in which orthologs are missing. Phase 2: Parallel studies screening each orthologous sequence for each missing species. Phase 3: The red arrow is a factorization of the different species studied. The

rendezvous agent aims at waiting that all species targeted in the study have been done, before continuing. During this phase, GLADX tries to find the reasons explaining why the orthologs were missing. Depending on the sequences conservation, saved orthologs are analyzed either at protein level, or at nucleotide level, or the both.
doi:10.1371/journal.pone.0038792.g002

carried out on this mode, but other databases can be integrated by short developments.

Screening for Orthologous Sequences

To check that a gene is in fact completely absent from the genome sequence of a candidate species, GLADX scans the gene phylogeny and takes an orthologous protein of the closest species. GLADX uses this orthologous protein as a reference to find putative homologous sequences within the genome of the candidate species, using the TBLASTN algorithm [35]. If no hit is detected, a gene loss is inferred; but if putative homologous sequences are found, GLADX checks among them for orthologs. To check orthology of the putative homologous sequences, a phylogenetic approach is used again. At this step, to get the best sequence signal as possible to produce reliable gene phylogenies, an implemented method screens the TBLASTN hits, searching those that are tidy, carried by the same chromosome, sharing an identical direction, and that are close to each other, to concatenate them together. Afterwards, the created sequences are sorted by blast score decreasing order in a list (Text S1, C). Gene phylogenies from hits are built one after another, until one hit is found orthologous to at least one protein of the orthologous group defined at the first step. If no orthologous sequence stands out from candidate sequences, it then infers a gene loss. But if an ortholog is retrieved, the software still have to confirm whether this orthologous sequence corresponds to an already annotated gene eliminated by the gene phylogeny [31], an un- or mis-annotated gene, or to a pseudogene.

Analysis of Saved Orthologs

Coding sequences are better conserved at protein level than at nucleotide level. To avoid running an analysis at nucleotide level when observations are made impossible by high involved divergence, the analysis of each recovered orthologous sequence is first done at protein level. For this purpose, a protein prediction is built from a large piece of DNA containing the orthologous signal found by TBLASTN. Predictions are built using a pipeline embedded to the GenePredix pipeline [31], modified to take a reference protein sequence and a DNA sequence as inputs. Its aim is to predict the most similar protein to the reference protein from the DNA sequence. Among the orthologous sequences group defined at the first step, GLADX chooses as reference a protein of a species that is phylogenetically closer to the species from which the DNA sequence studied comes from. Once the prediction is done, GLADX systematically screens the database used to test whether a similar gene in the genome has already been described at the same location (Text S1, F). It allows verifying that the found gene does not correspond to a gene which is missing in the gene phylogeny of the study starting point. When a similar gene is already present, GLADX concludes that no lineage-specific loss occurred in the studied species. At this point, GLADX tests the orthology annotation of the regained gene, and saves the knowledge of this new orthology if the information was missing in the used database. Whereas no similar gene was described at the location, GLADX concludes that the predicted protein was never described. Then, an analysis will be performed to discern if the predicted protein comes from a putative gene, unless the orthologous nucleotide sequence found is a pseudogene and the predicted protein should not exist. To choose the depth of the next analysis (protein or nucleotide level), a test of length and similarity of the orthologous protein sequence found by TBLASTN and of the protein prediction that followed from it, is performed, using the known orthologous proteins as reference (Figure 2, test 1; Text S1, D). The similarity test is performed using the Needleman-Wunsch algorithm [36]. The length ratio, expressed in percentages, is calculated using the length of the reference protein. When the length and the similarity percentage of both tested orthologous protein sequences are under the user-defined thresholds, the study remains at protein level; otherwise, when the features of one of analyzed orthologous protein sequences exceeds the user-defined thresholds, a study is performed at nucleotide level. The reason both protein sequences are tested is that the prediction may not be sufficient and worse than the "hit" sequence when the DNA sequence has nonsense mutations. Indeed, if nonsense codons are present in the nucleotide sequence, the prediction must avoid them correctly, by moving the prediction start or end, splicing them into introns, or changing the reading frame. Therefore the predicted protein will be shortened or no protein will be predicted. Alternatively, blast hits are unhampered by nonsense codons. There they have a low impact on sequence recovered.

Analysis at Protein Level

In cases involving a protein-level analysis, GLADX uses the best protein predicted from the DNA of the orthologous TBLASTN hits found. To check whether or not each retrieved orthologous sequence is a putative pseudogene, it is necessary to check whether the conservation of the predicted protein is consistent with the divergence time observed among its orthologous sequences (Figure 2, test 2). We assume that in an orthologous group the protein sequence conservation should remain proportional to the divergence time between species that carry them. This consistency can be tested in two ways depending on species that possess known orthologous proteins. In the first case, it exists two species which do not share the same LCA with the species in which the protein sequence is being investigated (Figure 3, A). The test will be positive when the similarity percentage between the recovered orthologous protein sequence and the less-diverging protein is

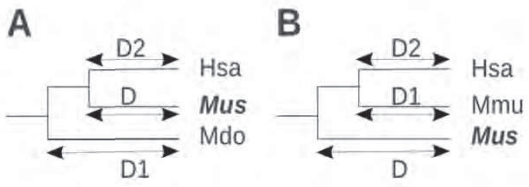


Figure 3. Test used at protein level to conclude on putative gene or putative pseudogene. The species where the protein is tested is highlighted in bold type. Other species have known proteins. *D1* and *D2* are age of divergence, in millions of years. Note that all abbreviations concerning species name and their ancestors are provided in Figure 6 (A) The species in which the protein is analyzed is surrounded by species having different LCAs. If the sequence identity is higher between *Mus* and *Hsa* as compared to *Mus* and *Mdo* thus, we can conclude on a putative gene; otherwise on a putative pseudogene. (B) If all species with known protein have the same LCA as the species under investigation, a calculation step is necessary. *Value1* is the percent identity between *Mus* and *Hsa*. *Value2* is the percent identity between *Mmu* and *Hsa*. A minimum relative threshold is calculated by multiplying *D2* distance by the similarity percentage *Value2* and by dividing the total by the distance *D*. If the similarity percentage *Value1* is superior to the minimum relative threshold calculated, we conclude on a putative gene; otherwise, we conclude on a putative pseudogene.
doi:10.1371/journal.pone.0038792.g003

higher than the similarity percentage with the protein from the more divergent species. In the second case, all available orthologous proteins of the study come from species that shares the same LCA with the studied species (Figure 3, B). The divergence times between the protein under investigation and the known proteins are identical. In this case, an average of decreasing of similarity percentage per million years according to the divergence observed between the known proteins is calculated. The test will be positive when the similarity percentage of the tested protein with a known protein is higher than the minimum similarity percentage expected according to their divergence time. When the conservation of the predicted protein gives a consistent result, we conclude on the putative existence of that orthologous gene in the candidate genome under study; otherwise, we conclude that the orthologous sequence undergoes a pseudogenization (Figure 2, test 2).

Analysis at Nucleotide Level

When the orthologous TBLASTN hit or the orthologous protein prediction successfully passes the first test (Figure 2, test 1), an analysis at the nucleotide level is performed in order to decipher and unravel the genetic events that affected the sequence during evolution. All recovered orthologous sequences that must be analyzed at nucleotide level, are tested together in one step. Together, the DNA sequence of all the recovered orthologous sequences, and those from known orthologous genes that are found not too divergent (Figure 2, test 3, Text S1, D), are sent to an ancestral sequence reconstruction-dedicated pipeline. No more as one sequence by species is used. This pipeline breaks down into two steps. Step one uses "LaganM" [37], a multiple aligner based on the CHAOS local alignment tool that combines speed and high accuracy for large sequences. It aligns the orthologous DNA sequences and compacts the coding sequences areas (Figure 4). The compaction process consists in only keeping the coding region signal in order to cut computation time and improve the quality of the subsequent processing work. Step two uses "Ortheus" to perform ancestral sequence reconstruction from the sequence alignment [38]. It builds a phylogenetic tree, and using efficient stochastic graph-based dynamic programming methods, it builds a multiple-sequence ancestor alignment, which contains explicit ancestor sequences for every node of the phylogeny. Identifying the ancestral sequence is highly valuable for revealing the step-by-step series of genetic events occurring during the gene evolution in a lineage. However, unlike routine alignments where indels are not interpreted, this multiple-sequence ancestor alignment is able to differentiate insertions from deletions. It is important to consider that to take in account the phenomenon of allele sorting [39] and to follow the real gene history, "Ortheus" is configured to reconstruct the ancestral sequences following its proper phylogeny, estimated by neighbor joining method with HKY model. Despite one sequence by species is used, the outcome may be a tree that is different to the species tree (a parameter can be added to force reconstruction based on a selected phylogeny). Working from this alignment each sequence, requiring an analysis, and its closest ancestral sequence built are sent to an agent dedicated to reveal the genetic mutations that appear between the ancestral sequence and the sequence analyzed (ancestral or contemporary). As we do not know the exons on the sequence retrieved and also in the reconstructed ancestor, the agent uses the exon positions inferred from a known gene and present in the alignment (Figure 5). Sequence comparison then makes it possible to test: i) the presence or absence of the start and/or final nonsense codons; ii) the occurrence of nonsense, insertion and/or deletion mutations in the open reading frame; iii) the loss and the modification of splice sites; and iv) the loss of exons. Thus, if the analysis does not return any degenerate mutations, we conclude on the putative existence of this orthologous gene in the genome of the candidate species; otherwise, it concludes on a pseudogenization.

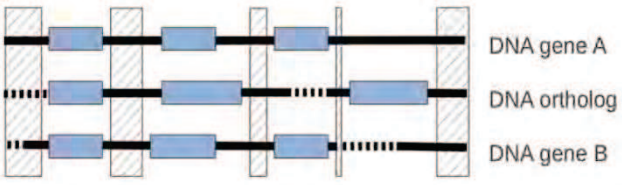


Figure 4. Processing of an alignment and conservation of informative signal of sequences. The hatched area is deleted area, blue boxes are exons (exons of retrieved orthologs come from prediction), bold black lines are DNA, and bold dotted lines are gaps.
doi:10.1371/journal.pone.0038792.g004

Synthesis of Results

The obtained results are summarized on the species tree used in GLADX on which all information is highlighted. On one hand, for each species, the state of presence of one representative of the gene-of-interest family is indicated on leaves of the tree by the character Present, Saved, Pseudogene or Lost. These characters may give insight about the current state of the function associated to the family of the gene-of-interest in the species studied. Sankoff parsimony [40,41] is used to highlight the ancestral and derived traits, making it possible to highlight the evolutionary aspect by defining the event occurrence dates of lineage-specific pseudogenization and loss of gene. Moreover, it allows calculating the ancestor from which the studied gene family seems to be born. On the other hand, the mutations found at nucleotide level are displayed directly on the phylogenetic tree. These genetic mutations are directly observable in an evolutionary dimension, as they show branch-by-branch the mutations found at nucleotide level, which occurred since the last ancestor having the gene intact, until the contemporary sequences that were investigated (Figure 6).

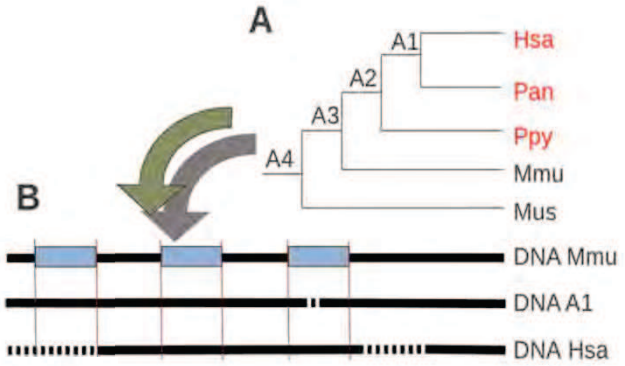


Figure 5. Example of processing sequences from a multiple sequences alignment and their ancestral sequences reconstructed. (A) Tree of species used for the reconstruction step. Species in red have sequence orthologs retrieved by GLADX. *A1*, *A2*, etc. correspond to ancestors reconstructed by Ortheus from orthologous sequences. Once this ancestral reconstruction is finished, the scan step is launched. (B) Sequences considered for an analysis at nucleotide level. Five scans will be carried out (*Hsa* vs *A1*, *Pan* vs *A1*, *Ppy* vs *A2*, *A1* vs *A2*, and *A2* vs *A3*). Only *Hsa* against *A1* is described here. The exon projection is necessary to each scan. Bold black lines are DNA, bold dotted lines are gaps, and red lines are projections from reference exons. Note that all abbreviations concerning species name and their ancestors are provided in Figure 6.
doi:10.1371/journal.pone.0038792.g005

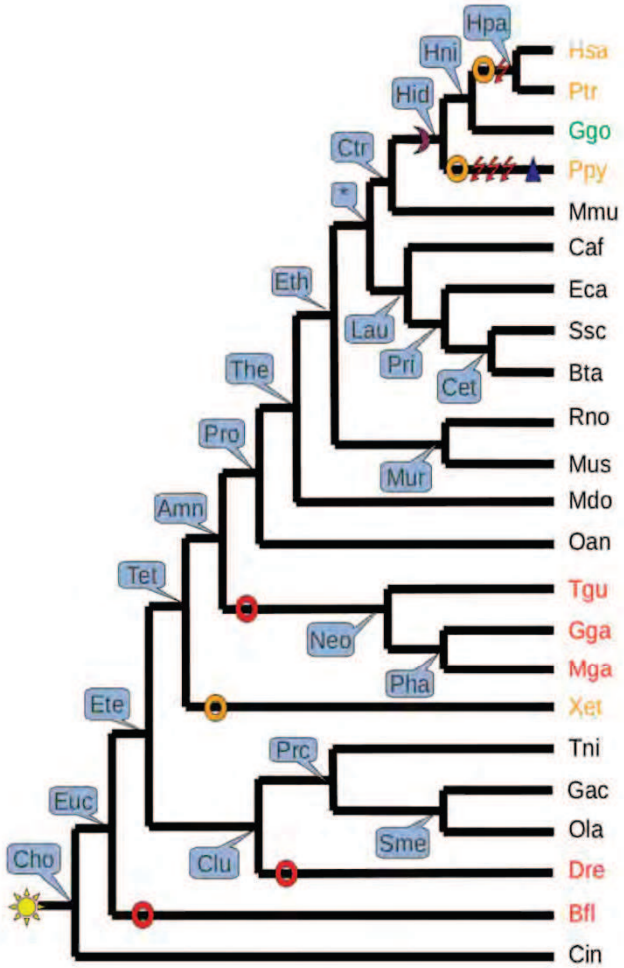


Figure 6. Summary of events occurred during evolution of *Acyl3* in 23 species. Each ancestor is indicated in blue frame by their abbreviation. All events are pinpointed on the specific branches. The asterisk indicates ancestor subject of controversy. Red rings show loss events; Orange rings show pseudogenization events; Magenta moon is a splice site mutation; Red flashes represent nonsense codon appearance; Blue triangle is insertion. The yellow star is the point of gene appearance in the phylogenetic tree. The species in red are species that have lost the gene. The species in orange have an orthologous sequence considered as pseudogene, and species in green is species where the recovered gene is considered as potentially intact. The complete name of species and ancestors are described in the following abbreviation paragraph. Name abbreviation of species: Bfl: Branchiostoma floridae; Bta: Bos taurus; Caf: Canis familiaris; Cin: Ciona intestinalis; Dre: Danio rerio; Eca: Equus caballus; Gac: Gasterosteus aculeatus; Gga: Gallus Gallus; Ggo: Gorilla gorilla; Hsa: Homo sapien; Mdo: Monodelphis domestica; Mga: Meleagris gallopavo; Mmu: Macaca mulatta; Mus: Mus musculus; Oan: Ornithorhynchus anatinus; Ola: Oryzias latipes; Ptr: Pan troglodytes; Ppy: Pongo pygmaeus Abelii; Rno: Rattus norvegicus; Ssc: Sus scrofa; Tgu: Taeniopygia guttata; Tni: Tetraodon nigroviridis; Xet: Xenopus tropicalis; Name abbreviation of ancestors: Amn: Amniota; Cet: Cetartiodactyla; Cho: Chordata; Clu: Clupeocephala; Ctr: Catarrhini; Ete: Euteleostomi; Eth: Eutheria; Euc: Euchordata; Hid: Hominidae; Hni: Homininae; Hpa: Homo/Pan ancestor; Lau: Laurasiatheria; Mur: Murinae; Neo: Neognathae; Pha: Phasianidae; Pri: Perissodactyla; Prc: Percomorpha; Pro: Prototheria; Sme: Smegmamorpha; Tet: Tetrapoda; The: Theria;
doi:10.1371/journal.pone.0038792.g006

Results

To demonstrate the efficiency of GLADX, we used it in complete mode on 23 Chordate species (see Figure 6). It was first benchmarked on 14 cases of unitary pseudogenes or gene losses in the human lineage described in the literature. This benchmark is a positive control demonstrating the tool's capacity to detect and correctly analyze the lineage-specific events occurring during the evolution of orthologous genes. We then performed a negative control to verify that GLADX does not overpredict gene loss or pseudogenization events. We obtained convincing results, as described below.

The 14 results obtained during the positive control test-runs are described in *Table S1* and summarized in *Table 1*. They are also available on the IODA (Interface for Ontological Data Analysis) user-friendly website (http://ioda.univ-provence.fr/). From 14 literature cases, our results are in line with results previously published in 13 cases. The events analyzed occurred at different times in the human lineage. Some unitary pseudogenes are human-specific, while others have undergone pseudogenizations that have begun in an ancestor. Our analysis is in fact more complete, as GLADX makes it possible to show what happened in all lineages and species of the dataset.

In 2 cases, GLADX found exactly the same results on pseudogenization in the human lineage, and in 9 cases, GLADX identified interesting new information and sometimes further details allowing us to refine the previous descriptions. Furthermore, it identified and described more genetic mutation events, and in 4 cases it was able to date the beginning of pseudogenizations, more ancestrally than those previously described. These refined results were not surprising given that we often used more species in our analysis. It is due also to the fact that the ancestral sequence reconstruction step enables a sharp detail of genetic events that occurred during the evolutionary course, which is a feature that guarantees increased usage of this method in the future. To illustrate this precision gain, take the example of the gene coding for acyltransferase 3 (*Acyl3*) protein [26], which the authors described as an unitary pseudogene in *Homo* and *Pan* due to a common nonsense mutation that appeared in exon seven after *Gorilla* diverged from the human lineage and before the Homo-Pan split. With GLADX, we not only found the mutations already described but also many other hitherto not described mutations (Figure 6). We discovered that a splice mutation appeared before the LCA of *Hominidae* and after *Macaca* diverged from the *Hominidae* lineage that seems to be the first event leading to the pseudogenization. Independently, four nonsense mutations and an insertion of four bases occurred in *Pan* after the Homo-Pan split and one nonsense mutation (previously described) occurred before the LCA of *Homo* and *Pan* and after *Gorilla* diverged from the human lineage. In addition, the analysis also revealed a loss of the *Acyl3* gene in the lineage leading to *Neognathae* from the LCA of *Amniota*. The fact that three species have lost the gene in *Neognathae* reinforces the idea that it is not a sequencing artifact. Loss of this gene also occurred in *Branchiostoma* lineage after the split with the LCA of *Chordata*, and in the *Danio* lineage after the split with the LCA of *Clupeocephala*. We also found a pseudogenization in the *Xenopus* lineage occurring after the split with the LCA of *Tetrapoda*. Indeed, we found an orthologous sequence of the *Mus* gene that was still present in the Xenopus genome, but the signal was too low and it was under the threshold configured to make it analyzable at the nucleotide level. Three possibilities could explain this fact: the first would be a pseudogenization that has begun in the not too distant past, with the result that the signal has not yet been totally erased; the second possibility is that the gene has evolved more rapidly than in other species, making its similarity percentage lower than the average of most other species; and the third explanation may be that another type of event occurred during this gene's evolution, such as a shuffling, partial gene loss,

Table 1. Benchmark results of 14 pseudogenes cases described in literature.

Gene symbol	Publications	Agree	Precision	Artifacts highlighted	Comments
2310042E22Rik	[26,27]	no	/	/	The gene was saved and detected as intact.
Gulo	[5,26,27]	yes	-	**	Problem of sensibility, tracks of pseudogenes are not detected
Acyl3	[26,27]	yes	+	/	/
Uox	[6,26,27,45]	yes	=	/	/
Ctf2	[26,27,46]	yes	+	/	/
Nradd	[26,27,47]	yes	+	*	/
Nepn	[26,27,47]	yes	+	*	/
Mup4	[27,48]	yes	+	*	/
T2r2	[49]	yes	+	/	Pseudogene was described as polymorphic
Tas2R134	[27,49]	yes	=	/	Tas2R134 and Tas2R143 *Mus* genes are co-orthologs to human pseudogene
1110012D08Rik	[27]	yes	+	*	/
Gpr33	[26,27,47]	yes	+	/	/
Slc7a15	[26,27,47]	yes	+	* **	Some mutations are not seen due to artifacts
Sult3a1	[42]	yes	-	**	Artifacts leads to a scan in false frame, but the pseudogenization was confirmed manually

The ***Publications*** column indicates references of studies of the literature used as reference for comparison with GLADX results.
The ***Agree*** column contains yes if the case is consistent with the literature results and no when the result is in contradiction.
The ***Precision*** column indicates the quality of results obtained: "-" means low precision, "+" means a better precision, "=" means we found exactly the same results, and "/" means it can't be interpreted.
The ***Artifacts highlighted*** column indicates cases where artifacts are present: "*" are cases where GLADX found artifacts in databases (Text S2); "**" means artifacts caused by tools implemented in GLADX, and "/" means no artifact was observed.
The ***Comments*** column indicates some particularities, "/" means no particular comment.
doi:10.1371/journal.pone.0038792.t001

etc. Neither the losses nor the pseudogenization in *Pongo* and *Xenopus* were described before.

In 2 cases our results are in line with published results, although GLADX provides less accurate conclusion. First, take the example of the *Sult3a1* gene described as a pseudogene in *Homo* [42]. Results given by GLADX show a pseudogenization that began in the LCA of *Catarrhini*. However, the mutations found in *Homo* are different to those described in the literature study. Exploiting manual expertise, we found that in contrast to other species, primates have orthologous to *Sult3a1 Mus* sequences without introns. We can deduce that the LCA of *Eutheria* should, in the most likely scenario, display the gene with introns. During the reconstruction step, only *Canis* gene was kept as primates outgroup. Sequences without introns came out over-represented, and the reconstructed ancestral sequence of the LCA of *Eutheria* did not have introns. Despite this prediction hiccup, no error was introduced into the reconstructed exon sequences. As the gene was unknown in primates, to perform the scan at nucleotide level, the exons that were used in primates were modeled by the exon structure of the *Canis* gene. However, studied sequences are not perfectly aligned with the *Canis* gene. Indeed, one or two bases of studied exons that should be positioned in front of the end of *Canis* exons 2, 3 and 4, are positioned in front of introns (Figure 7). These misplaced bases cannot be observed by GLADX, which studies the sequences only opposite to known positions of reference exons. Consequently these missing bases cause a reading frame-shift, and the mutations found by GLADX does not reflect reality of events occurring at nucleotide level during evolution of the gene. With a manual verification, we observe clearly that primates sequences contain numerous harmful mutations. Primates have pseudogenes, and pseudogenization seems to have begun at least in the LCA of *Catarrhini*. The pseudogenes in *Gorilla*, *Pongo*, *Pan* and *Homo* do not have introns and surely represent retroposon fixed at least since the LCA of *Catarrhini*. It is interesting to note that the original orthologs to *Mus Sult3a1* gene (ortholog from position) have been lost in these species, keeping only the retroposon which has become the unitary pseudogene in primates.

Another case in which GLADX results are less precise is the one of *Gulo* gene. It is an undergoing pseudogenization since at least the LCA of *Catarrhini*, and the bit of gene that remains in *Homo* contains a high number of mutations [5,26,27]. GLADX, via TBLASTN, detects the pseudogenized sequences in the *Catarrhini* species used, but their signals are too low to build gene phylogenies, and without phylogenetic confirmation of the orthology, GLADX concluded on a loss of *Gulo* in *Catarrhini*.

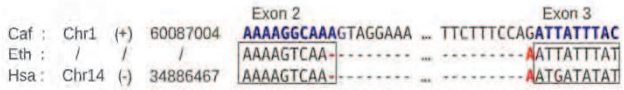


Figure 7. Inherent error generated during alignment processing of *Sult3a1* gene. Section of multiple alignment of the *Sult3a1* gene retrieved from the output result of ancestral sequences reconstruction step. From left to right, there is the species' name abbreviation, the chromosome's number, the strand in parenthesis, the position and the DNA sequence. The three dots represent parts of sequence not shown here. In bold and blue, are the exons described in *Caf* gene, with their number written above. The fragments of sequences that will be scanned are defined from the exon inference of the *Canis* gene, and are highlighted by a frame. The mis-position of nucleotides is highlighted in red. As consequence there is a frame shift which will not be detected, during the sequence scan.
doi:10.1371/journal.pone.0038792.g007

The loss of *Gulo* function detected by GLADX is consistent with the high pseudogenization of the *Gulo* gene reported in other studies.

In contrast, the only result that was discordant with previously published information concerns the *2310042E22Rik Mus* gene. It has been described as a pseudogene in *Homo* [26,27], but GLADX detected it as intact without harmful mutation. This gene is also saved in *Macaca*, is a pseudogene in *Pan*, and is lost in *Sus*. It appeared by duplication in the LCA of *Eutheria*, and a gene phylogeny of *Euteleostomi* phylum shows that the gene family exists at least since the LCA of *Tetrapoda*.

We then ran a second series of studies as a negative control testing a set of random genes, which are known to have an ortholog present in the human genome. In a majority of cases, the results agree with Ensembl annotation. When a gene is noted as present or pseudogene, the gene is effectively found present or pseudogene. In rare cases, GLADX results disagree with Ensembl as it found a potential functional gene that Ensembl has noted as pseudogene (as was the case for ortholog to *2310042E22Rik Mus* gene in *Homo*).

Numerous missing annotations and annotation errors are present in databases. The missing annotations are not a problem with GLADX because of the systematic annotation step which enables the finding of novel genes. During our analyses, it annotated several putative novel genes (*2310042E22Rik, Gpr33, Ctf2, Slc7a15, T2R2*) in different species (*Homo, Gorilla, Macaca, Mus, Bos, Equus, Ornithorhynchus, Oryzias*), which further demonstrates the tool's ability to re-annotate sequences. Currently, the annotation errors as over-predictions are not detected automatically by GLADX. A review of the results is necessary to find suspicious cases. In our analyses, some results do not fit with already published results and/or give non-parsimonious results due to the presence of a gene in an unexpected species. After manual expertise, these outcomes seem be the result of over-predicted genes in the database. These suspicious annotations have been found for *Gorilla* in predictions of *Nradd, Nepn, Mup4, 1110012D08Rik and Slc7a15* genes, as well as for *Macaca* in predictions of *Slc7a15* gene (Text S2). They have been re-annotated by relaunching GLADX omitting their presence in database (Text S1, B). The relaunched studies seem to give better results that are more parsimonious, and these are the results being analyzed in this paper. The last kind of artifacts that can be encountered using GLADX, are those that occurred due to limits of the GLADX-integrated tools. As we found this type of problem for alignment in the case of *Sult3a1*, we also found a problem on the ancestral sequence reconstruction step in the case of *Slc7a15* (Text S2). Artifacts do not necessarily have a dramatic impact on the results; nevertheless unusual results must be interpreted cautiously.

Discussion

We created the GLADX module implemented in the DAGOBAH framework as an attempt to totally automate the analysis of lineage-specific gene losses. The performed benchmark demonstrated the efficiency and power of GLADX to answer a majority of cases with details on gene losses or pseudogenization events. Its use has underlined the importance of the quality of genomic data and annotations available in databases. We have already seen that missing annotations are not a problem for GLADX which is able to annotate novel genes. As for mis-predicted and over-predicted genes in databases, they can be a real problem for analyses, as they not only give a false view of gene presence in the species concerned but also engender a mis-reconstruction of states of presence and absence in ancestors. The suspicious predictions can easily be detected upstream of the study by testing the intron size or the presence of initiator codon; or downstream by detection of particular and unusual patterns in the results of the phylogenetic tree produced as an output of the GLADX study. We have also seen that GLADX offers the possibility to easily and accurately re-annotate these selected suspicious annotations. GLADX represents an essential tool for analyzing the evolutionary history of orthologous genes groups, more specifically the gene family's retention in lineages. As GLADX is completely automated, it can be used at high-throughput to analyze a wide-range of gene datasets, with the additional strength that it can also be used on any *Metazoan* species dataset. As the number of complete genomes increases, the quality of analyses performed with GLADX will increasingly improve.

The fact that GLADX was developed in the DAGOBAH framework eases adding of new functionalities, and several new sources of data can be used. Moreover, it is possible to implement additional manual expertise. These two features can improve the quality of the results and their interpretation. For example, use of EST or mRNA databases can confirm the transcriptional activity of a pseudogene or saved gene. In the case of a pseudogene, the impact of any mutation detected by GLADX, on the transcript formed can be demonstrated. These databases can also be helpful to confirm any mutations detected, to analyze polymorphism and found potential mistakes, on the genomic sequences used. Furthermore, other databases may contain useful information such as those specialized in the sequence polymorphism [43], although outside of the human genome, which has been extensively researched, there is currently still insufficient data. Integration in GLADX of tools such as PAML [44] can highlight the kinds of selective pressures that sequences are subjected to. A pseudogene will be confirmed by neutral evolution, whereas a saved gene may be confirmed, and its behavior better understood, by positive or purifying selection. It is also possible to slightly modify GLADX to answer other questions or provide a different field view. In the near future, by integrating concepts linked to lateral gene transfers, it should be possible to create a specific version dedicated to studying bacterial genomes.

To conclude, in addition to GLADX being dedicated specifically to studying gene loss and pseudogenizations that are lineage-specific, other DAGOBAH agents are specialized in identifying through phylogenetic analyses, other event types, such as new protein architectures, duplications, and more. All these events are saved in an ontological database allowing to cross-check the evidences and deduce events of higher-level. Analyses based on evolutionary biology approaches allow to detect if several events occur at the same time, and precisely to show convergence and co-convergence. This brings to recognize links between environmental shifts and genetic and functional shifts, to better understand the evolutionary processes.

Supporting Information

Figure S1 Class Diagram of GLADX ontology.
(TIF)

Table S1 Summary of benchmarking results.
(RTF)

Text S1 Description of GLADX parameters.
(RTF)

Text S2 Analyses of artifacts.
(RTF)

Text S3 GLADX user's manual.
(PDF)

Acknowledgments

Benjamin Dainat, Lourdes Hernandez, Antonio Hernandez Lopez and Anthony Levasseur for reading and editing earlier versions of the manuscript. Thanks to both anonymous reviewers for improving the manuscript.

Author Contributions

Conceived and designed the experiments: JD PG PP. Performed the experiments: JD. Analyzed the data: JD. Contributed reagents/materials/analysis tools: JD PG PP. Wrote the paper: JD. Designed GLADX: JD PG PP. Developed GLADX: JD PG. Developed the website to visualize the results: JP PG. Prepared the GLADX image for the distribution: PG. Carried out revision of the manuscript: PG PP.

References

1. Hahn MW, Kern AD (2005) Comparative genomics of centrality and essentiality in three eukaryotic protein-interaction networks. Molecular Biology and Evolution 22: 803–806.
2. Jordan IK, Rogozin IB, Wolf YI, Koonin EV (2002) Essential Genes Are More Evolutionarily Conserved Than Are Nonessential Genes in Bacteria. Genome Research 12: 962–968.
3. Zhang L, Li W-H (2004) Mammalian housekeeping genes evolve more slowly than tissue-specific genes. Molecular Biology and Evolution 21: 236–239.
4. Lynch M (2000) The evolutionary fate and consequences of duplicate genes. Science 290: 1151–1155.
5. Nishikimi M, Fukuyama R, Minoshima S, Shimizu N, Yagi K (1994) Cloning and chromosomal mapping of the human nonfunctional gene for L-gulono-gamma-lactone oxidase, the enzyme for L-ascorbic acid biosynthesis missing in man. The Journal of Biological Chemistry 269: 13685–13688.
6. Wu XW, Lee CC, Muzny DM, Caskey CT (1989) Urate oxidase: primary structure and evolutionary implications. Proceedings of the National Academy of Sciences of the United States of America 86: 9412–9416.
7. Varki A (2001) Loss of N-Glycolylneuraminic Acid in Humans: Mechanisms, Consequences, and Implications for Hominid Evolution. Yearbook of Physical Anthropology 69: 54–69.
8. Olson M (1999) When less is more: gene loss as an engine of evolutionary change. American Journal of Human Genetics: 18–23.
9. Mitchell A, Graur D (2005) Inferring the pattern of spontaneous mutation from the pattern of substitution in unitary pseudogenes of Mycobacterium leprae and a comparison of mutation patterns among distantly related organisms. Journal of Molecular Evolution 61: 795–803.
10. Goodman M, Czelusniak J, Moore GW, Romero-Herrera AE, Matsuda G (1979) Fitting the Gene Lineage into its Species Lineage, a Parsimony Strategy Illustrated by Cladograms Constructed from Globin Sequences. Systematic Biology 28: 132–163.
11. Page RDM (1994) Maps between trees and cladistic analysis of historical associations among genes, organisms, and areas. Systematic Biology 43: 58–77.
12. Eulenstein O, Mirkin B, Vingron M (1997) Comparison of Annotating Duplication, Tree Mapping, and Copying as Methods to Compare Gene Trees with Species Trees. In B. Mirkin, F. McMorris, F. Roberts, and A. Rzhetsky, editors. Mathematical Hierarchies and Biology. Providence: American Mathematical Society. 71–94.
13. Ronquist F, Huelsenbeck JP (2003) MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics 19: 1572–1574.
14. Zmasek CM, Eddy SR (2001) A simple algorithm to infer gene duplication and speciation events on a gene tree. Bioinformatics 17: 821–828.
15. Brawand D, Wahli W, Kaessmann H (2008) Loss of egg yolk genes in mammals and the origin of lactation and placentation. PLoS Biology 6: 508–517.
16. Ponting CP (2008) The functional repertoires of metazoan genomes. Nature Reviews Genetics 9: 689–698.
17. Aravind L, Watanabe H, Lipman DJ, Koonin EV (2000) Lineage-specific loss and divergence of functionally linked genes in eukaryotes. Proceedings of the National Academy of Sciences of the United States of America 97: 11319–11324.
18. Moran NA (2002) Microbial minimalism: genome reduction in bacterial pathogens. Cell 108: 583–586.
19. Koonin EV, Fedorova ND, Jackson JD, Jacobs AR, Krylov DM, et al. (2004) A comprehensive evolutionary classification of proteins encoded in complete eukaryotic genomes. Genome Biology 5: R7.
20. Hughes AL, Friedman R (2005) Loss of ancestral genes in the genomic evolution of Ciona intestinalis. Evolution & Development 7: 196–200.
21. Hughes AL, Friedman R (2004) Shedding genomic ballast: extensive parallel loss of ancestral gene families in animals. Journal of Molecular Evolution 59: 827–833.
22. Zdobnov EM, von Mering C, Letunic I, Torrents D, Suyama M, et al. (2002) Comparative genome and proteome analysis of Anopheles gambiae and Drosophila melanogaster. Science 298: 149–159.
23. Krylov DM, Wolf YI, Rogozin IB, Koonin EV (2003) Gene loss, protein sequence divergence, gene dispensability, expression level, and interactivity are correlated in eukaryotic evolution. Genome Research 13: 2229–2235.
24. Kuraku S, Kuratani S (2011) Genome-wide Detection of Gene Extinction in Early Mammalian Evolution. Genome Biology and Evolution: 1–45.
25. Wang X, Grus WE, Zhang J (2006) Gene losses during human origins. PLoS biology 4: e52.
26. Zhu J, Sanborn JZ, Diekhans M, Lowe CB, Pringle TH, et al. (2007) Comparative genomics search for losses of long-established genes on the human lineage. PLoS Computational Biology 3: e247.
27. Zhang ZD, Frankish A, Hunt T, Harrow J, Gerstein M (2010) Identification and analysis of unitary pseudogenes: historic and contemporary gene losses in humans and other primates. Genome Biology 11: R26.
28. Gouret P, Paganini J, Dainat J, Louati D, Darbo E, et al. (2011) Integration of evolutionary biology concepts for functional annotation and automation of complex research in evolution: The multi-agent software system DAGOBAH. In: Springer-Verlag, editor. Evolutionary Biology – Concepts, Biodiversity, Macroevolution and Genome Evolution. 71–87.
29. Flicek P, Aken BL, Ballester B, Beal K, Bragin E, et al. (2010) Ensembl's 10th year. Nucleic Acids Research 38: D557–D562.
30. Sayers EW, Barrett T, Benson DA, Bryant SH, Canese K, et al. (2009) Database resources of the National Center for Biotechnology Information. Nucleic Acids Research 37: D5–D15.
31. Gouret P, Vitiello V, Balandraud N, Gilles A, Pontarotti P, et al. (2005) FIGENIX: intelligent automation of genomic annotation: expertise integration in a new software platform. BMC Bioinformatics 6: 198.
32. Paganini J GP (2012) Reliable phylogenetic trees building: a new web interface for FIGENIX. Evolutionary Bioinformatics. In press.
33. Gouret P, Thompson JD, Pontarotti P (2009) PhyloPattern: regular expressions to identify complex patterns in phylogenetic trees. BMC Bioinformatics 10: 298.
34. Li W-H, Gojobori T, Nei M (1981) Pseudogenes as a paradigm of neutral evolution. Nature 292: 237–239.
35. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. Journal of Molecular Biology 215: 403–410.
36. Needleman SB, Wunsch CD (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. Journal of Molecular Biology 48: 443–453.
37. Brudno M, Do CB, Cooper GM, Kim MF, Davydov E, et al. (2003) LAGAN and Multi-LAGAN: Efficient tools for large-scale multiple alignment of genomic DNA. Genome Research: 721–731.
38. Paten B, Herrero J, Fitzgerald S, Beal K, Flicek P, et al. (2008) Genome-wide nucleotide-level mammalian ancestor reconstruction. Genome Research 18: 1829–1843.
39. Degnan JH, Rosenberg NA (2009) Gene tree discordance, phylogenetic inference and the multispecies coalescent. Trends in Ecology & Evolution 24: 332–340.
40. Sankoff D, Rousseau P (1975) Locating the vertices of a Steiner tree in an arbitrary metric space. Mathematical Programming 9: 240–246.
41. Sankoff D (1975) Minimal mutation trees in sequences. Society for Industrial and Applied Mathematics 28: 35–42.
42. Freimuth RR, Wiepert M, Chute CG, Wieben ED, Weinshilboum RM (2004) Human cytosolic sulfotransferase database mining: identification of seven novel genes and pseudogenes. The Pharmacogenomics Journal 4: 54–65.
43. Altshuler DM, Gibbs RA, Peltonen L, Dermitzakis E, Schaffner SF, et al. (2010) Integrating common and rare genetic variation in diverse human populations. Nature 467: 52–58.
44. Yang Z (2007) PAML 4: phylogenetic analysis by maximum likelihood. Molecular Biology and Evolution 24: 1586–1591.
45. Oda M, Satta Y, Takenaka O, Takahata N (2002) Loss of urate oxidase activity in hominoids and its evolutionary implications. Molecular Biology and Evolution 19: 640–653.
46. Derouet D, Rousseau F, Alfonsi F, Froger J, Hermann J, et al. (2004) Neuropoietin, a new IL-6-related cytokine signaling through the ciliary neurotrophic factor receptor. Proceedings of the National Academy of Sciences of the United States of America 101: 4827–4832.
47. IHGSC (2004) Finishing the euchromatic sequence of the human genome. Nature 431: 931–945.
48. Chamero P, Marton TF, Logan DW, Flanagan K, Cruz JR, et al. (2007) Identification of protein pheromones that promote aggressive behaviour. Nature 450: 899–902.
49. Go Y, Satta Y, Takenaka O, Takahata N (2005) Lineage-specific loss of function of bitter taste receptor genes in humans and nonhuman primates. Genetics 170: 313–326.

Supplementary Text S1 : Description of GLADX parameters. ⇔ Annexe 4

Supplementary Figure S1 : Class Diagramm of GLADX ontology. ⇔ Annexe 5

Supplementary Table S1 : Summary of benchmarking results. ⇔ Annexe 6

Supplementary Text S2 : Analyses of artifacts. ⇔ Annexe 7

Supplementary Text S3 : GLADX user's manual. ⇔ Annexe 8

Chapitre III - Étude à grande échelle de pertes de gènes unitaires dans la lignée humaine depuis l'ancêtre des Eucaryotes

L'efficacité du module GLADX pour l'étude des pertes de gènes unitaires et de la pseudogénisation a été démontrée dans le « *benchmark* » décrit dans la publication portant sur GLADX (Article 2). Après la mise au point de cet outil, qui permet une analyse de qualité et automatisée des pertes de gènes unitaires, le second objectif de cette thèse est de développer une stratégie permettant une analyse à grande échelle de ce type d'événement dans une lignée spécifique. Je me suis alors lancé dans l'étude des pertes de gènes unitaires dans la lignée menant à l'homme. Le choix de la lignée humaine s'est fait en raison de l'intérêt que suscitent naturellement les études anthropocentriques, et de sa cohérence avec le projet EvolHHuPro du laboratoire qui porte sur l'évolution des protéomes des Chordés. L'ajout dans le projet EvolHHuPro de l'analyse des pertes de gènes unitaires dans la lignée humaine apporte des informations sur l'évolution du protéome des Chordés et rend exhaustive l'analyse de l'évolution du génome humain. Le génome humain a également été choisi car il est très bien séquencé, ce qui rend les analyses de pertes de gènes unitaires plus fiables. J'ai choisi d'élargir le champ d'investigation de l'étude jusqu'à l'ancêtre des Eucaryotes afin d'avoir une vision évolutive complète du phénomène.

Grâce à la méthode d'analyse de GLADX décrite dans le chapitre précédent, la détection de chaque perte de gène unitaire entraîne une recherche systématique de la présence d'un pseudogène unitaire. On peut ainsi détecter parmi les pertes de gènes unitaires, les pseudogènes unitaires et les processus de pseudogénisations.

Quelques équipes de recherche ont analysé les pertes de gènes dans la lignée humaine, en recherchant les gènes orthologues absents dans les bases de données (méthode des GOs), ou en recherchant les traces des pseudogènes dans les séquences génomiques.

J'ai comparé mes résultats à ceux de ces équipes, afin d'évaluer leur qualité et l'efficacité de GLADX. Les résultats enrichissent et affinent les études existantes ; ils peuvent également permettre d'observer les cas de pertes dans d'autres perspectives, voire de les réévaluer.

1 Stratégie

Dans la stratégie développée pour mener à bien l'étude à grande échelle et exhaustive des pertes de gènes unitaires au sein d'une lignée spécifique, l'utilisation du module GLADX est un élément central. Cependant, GLADX utilise la phylogénie : il est difficile de résoudre la phylogénie de l'ensemble des gènes des espèces d'un phylum pour savoir si chacun des gènes

était présent dans un ancêtre donné et s'il a été perdu dans la lignée étudiée. Il faut donc sélectionner attentivement les gènes à étudier pour optimiser le travail de GLADX. Mais pour l'étude des pertes de gènes au sein d'une lignée, il n'est pas possible de partir directement des génomes des espèces de la lignée étudiée puisque les gènes perdus en sont absents. Pour contourner la difficulté il faut utiliser les génomes d'autres espèces, et les traiter par la génomique comparative afin d'en déduire les génomes ancestraux. C'est à partir de ces génomes ancestraux que l'on peut mettre en évidence toutes les pertes qui se sont produites par la suite dans les génomes de la lignée étudiée. Pour estimer rapidement les génomes ancestraux, l'approche passe par une méthode de génomique comparative basée sur la clustérisation de séquences similaires, pour créer des groupes de séquences orthologues (GOs), et par l'analyse de parcimonie des espèces représentées dans ces GOs.

Comme souligné précédemment (Chapitre I, 2.2.2), la création de groupes d'orthologues par un algorithme de clustérisation est relativement rapide et aisément utilisable. Ensuite, l'analyse des GOs ainsi créés par l'inférence de l'état Présent/Absent des espèces représentées dans un GO sur un arbre des espèces, permet de détecter l'apparition et les pertes de gènes (Illustration 16). Ainsi, l'analyse de ces GOs permet de reconstruire les états Présent/absent ancestraux des gènes ; on détecte ainsi les GOs qui contiennent un gène apparu dans un ancêtre de l'homme, mais dont tous les représentants ont disparu dans la lignée humaine (perte de gènes unitaires). Bien que les algorithmes de clustérisation de groupes d'orthologues manquent de précision, ils restent un formidable outil pour filtrer rapidement les pertes de gènes putatifs, qui sont étudiées ensuite avec plus de précisions au moyen de GLADX.

1.1 État des connaissances

L'état des connaissances sur les études de pertes de gènes unitaires à grande échelle (Chapitre I, 2.2.2) montre que, dans la majorité des cas, les chercheurs utilisent des méthodes d'analyse de GOs d'Eucaryotes (KOGs) créées par clustérisation de séquences similaires. Ces méthodes, d'abord utilisées en microbiologie, sont devenues courantes pour étudier de nombreux génomes et observer le flux de gènes dans les lignées majeures d'Eucaryotes.

Parmi les séquençages des génomes « complexes », celui de l'homme a été précocement accessible : il a été utilisé dans de nombreuses études de pertes de gènes unitaires. Parmi les études publiées sur la perte de gènes unitaires chez l'homme, j'ai sélectionné les 10 études les plus pertinentes (Illustration 26). Dans l'illustration, les colonnes indiquent les ancêtres

utilisés comme référentiels. Les 10 premières colonnes correspondent aux études sélectionnées, la onzième surmontée d'une flèche rouge correspond à l'étude qui fait l'objet de cette thèse. Des informations complémentaires sont données en annexe (Annexe 1).

Parmi les études sélectionnées, qui abordent l'analyse des pertes de gènes unitaires dans la lignée humaine, il faut distinguer :

- Huit études qui ignorent les pseudogènes et le phénomène de pseudogénisation

  Ces 8 études se fondent sur l'absence d'orthologues dans les bases de données utilisées. Parmi ces études, 6 utilisent les méthodes de KOG (Danchin, Gouret, & Pontarotti, 2006 ; Hughes & Friedman, 2004a, 2004b, 2005 ; Koonin *et al.*, 2004 ; Krylov, Wolf, Rogozin, & Koonin, 2003 ; Wyder *et al.*, 2007), et 2 utilisent des méthodes d'analyse phylogénétique (Blomme *et al.*, 2006 ; Kuraku & Kuratani, 2011).

- Deux études qui recherchent systématiquement les pseudogènes et analysent la pseudogénisation.

  Une première étude (Zhu *et al.*, 2007) se base sur la conservation de la structure des gènes entre la souris et le chien pour rechercher des gènes équivalents chez l'homme. Les auteurs trouvent ainsi près de 76 pseudogènes unitaires chez l'homme. Elargie à d'autres Primates, l'analyse des pertes leur permet de préciser les dates de « désactivations » de ces gènes.

  Une deuxième étude (Z. D. Zhang *et al.*, 2010) fait figure d'exception car elle s'organise en deux parties distinctes. Dans un premier temps les auteurs utilisent la méthode de KOG comme filtre pour détecter les pertes putatives puis ils recherchent systématiquement la présence d'un pseudogène pour confirmer les pertes de gènes unitaires. Cette méthode permet une étude précise dans un phylum restreint (Homme, Souris), et nécessite beaucoup de manipulations et d'expertises manuelles.

Il existe 2 autres études pertinentes sur la recherche des pseudogènes chez l'homme (Annexe 1) (Hahn & Lee, 2005 ; X. Wang *et al.*, 2006). Je ne les ai pas représentées dans l'Illustration 26 pour les raisons qui suivent. Ces études comparent deux à deux les séquences du génome du chimpanzé et de l'homme. Cette approche permet de découvrir des pertes par pseudogénisation spécifique à l'homme. Mais aucune information n'est donnée sur le temps d'établissement de ces gènes dans les génomes ancestraux. Proviennent-ils de duplications récentes ou sont-ils des gènes établis depuis longtemps ? En effet, on sait que la séparation

homme-chimpanzé date d'environ 6 millions d'années. En raison de ce temps très court, les pseudogènes observés pourraient être des gènes apparus par duplication juste avant la séparation des deux espèces, et ces nouveaux gènes se seraient fixés chez le chimpanzé tandis qu'ils auraient subi une pseudogénisation dans la lignée humaine. Pour obtenir plus d'informations, les auteurs auraient pu utiliser, dans ces études, des espèces plus éloignées telle *M. musculus*, qui leur auraient permis de savoir si ces gènes sont établis de longue date. Dans ces études, en ajoutant des espèces proches, comme l'homme de Neandertal (Green *et al.*, 2010) disponible récemment, il est possible de spécifier si les pseudogénisations ont eu lieu directement après la séparation homme-chimpanzé, ou après la séparation homme-Neandertal.

Parmi les 10 études sélectionnées, j'ai relevé que dans 7 études, un seul génome ancestral de la lignée humaine est reconstruit. Les analyses qui en découlent étudient uniquement les pertes des gènes présents chez cet ancêtre. Trois autres études utilisent des reconstructions du génome de plusieurs ancêtres de la lignée humaine (Blomme *et al.*, 2006 ; Koonin *et al.*, 2004 ; Wyder *et al.*, 2007). Celle de Koonin (Koonin *et al.*, 2004) ne précise pas de quels ancêtres proviennent les gènes perdus. Les études de Blomme (Blomme *et al.*, 2006) et de Wyder (Wyder *et al.*, 2007) sont les seules qui différencient explicitement les pertes de gènes en fonction des ancêtres pour lesquels les gènes semblent être établis.

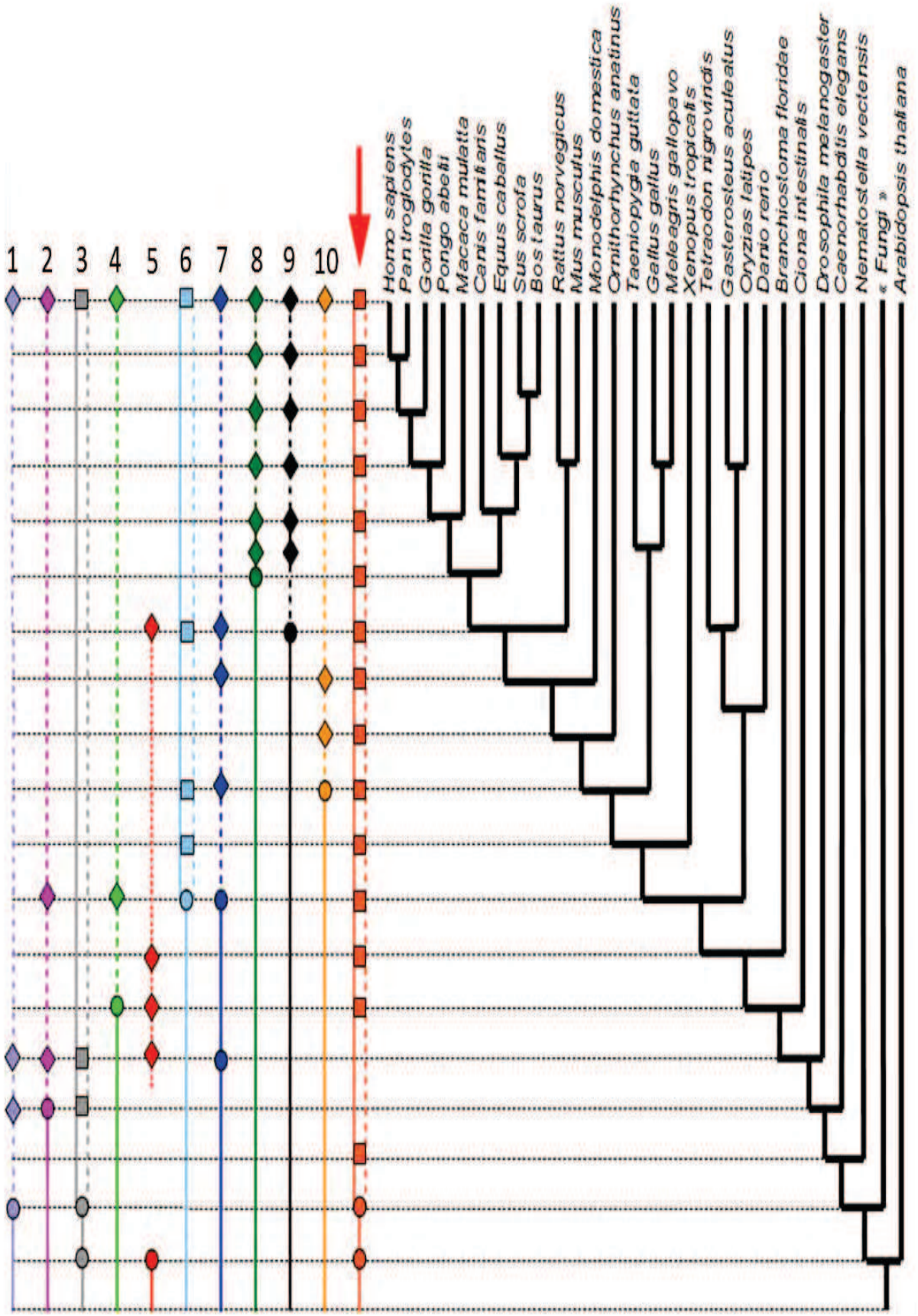


Illustration 26 : État des connaissances des principales études qui ont traité de la perte de gènes dans la lignée humaine

Cette illustration permet de voir l'ensemble des ancêtres utilisés comme référentiel pour étudier les gains et les pertes de gènes.

1)Krylov *et al.*, 2003 ; 2)Hughes & Friedman, 2004a, 2004b ; 3)Koonin *et al.*, 2004 ; 4)Hughes & Friedman, 2005 ; 5)Danchin *et al.*, 2006* ; 6)Blomme *et al.*, 2006 ; 7)Wyder *et al.*, 2007 ; 8)Zhu *et al.*, 2007 ; 9)Z. D. Zhang *et al.*, 2010 ; 10)Kuraku & Kuratani, 2011

○ *Gains de familles de gènes observés ;* ◇ *Pertes de familles de gènes observées. Lorsque ce signe ne figure en face d'aucun ancêtre, cela signifie que des espèces non représentées ici ont été utilisées ;* □ *Gains et pertes de familles de gènes observés.*

Les traits pleins verticaux (hors arbre phylogénétique) représentent la période où les familles de gènes présentes ont pu apparaître. Les traits en pointillés verticaux représentent la période pendant laquelle les gènes considérés comme perdus ont pu disparaître. La flèche rouge indique l'analyse effectuée dans le cadre de la thèse.

** Étude où C. elegans & D. melanogaster sont utilisés avec une topologie où ils possèdent le même plus proche dernier ancêtre commun avec la lignée humaine.*

L'Illustration 26 montre que la reconstruction de nombreux génomes ancestraux dans mon étude (colonne surmontée d'une flèche rouge) permet de gagner en précision. Ainsi, l'étude des gains et des pertes de gènes au cours de l'évolution dans la lignée humaine, est plus exhaustive. L'utilisation de l'ensemble des génomes ancestraux permet de définir précisément la période de gain ou de perte d'un gène. Elle permet également de savoir, pour un gène unitaire perdu, depuis quel ancêtre le gène était établi, et d'en déduire le temps de fixation du gène avant sa disparition.

1.2 Choix des espèces

Le choix des espèces peut dépendre du type de fonction que l'on souhaite étudier. L'étude que j'ai menée cherche à être la plus exhaustive possible, et ne cherche pas à cibler un type de fonction en particulier. L'étude des événements de gains et de pertes de gènes unitaires dans une lignée nécessite la connaissance des génomes ancestraux. Pour examiner ces événements dans la lignée humaine depuis l'ancêtre des Eucaryotes, au travers de nombreux ancêtres pour obtenir une information la plus précise possible, j'ai choisi 26 espèces de diverses lignées dont le génome est complètement séquencé (Illustration 27). En réalité j'en ai utilisé beaucoup plus. Par exemple, l'espèce annotée « Fungi » sur l'arbre d'espèces utilisé, représente un pan-génome formé à partir de 18 génomes de champignons (Annexe 9). J'ai profité d'une étude collaborative déjà effectuée au laboratoire sur ces génomes pour les intégrer à l'étude. Les caractéristiques des génomes et protéomes des espèces utilisées sont disponibles en annexe (Annexe 10). En revanche j'ai fait le choix de ne pas utiliser de génomes de Protostomiens comme celui de *C. elegans* ou *D. melanogaster*. En effet, ces génomes semblent trop réduits par des pertes massives au cours de leur évolution après leur séparation de la lignée humaine (Hughes & Friedman, 2005 ; Wyder *et al.*, 2007). En conséquence, j'ai estimé que la reconstruction de génomes ancestraux en utilisant ces espèces aurait amené peu d'informations pour étudier les pertes dans la lignée humaine.

Les 26 espèces sélectionnées, me permettent de reconstruire 16 génomes ancestraux de la lignée humaine (du LCA des Eucaryotes au LCA Homo-Pan). J'ai pu analyser les gains de gènes dans les 16 génomes (du LCA du groupe *Fungi-Metazoan* à l'homme) et les pertes dans 15 de ces génomes (du LCA des Eumetazoaires à l'homme) (Illustration 26). L'étude des gènes qui composent ces génomes, permet de déduire la période d'apparition des gènes et celle de leur disparition. Je peux ainsi classer les pertes de gènes en fonction de leur temps d'établissement dans les génomes, et mettre en évidence des pertes de gènes unitaires, de gènes qui sont plus ou moins stables dans les génomes. Le grand nombre d'ancêtres reconstruits, confère une bonne précision à ces études.

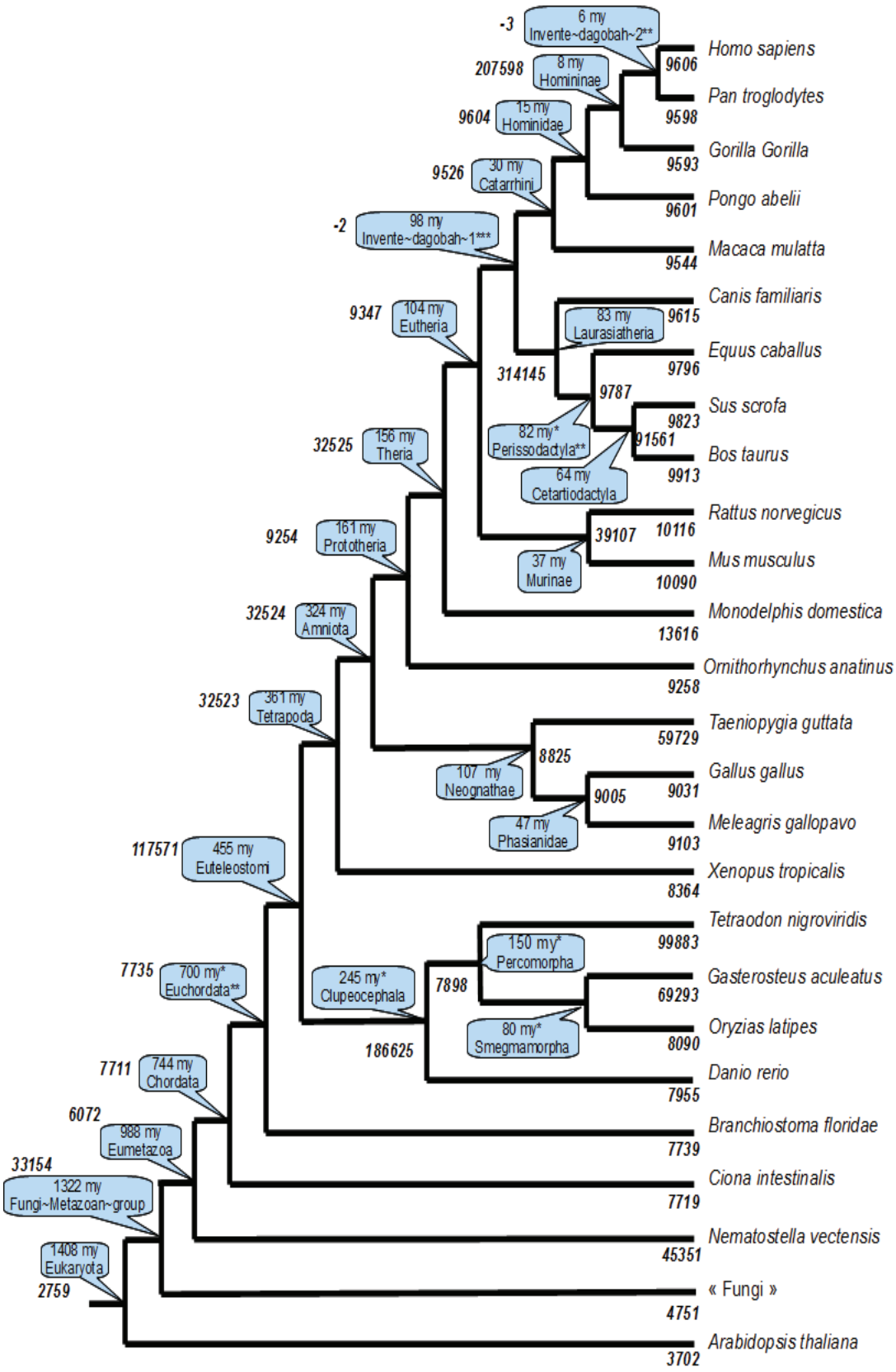


Illustration 27 : Arbre des 26 espèces utilisées

Dates de spéciations : Elles proviennent du projet TimeTree *(http://www.timetree.org/) de novembre 2011. Ces dates sont appelées à évoluer en fonction des informations provenant de nouvelles publications. Ces dates sont exprimées en millions d'années (ici notées my). Lorsque aucune date n'est disponible pour le nœud (date incohérente, ou inconnue), il est noté « * » et une date cohérente (date plus récente que le nœud ancestral et plus ancienne que les descendants) a été choisie.*

*Noms d'espèces et d'ancêtres : Les noms proviennent du site de taxonomie du NCBI (http://www.ncbi.nlm.nih.gov/Taxonomy/). Lorsqu'un ancêtre commun est peu clair il est noté par un râteau sur le site du NCBI, ce qui représente plus de deux espèces ayant comme plus proche ancêtre commun le même ancêtre. Ces cas sont notés par « ** » dans l'arbre et une topologie dichotomique est définie selon la topologie adoptée au laboratoire et un nom d'ancêtre est défini (noté «-» lorsque aucun nom est défini). Dans certains cas la définition de certaines topologies est encore discutée dans la communauté scientifique, et lorsque la topologie du laboratoire est différente de celle utilisée par le NCBI, elle est notée « *** » dans l'arbre.*

Identifiant de taxonomie : Pour chaque feuille et chaque nœud ancestral il existe un identifiant numérique unique inscrit en gras. Ces identifiants proviennent également du NCBI. Pour les nœuds ancestraux de notre topologie n'ayant pas de correspondance dans le site du NCBI, un nombre négatif unique a été choisi.

Etudier les pertes de gènes unitaires nécessite un triptyque minimum de 3 espèces. Une espèce où le gène est perdu, et deux autres qui possèdent un orthologue confirmant la présence ancestrale du gène. De par ce fait, les pertes, de gènes apparus après le dernier ancêtre commun des Homininés (*Homininae*), ne peuvent être étudiées.

1.3 Choix d'un algorithme de clustérisation de groupes d'orthologues

L'étude de GOs créés par un algorithme de clustérisation permet de filtrer rapidement les pertes potentielles. Pour créer ces GOs, plusieurs algorithmes existent :

- OrthoMCL (L. Li, Stoeckert, & Roos, 2003)
- INPARANOID (Remm, Storm, & Sonnhammer, 2001)
- EGO anciennement appelé TOGA (Lee *et al.*, 2002)
- SYNERGY (Wapinski, Pfeffer, Friedman, & Regev, 2007a)

L'algorithme OrthoMCL (L. Li *et al.*, 2003) a été choisi pour plusieurs raisons.

- L'algorithme INPARANOID utilise une stratégie basée sur le BLAST pour identifier les meilleurs hits réciproques entre seulement deux espèces. La possibilité de travailler avec plusieurs génomes à la fois avec OrthoMCL semblait plus en adéquation avec les perspectives choisies. De plus, INPARANOID est moins performant pour identifier les duplications récentes, c'est-à-dire que deux GOs issus d'une duplication récente ont

plus tendance à se retrouver dans un seul GO regroupant les deux. En utilisant OrthoMCL je cherche à analyser le plus grand nombre possible de familles de gènes.

- L'algorithme EGO est basé sur la méthode des COGs qui clustérise les protéines, en passant par une étape de BLAST réciproque, et en utilisant une approche triangulaire. La taille d'un GO est au minimum de trois protéines. EGO est facilement induit en erreur lorsqu'il traite des paralogues multiples et lorsque des orthologues sont absents dans les génomes incomplets. Les manques d'annotations, encore nombreux dans les bases de données, sont mieux gérés par OrthoMCL. De plus OrthoMCL permet de clustériser des familles composées seulement de deux gènes annotés.
- SYNERGY semble pouvoir apporter de meilleurs résultats, grâce à l'utilisation d'information de synténie et de phylogénie. SYNERGY utilise un arbre des espèces pour clustériser en prenant en compte la proximité phylogénétique des espèces utilisées. Ceci nécessite de connaitre la phylogénie des espèces, sinon SYNERGY peut engendrer des erreurs lors de la création des GOs. Il faut constater que des nœuds de spéciations sont mal identifiés parmi les espèces que j'étudie (entre le plus proche ancêtre commun de *C. familiaris* et de *M. musculus* avec la lignée humaine et entre le plus proche ancêtre commun de *N. vectensis* et de *C. intestinalis* avec la lignée humaine). OrthoMCL est donc plus sûr.

De plus, OrthoMCL est utile pour adjoindre de nouveaux génomes, même peu annotés, aux analyses déjà effectuées sans refaire tous les calculs.

La création des GOs par OrthoMCL se déroule en deux étapes essentielles. La première utilise un BLAST réciproque entre toutes les protéines de tous les génomes utilisés d'une manière similaire aux autres méthodes, et la seconde étape clustérise ces séquences avec le Markov Cluster Algorithm (MCL)(Van Dongen, 2000) en se basant sur la théorie des probabilités et celle des graphes (Illustration 28).

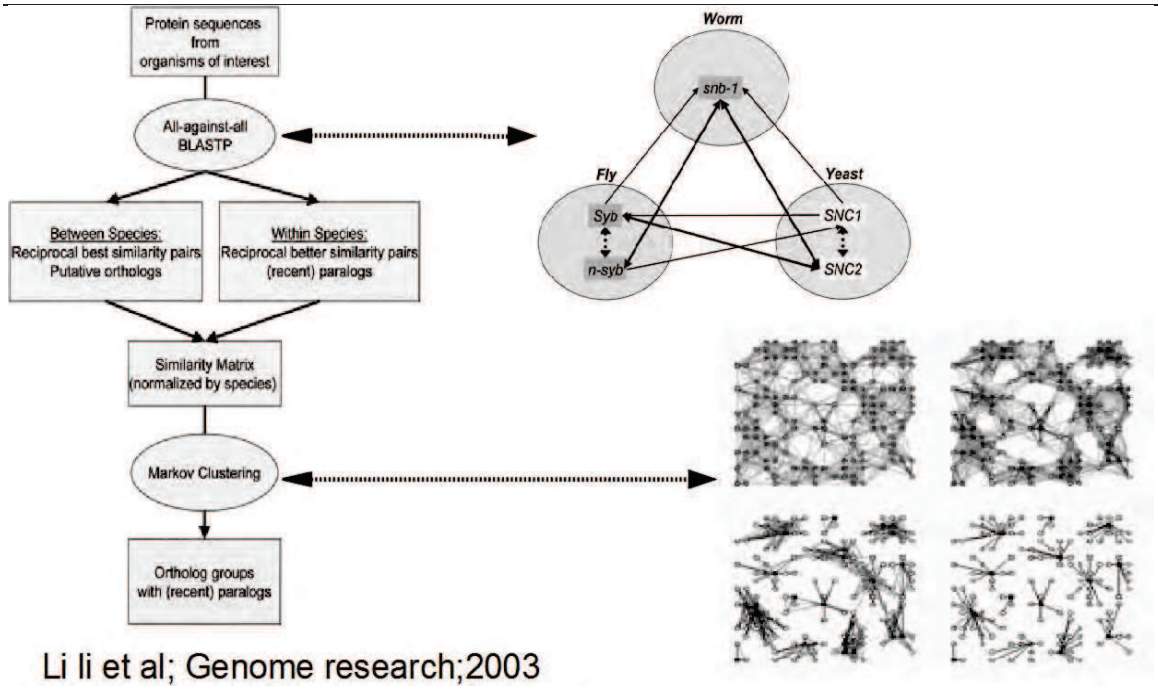


Illustration 28 : Présentation des étapes importantes d'OrthoMCL
Cette mise en page utilise les figures de la publication présentant OrthoMCL (L. Li *et al.*, 2003).

1.4 Filtrage des pertes putatives par l'analyse de groupes d'orthologues

La première partie de la stratégie développée pour analyser les pertes de gènes unitaires chez l'homme est similaire à celles des méthodes traditionnelles qui se basent sur la création de groupes d'orthologues (GOs). J'ai créé des GOs à partir du génome des 26 espèces choisies (voir ci-dessus) à l'aide de l'outil de clustérisation OrthoMCL. L'examen de ces GOs par l'inférence des espèces présentes dans chaque GO sur un arbre d'espèces, permet de savoir quand le gène qui a formé un GO est apparu, et dans quelles lignées et quelles espèces il a disparu (Illustration 16). Parmi les nombreuses méthodes de parcimonie pour inférer les états ancestraux, j'ai utilisé la parcimonie de Dollo. L'utilisation de cette parcimonie permet d'être exhaustif et de ne rater aucune perte putative. Les reconstructions par la parcimonie de Dollo ne sont pas troublées par les problèmes d'annotations des génomes utilisés. Par contre, avec cette parcimonie il est possible d'obtenir de nombreux faux positifs lors de l'analyse de GOs « artefactuels » produits à partir d'artefacts d'annotations de gènes (gènes mal prédits, contaminations) ou aux limites de l'outil de clustérisation. Toutefois, elle n'omet pas de pertes réelles. Je sélectionne ainsi tous GOs représentant des familles de gènes d'un ancêtre de la lignée humaine et qui semblent être perdues par la suite dans la lignée menant à l'homme. Ces

GOs représentent des pertes putatives dans la lignée humaine, et sont utilisés pour des analyses plus approfondies avec GLADX. Parmi ces pertes putatives, les éventuels faux positifs sont filtrés par la suite avec GLADX.

1.5 Analyse approfondie avec GLADX

J'ai sélectionné les GOs représentant des pertes putatives chez l'homme avant de les étudier avec GLADX. La taille des GOs est très hétérogène. Certains GOs possèdent deux protéines, d'autres plusieurs dizaines ; pour étudier un GO avec GLADX, il faut choisir comme référence une des protéines qui le constitue. Elle est choisie en prenant systématiquement la protéine de l'espèce phylogénétiquement la plus proche de l'homme.

Pour optimiser le temps de calcul, et ne pas faire de recherches inutiles de pseudogènes lorsque les pertes sont très anciennes, j'ai divisé les GOs en deux groupes. Le premier, dit « perte récente », rassemble les pertes putatives de gènes unitaires apparemment survenues après le LCA des Amniota (< 324 My), et un second, dit perte ancienne, où les pertes semblent survenues avant cet ancêtre (pertes > 324 My). Dans le groupe des pertes récentes, GLADX effectue systématiquement la recherche de pseudogènes. Dans le groupe des pertes anciennes, GLADX ne recherche pas *a priori* la présence de pseudogènes. Dans ce deuxième groupe, GLADX a mis en évidence des événements de pertes datés par l'analyse des GOs avant le LCA des Amniotes, qui étaient en réalité survenus plus récemment. Ces types de pertes sont soumis *a posteriori* à une analyse plus approfondie de GLADX incluant la recherche de pseudogènes (Illustration 31).

2 Résultats

2.1 Résultats de la création de groupes d'orthologues

La création de groupes d'orthologues (GOs) par OrthoMCL a été faite par agrégation, en ajoutant par étapes les protéomes des espèces partageant un ancêtre commun de plus en plus ancien. J'ai appelé ce processus « agrégation par strate ». Ce processus est réalisé en 11 étapes décrites ci-dessous (Illustration 29). Les 26 protéomes sont préalablement filtrés pour garder une seule protéine par gène (Annexe 10). La protéine choisie est l'isoforme le plus long décrit.

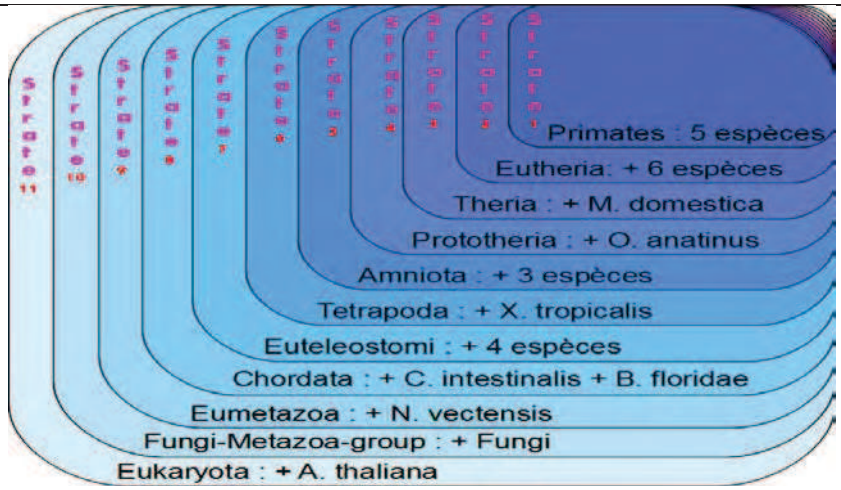


Illustration 29 : Clustérisation par strate permettant d'analyser des phyla de plus en plus anciens

A la fin des 11 strates de traitement par OrthoMCL, les protéomes des 26 espèces utilisées sont clustérisés ensemble. A partir de **555 449** protéines provenant des 26 protéomes, j'ai obtenu 22 558 GOs qui semblent représenter des familles de gènes présentes dans la lignée humaine. Ces 22 558 GOs ne contiennent plus que 389 236 protéines, soit 70% des protéines données en entrée. Les protéines non retenues (30%) sont :

- Les protéines en dessous du seuil de longueur minimum définie à 100 acides aminés. J'ai fixé ce seuil à 100 aa dans le but d'avoir un signal phylogénétique permettant de construire par la suite des phylogénies de bonne qualité. Avec ce seuil j'élimine de l'étude les protéines courtes, mais je couvre cependant la grande majorité des protéines, car la médiane de la taille des protéines chez l'homme est de 416 acides aminés (Scherer, 2008).
- Les protéines n'ayant pas d'homologues chez les autres espèces et que l'on peut appeler gènes orphelins.
- Les protéines contenues dans des GOs éliminés. En effet, j'ai éliminé certains GOs car ils représentaient des familles de gènes spécifiques à une lignée différente de celle de l'homme. Pour savoir si un GO représente une famille de gènes présente dans la lignée de *H. sapiens*, j'ai reconstruit l'état de présence des gènes dans l'ancêtre commun de toutes les espèces représentées au sein du GO par une approche de parcimonie. Quand cet ancêtre fait partie d'un ancêtre commun à *H. sapiens*, le GO est retenu dans l'étude ; sinon il est éliminé. Pour reconstruire des gènes ancestraux, j'ai utilisé la parcimonie de Dollo (Farris, 1977 ; Lequesne, 1972). J'ai opté pour la parcimonie de Dollo, car elle se base sur le postulat qu'il est plus facile de perdre un gène que de le

gagner : elle permet donc de reconstruire une seule apparition du gène.

Parmi les 22 558 GOs ainsi retenus représentant des familles de gènes présentes dans la lignée humaine, 5 901 sont des GOs où aucune protéine humaine n'est présente. Ces 5 901 GOs représentent des pertes putatives de gènes unitaires.

Pour être le plus exhaustif possible dans la détection des pertes putatives de gènes unitaires, j'ai effectué une recherche supplémentaire. J'ai comparé les 22 558 GOs de la strate finale (la plus ancienne) avec les GOs d'autres strates pour détecter si la clustérisation n'a pas regroupé ensemble des groupes paralogues proches à cause d'une protéine d'une espèce plus ancestrale. Par exemple (Illustration 30), j'ai comparé deux strates (strates 3 et 4) pour voir si dans une strate plus profonde, 1 GO ne contient pas deux GOs d'une strate moins profonde. En effet, les GOs d'une strate moins large peuvent correspondre à des sous-familles apparues par exemple par duplication. Par ce biais, j'ai pu mettre en évidence des sous-groupes d'orthologues qui n'étaient pas visibles dans la strate finale. Parmi eux, 336 représentent des groupes n'ayant pas de protéine de *H. sapiens*. Je les ai rajoutés à l'étude en suspectant que les pertes auraient pu avoir lieu dans une sous-famille apparue par duplication.

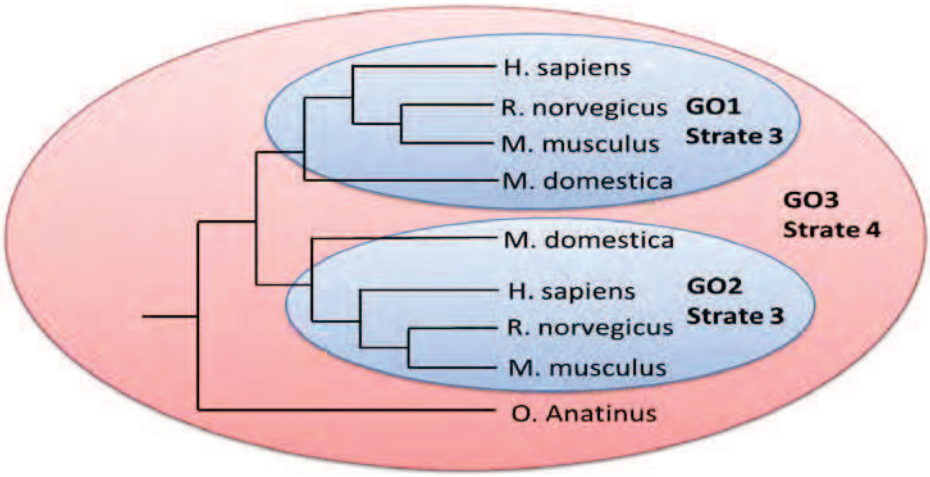


Illustration 30 : Incidence potentielle de l'ajout d'espèces extérieures lors de création de GOs
La strate 3 représente une création de GOs sans l'espèce O. Anatinus. Lorsqu'on rajoute l'espèce extérieure O. Anatinus, le groupe GO3 contient les deux groupes GO1 et GO2. GO1 et GO2 peuvent par exemple être des sous-familles apparues par duplication.

Aux 5 901 GOs représentant des pertes putatives de gènes unitaires détectés parmi les 22 558 GOs de la strate de clustérisation la plus ancienne, j'ai rajouté les 336 GOs issus de l'étude comparative des GOs entres les différentes strates, soit un total de 6 237 GOs, qui représentent des pertes putatives de familles de gènes dans la lignée humaine (Illustration 31).

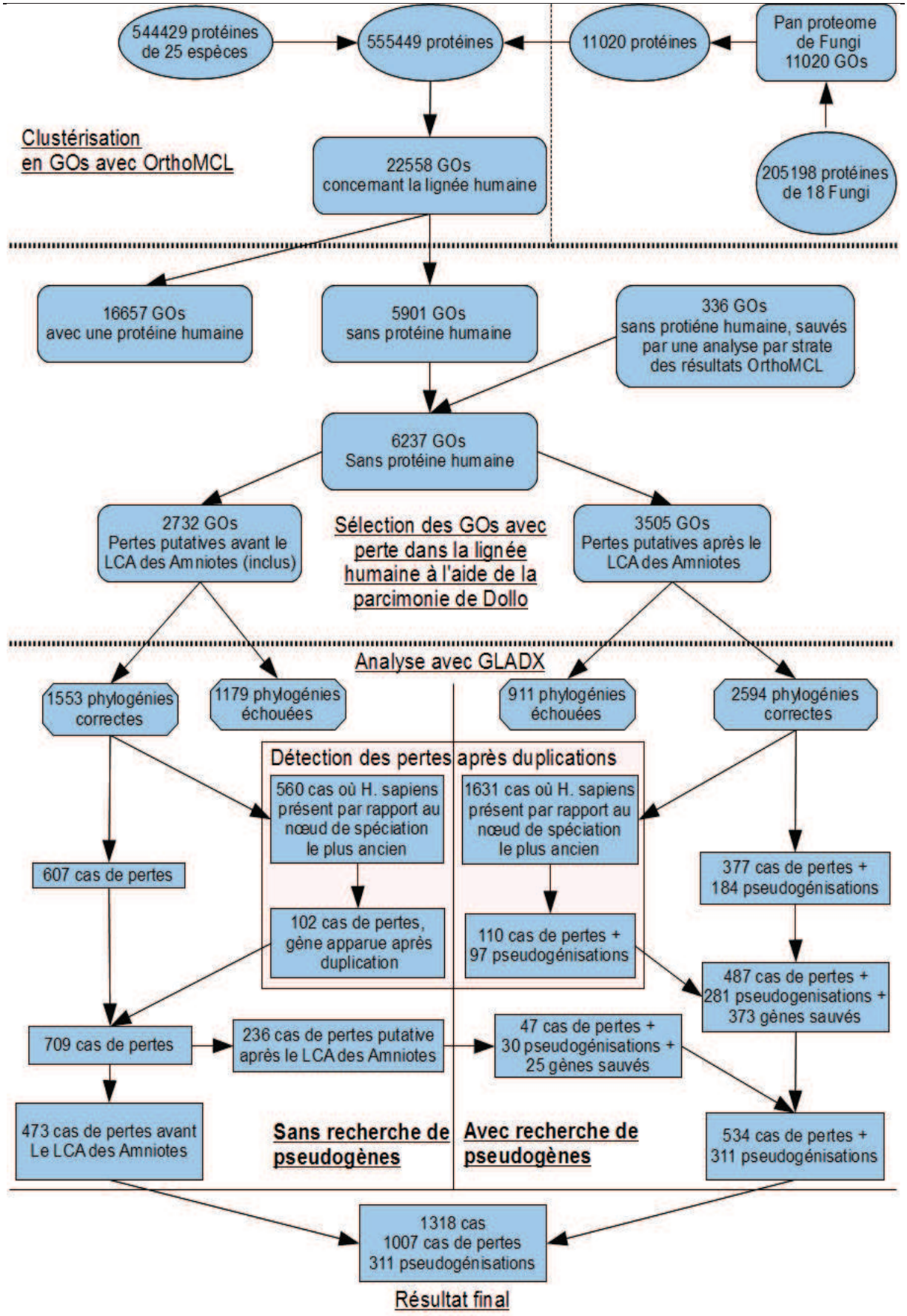


Illustration 31 : Résultats et résumé de la méthode utilisée pour détecter les pertes dans la lignée humaine

2.2 Résultats des analyses faites avec GLADX

L'analyse avec GLADX porte sur les 5 901 GOs représentant des pertes putatives de gènes unitaires détectés parmi les 22 558 GOs de la strate de clusterisation la plus ancienne, et les 336 GOs issus de l'étude comparative des GOs entres les différentes strates, soit 6 237 GOs sélectionnés. Sur ces 6 237 GOs, 2 732 GOs présentent des pertes putatives avant le LCA des Amniotes et 3 505 GOs des pertes putatives après le LCA des Amniotes. J'ai pris une protéine de référence dans chacun de ces 6 237 GOs, et lancé 6 237 études (Illustration 31). Parmi les 6 237 études, GLADX a initié seulement 4 147 études avec l'obtention d'une phylogénie (1 553 des 2 732 GOs + 2 594 des 3 505 GOs), soit 66,5% du total des études. Sur ces 4 147 GOs étudiés, GLADX trouve 1 318 pertes de gènes unitaires dans la lignée humaine. Sur ces 1 318 pertes de gènes unitaires, GLADX trouve 311 pseudogènes chez l'homme. GLADX annote également 398 (373 + 25) nouveaux gènes putatifs chez l'homme (Tableau 2 et Illustration 31).

Type d'événement	Nombre d'événements total	Evénements chez l'homme
Pseudogénisation	2 878	311
Perte sans pseudogène	8 382	1 007
Gène putatif annoté	1 967	398

Tableau 2 : Evénements détectés par GLADX dans l'ensemble des espèces utilisées

Si l'on additionne les événements détectés, les 2 878 pseudogénisations, les 8 382 pertes sans pseudogène et les 1 967 gènes sauvés, on obtient 13 227 événements répartis sur 43 espèces (dont 18 *Fungi*). Ces événements sont détectés au niveau des nœuds internes de la phylogénie ou sur les feuilles.

Lorsque GLADX effectue une recherche d'événements de pertes, il étudie toutes les espèces ayant une perte putative dans la lignée étudiée, à la lecture de la phylogénie de départ. Dans une petite proportion de GOs (638 sur 4 147), GLADX a recherché systématiquement les événements pour chaque étude (lorsque la phylogénie de départ fonctionne) sans prendre en compte la présence ou l'absence de *H. sapiens* comme facteur déterminant. Ces recherches sont plus gourmandes en temps de calcul car tous les événements sont systématiquement analysés. Dans un souci de gain de temps dans l'analyse des événements de la lignée humaine, j'ai paramétré GLADX pour qu'il ne continue les recherches d'événements que lorsqu'au moins l'orthologue de *H. sapiens* est absent dans la lignée étudiée.

L'annotation par GLADX de nouveaux gènes et de pseudogènes entraîne, lorsque c'est possible, une recherche de mutations génétiques. Pour GLADX, une séquence est considérée comme pseudogène lorsqu'un codon stop prématuré est présent. L'ensemble des mutations observées est synthétisé dans le tableau suivant (Tableau 3). Le nombre élevé de codons stops détectés vient du fait que les séquences sont analysées dans leurs cadres de lectures contemporains. Ainsi, nombre de mutations (indels) entraînent un décalage de phase qui provoque la détection de nombreux codons stop par le scanner dans la suite de la séquence. Ces codons stop apparus par décalage de phases peuvent être différenciés des codons stop apparus par substitution en comparant le codon contemporain et le trinucléotide ancestral correspondant. Comme les trinucléotides sont enregistrés dans l'ontologie, un codon stop dû à un décalage de phase sera déduit lorsque le trinucléotide ancestral correspondant est identique.

Type de mutation	**Nombre total d'événements**
Insertion	2 878
Délétion	8 382
Disparition d'exon	1 967
Disparition de codon stop final	55
Apparition de codon stop	12 228
Disparition de codon initiateur	74
Mutation de site d'épissage	525

Tableau 3 : Mutations observées par GLADX dans l'ensemble des espèces utilisées

2.3 Partage des résultats avec la communauté scientifique

Dans le cadre de l'ANR EvolHHuPro (*Evolutionary Histories of Human Proteome*) les études menées dans le laboratoire se décomposent en trois parties.

La première, consiste à produire dans la lignée des Chordés le phylome de toutes les protéines humaines (~20 000), et de tous les domaines protéiques contenus dans ces protéines. Au sein de DAGOBAH, les phylogénies du phylome ont permis d'analyser, au cours de l'évolution des Chordés, des modifications dans l'architecture en domaines des protéines.

La seconde consiste à créer des phylogroupes. Ce sont des groupes de GOs créés par OrthoMCL et agrégés ensemble après expertise des phylogénies disponibles les concernant.

La phylogénie sert à gagner en précision dans la création des GOs car elle est sensiblement plus précise et permet d'éviter un certain nombre d'artefacts liés à la création de GOs. Des phylogroupes ont été établis pour le phylum des Chordés.

La troisième partie porte sur l'analyse des pertes de gènes unitaires. Ce projet fait l'objet de cette thèse. J'ai mené l'analyse des pertes de gènes unitaires sur 23 espèces de Chordés, parmi lesquelles 13 espèces de Chordés utilisées pour le phylome du protéome humain. J'ai élargi le champ d'investigation en ajoutant les espèces *N. vectensis, Fungi, A. Thaliana* qui sont des Eucaryotes ayant divergé avant l'apparition des Chordés.

L'ensemble des données produites et analysées dans le laboratoire forme ce que l'on appelle le « *Chordate proteome history database* ». Les résultats contenus dans cette BD ontologique sont disponibles sur le site Internet I.O.D.A (***I****nterface for* ***O****ntological* ***D****atabases* ***A****nalysis*) à l'adresse suivante : « http://ioda.univ-provence.fr ». Le phylome du protéome ainsi que l'analyse de l'architecture en domaines des protéines sont disponibles dans l'onglet « *Domain event phylogenies* », les phylogroupes sont disponibles dans l'onglet « *Chordata phylogroups and gene loss light* », et enfin les études de pertes de gènes unitaires sont dans l'onglet « *Human lineage specific losses* » (Illustration 32).

L'interface Web de IODA développée au laboratoire permet notamment de récupérer la fonction des gènes par l'interrogation des bases de données fonctionnelles KEGG, ArrayExpress, String et QuickGO. Les données produites par GLADX dans le cadre de l'étude des pertes unitaires dans le génome humain sont facilement exploitables par l'interface du site Web. Il est ainsi possible d'observer et de contrôler toutes les étapes importantes des études faites avec GLADX. L'interprétation et l'exploitation des nombreuses données sont grandement facilitées (Illustration 33).

La « *Chordate proteome history database* », qui contient l'histoire complète du protéome humain au travers du phylome humain, l'analyse des événements de changement d'architecture en domaines des protéines ainsi que l'ensemble des études de pertes de gènes, a fait l'objet d'un article récent (Article 3 ci-après).

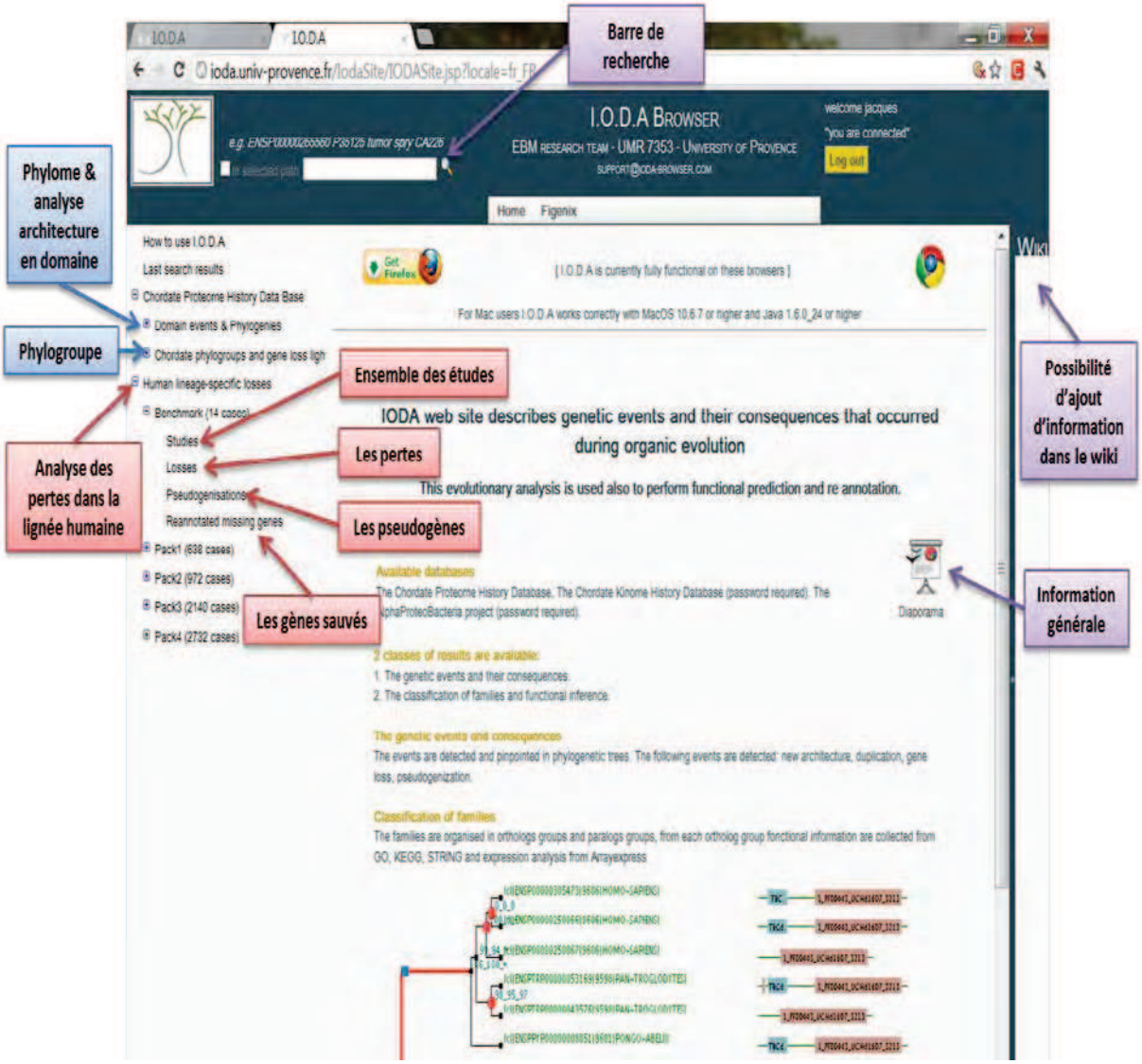


Illustration 32 : Page d'accueil du site IODA

Les cadres rouges indiquent la partie spécifique au projet d'analyse des pertes de gènes unitaires. Les cadres bleus indiquent les autres projets. Les cadres violets indiquent des éléments communs à tous les projets.

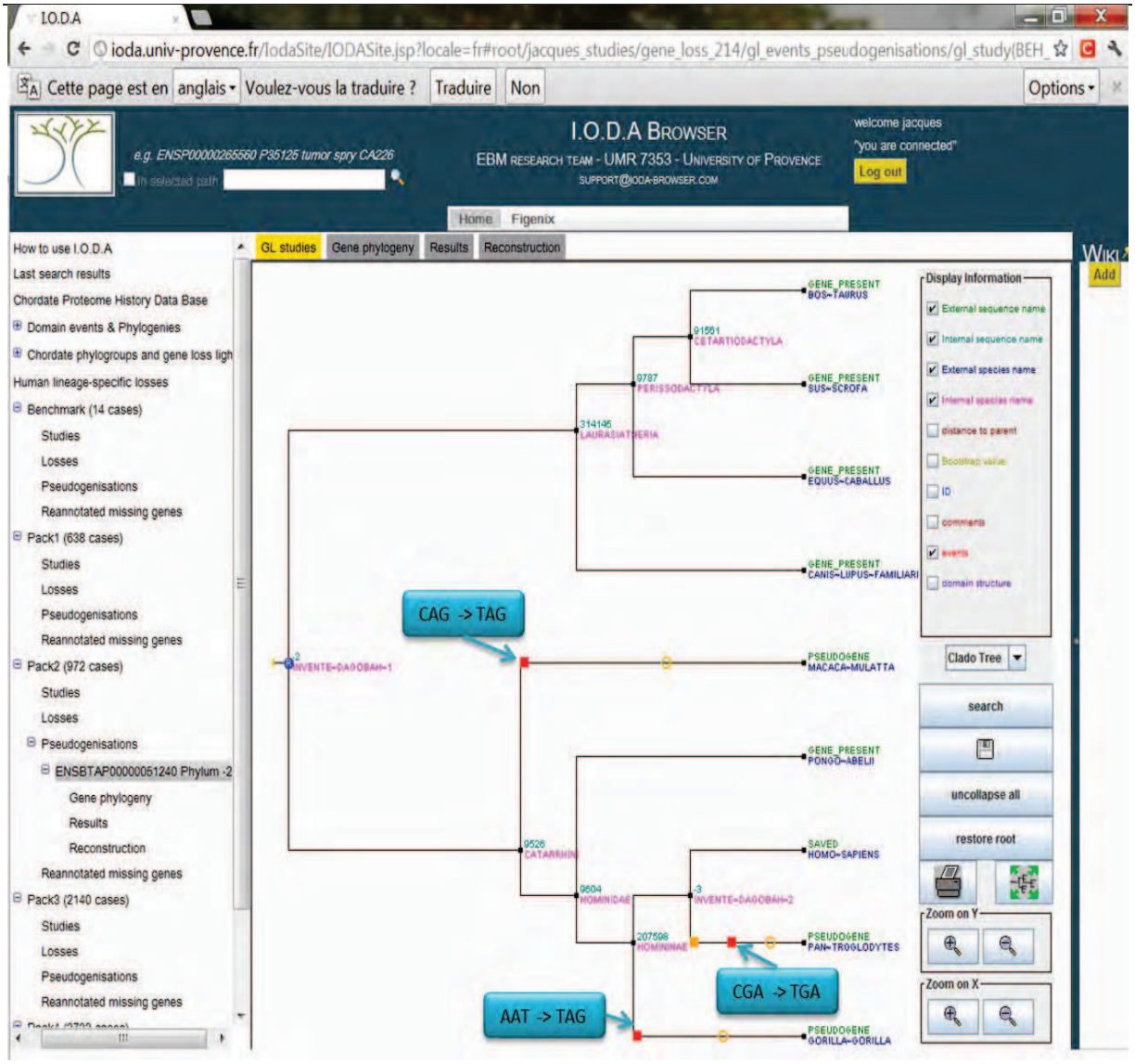


Illustration 33 : Page IODA synthétisant les résultats de GLADX pour l'étude ENSBTAP00000051240.
En rouge sont indiqués des événements d'apparition de codons stop. Il faut cliquer sur ces événements pour lire l'information associée. Dans cette illustration j'ai ajouté en bleu une partie de l'information liée aux événements d'apparition de codons stop. Les cercles jaunes présentent les macros événements de pseudogénisation. Le rectangle à la racine de la phylogénie montre quand le gène est apparu. Sur les feuilles de l'arbre sont indiqués les caractères de « présence », « absence » ou « sauvé » de la famille du gène étudié.

Article 3 - The chordate proteome history database

(*Evolutionary Bioinformatics*, 2012)

Evolutionary Bioinformatics

ORIGINAL RESEARCH

The Chordate Proteome History Database

Anthony Levasseur[1,2,*], Julien Paganini[3,*], Jacques Dainat[3], Julie D. Thompson[4], Olivier Poch[4], Pierre Pontarotti[3] and Philippe Gouret[3]

[1]INRA, UMR1163 Biotechnologie des Champignons Filamenteux, Aix Marseille Université, ESIL Polytech, 163 avenue de Luminy, CP 925, 13288 Marseille Cedex 09, France. [2]Aix Marseille Université, UMR1163 BCF, 163 avenue de Luminy, CP 925, 13288 Marseille Cedex 09, France. [3]UMR7353, Evolutionary Biology and Modeling, Aix Marseille Université, 3 place Victor-Hugo, 13331 Marseille, France. [4]Département de Biologie Structurale et Génomique, IGBMC (Institut de Génétique et de Biologie Moléculaire et Cellulaire), CNRS/INSERM/Université de Strasbourg, Illkirch, France.
*These authors contributed equally to this article. Corresponding author email: anthony.levasseur@univ-amu.fr

Abstract: The chordate proteome history database (http://ioda.univ-provence.fr) comprises some 20,000 evolutionary analyses of proteins from chordate species. Our main objective was to characterize and study the evolutionary histories of the chordate proteome, and in particular to detect genomic events and automatic functional searches. Firstly, phylogenetic analyses based on high quality multiple sequence alignments and a robust phylogenetic pipeline were performed for the whole protein and for each individual domain. Novel approaches were developed to identify orthologs/paralogs, and predict gene duplication/gain/loss events and the occurrence of new protein architectures (domain gains, losses and shuffling). These important genetic events were localized on the phylogenetic trees and on the genomic sequence. Secondly, the phylogenetic trees were enhanced by the creation of phylogroups, whereby groups of orthologous sequences created using OrthoMCL were corrected based on the phylogenetic trees; gene family size and gene gain/loss in a given lineage could be deduced from the phylogroups. For each ortholog group obtained from the phylogenetic or the phylogroup analysis, functional information and expression data can be retrieved. Database searches can be performed easily using biological objects: protein identifier, keyword or domain, but can also be based on events, eg, domain exchange events can be retrieved. To our knowledge, this is the first database that links group clustering, phylogeny and automatic functional searches along with the detection of important events occurring during genome evolution, such as the appearance of a new domain architecture.

Keywords: phylogenetic reconstruction, ortholog groups, protein architecture, functional inference, family size, genome evolution

Evolutionary Bioinformatics 2012:8 437–447

doi: 10.4137/EBO.S9186

This article is available from http://www.la-press.com.

Introduction

The genetic information encoded in the genome sequence contains the blueprint for an organism's potential development, physiology and activity. This information can only be fully comprehended in the light of the evolutionary events acting on the genome (duplication, gains and gene losses, nucleotide substitutions, genome recombination), reflected in changes in the sequence, structure and function of the gene products (nucleic acids and proteins) and ultimately in the organism's biological complexity.

The recent availability of the complete genome sequences of a large number of model organisms means we can now begin to unravel the mechanisms involved in the evolution of the genomes and their implications for the study of biological systems. At the same time, theoretical advances in biological information representation and management have revolutionized the way experimental information is collected, stored and exploited. Ontologies, such as Gene Ontology (GO) or Sequence Ontology (SO),[1] provide a formal representation of the data for automatic, high-throughput data parsing by computers. These ontologies are being exploited in new information management systems to allow large-scale data mining, pattern discovery and knowledge inference.

The vast number and complexity of the events shaping genomes means that a complete understanding of evolution at the genomic level is not currently feasible. At the lowest level, point mutations affect individual nucleotides. At a higher level, large chromosomal segments undergo duplication, lateral transfer, inversion, transposition, deletion and insertion. Ultimately, whole genomes are involved in processes of hybridization, polyploidization and endosymbiosis, often leading to rapid speciation (for a review see[2]).

Several databases dedicated to homologous gene families from vertebrates and microbial organisms have recently been developed for use in comparative genomics projects (for example,[3,4]). At present, the genomic context of a specific gene can be easily displayed using different user-friendly databases[5,6] and the evolutionary dynamics of gene clustering can be accurately inspected. In our study, the main objective is the characterization and study of the evolutionary histories of the chordate proteome, in particular the detection of genomic events and automatic functional searches. We make use of formal descriptions of biological data, together with recent developments concerning automated reliable protein sequence alignment and accurate phylogenetic reconstruction. These approaches have been combined in a multi-agent, expert system for the construction of evolutionary histories to facilitate the automatic definition of the important genetic events shaping a given protein. Here we present the computational strategies that we have developed and the first steps towards our final goal, in the form of a novel database: the chordate proteome history database. This database provides phylogenies for the chordate proteomes, reconstructed using a gene-based approach in which the same high quality phylogenetic pipeline is applied to each individual gene in a given genome. Genomic events, at the gene level or at the protein domain level, were detected automatically and localised on the gene phylogenies and on the genomic sequence, wherever possible. We focused on the orthologous relationships between sequences from 14 species: *Homo sapiens, Pan troglodyte, Pongo abelii, Macaca mulata, Canis lupus familiaris, Mus musculus, Rattus norvegicus, Monodelphis domestica, Gallus gallus, Xenopus tropicalis, Tetraodon nigriviridis, Oryzias lentipes, Amphioxus, Ciona intestinalis* and used these relationships for functional transfer wherever possible. We note that *Amphioxus* is used only in phylogroup analyses. The orthologous associations were obtained by clustering the protein sequences using OrthoMCL,[7] followed by correction based on a detailed phylogenetic analysis. All multiple alignments, phylogenetic trees, and tree-based functional predictions and genomic events affecting protein domain architecture can be easily accessed *via* a web-based user interface.

Materials and Methods

General features

The chordate proteome history database is deployed *via* a web application named Interface for Ontological Data Analysis (I.O.D.A), developed with the Google Web Toolkit technology, which uses Java and Javascript/AJAX languages. Each results menu on the left-hand side of the site can be annotated by a registered user *via* a wiki system on the right-hand side of the site. All users can read these wiki pages when they browse

the database. I.O.D.A is currently fully functional on the browsers Firefox and Google Chrome (download available on the I.O.D.A. homepage). For Macintosh users, I.O.D.A works correctly with MacOS 10.6.7 or higher and Java 1.6.0_24 or higher.

Data model

As its name suggests, I.O.D.A does not rely on a relational database model, but on a more accurate and flexible model structured by an ontology. The ontology used in the chordate database focuses on the specific evolutionary concepts manipulated in the laboratory. More specifically, we use an approach based on mathematical first-order logic named *Description Logic* (http://dl.kr.org/). The W3C-standardized OWL language (http://www.w3.org/TR/owl-ref/) is an XML representation of DL that we use to define our model. The database itself is formed by RDF triples persisting on an underlying relational database server (PostgreSQL: http://www.postgresql.org). The server is not accessed directly, but *via* a JAVA (http://www.java.com) API named Jena (http://jena.sourceforge.net/), which provides access to classes, instances and relationships. Wherever possible, we adopt the Relational Ontology terminology,[8] designed to standardize relationships in biological ontologies. The ontological database model scheme is described in.[9]

Phylogeny construction and event detection

All the phylogenetic trees present in the database were built automatically using the software platform FIGENIX[10,11] driven by the DAGOBAH expert system.[9] DAGOBAH is a multi-agent system in which specific agents have been developed for genetic event detection and verification. The phylogenetic trees were automatically analysed by a Java API: PhyloPattern.[12]

Identification of vertebrate homologs and construction of a multiple sequence alignment

The 19,837 human proteins defined by the Human Protein Initiative (http://expasy.org/sprot/hpi/) were used in this analysis. For each protein, database searches of the Swissprot and Ensembl databases[13] were performed using the BlastP program. Multiple alignments of complete sequences (MACS) were then constructed using the MAFFT program,[14] containing up to 500 full-length protein sequences. The quality of the MACS was then validated using the NorMD objective function, and unrelated sequences were excluded using the LEON program.[15]

Once a high quality MACS was obtained, the next step was to extract structural/functional information related to the protein family from the public databases. This was done using the in-house BIRD data retrieval system, and covered a wide range of information, from taxonomic data and functional descriptions (protein definition, EC number, GO, pFAM, Interpro) to sequence features, such as structural domains and active site residues. The retrieved data was integrated in the multiple alignment, together with a number of *ab initio* calculations (disordered regions, low-complexity segments and transmembrane helices), using the MACSIMS Information Management System.[16]

Construction of an accurate phylogenetic tree

Based on the main FIGENIX phylogeny pipeline, a new phylogeny pipeline was specifically developed to initiate phylogenetic studies from MACSIMS alignment files. In this pipeline, the alignment was intelligently cut to detect alignment areas associated with specific protein domains and repeats. For each domain, a phylogenetic tree was built and used for the study of domain architecture events.

In addition, a gene-level phylogeny was produced. All alignment areas associated with the domains in the protein query (the one that initiated the alignment) were concatenated and the resulting alignment was used for tree building. The gene phylogenies were used to study gene losses/gains and horizontal gene transfers and to compile duplication events and orthology and paralogy relationships.

Functional data

From all the homolog pages in I.O.D.A, the user can search functional data from: GO,[1] KEGG,[17] ArrayExpress,[18] String,[19] and QuickGO.[20] To do this, I.O.D.A converts Ensembl references to Uniprot references, which are all indexed in these databases.[21] To extract the functional data, these references are then sent to the web services associated with each of these databases. I.O.D.A presents the functional data, either directly on the web pages or through a link to these sites.

New protein domain architecture events, localization on the chordate species tree and verification at genome level

An apomorphic protein can be formed by any of five kinds of events detected by a dedicated DAGOBAH agent:

- *Gain*: one or more domains are gained at the beginning or end of the ancestral protein,
- *Loss*: one or more domains are lost at the beginning or end of the ancestral protein,
- *Insertion*: one or more domains are inserted between two domains of the ancestral protein,
- *Deletion*: one or more domains are deleted between two domains of the ancestral protein,
- *Shuffling*: one or more domains are exchanged at the beginning or end of the ancestral protein with a pendant protein.

The general strategy for domain event detection involved a nine-step process driven by the DAGOBAH multi-agents system:

1. Domain-annotated protein alignments built from a query protein are used to outsource phylogeny tree construction (domain trees and protein trees) to the FIGENIX pipeline.
2. The Mirkin parsimony algorithm[22] is used on each tree produced to infer ancestral domain architectures on internal nodes. Unfortunately, no efficient algorithm is currently available to infer the order of ancestral domain architectures.
3. The query's domain architecture is divided into a list of consecutive domain pairs. We note that two artificial domains (without any associated phylogenetic tree) are added at the tips, in order to study events occurring at the beginning and end of the protein. For example, for a protein with three domains A, B and C, our process studies each of the four pairs *[A-before, A], [A, B], [B, C]*, and *[C, C-after]*. For each pair, the phylogenetic trees produced at step 1 are used in the first steps of event detection (steps 4–6).
4. Ideally, a phylogenetic pattern consistent with the event should be found on each domain tree of a pair, which strengthens the event hypothesis. Nevertheless, events found only on one tree of the domain pair are considered as valid, but weaker, candidates. Two patterns are applied on the domain pair trees with our API: PhyloPattern,[12] one for deletion events and one for other events (Fig. 1). A pattern is a triple, ie, a well-supported ancestral node with two children: a plesiomorphic node corresponding to a domain architecture close to the ancestral one, an apomorphic node corresponding to the derived domain architecture.

   The pattern associated with a deletion event candidate is an ancestral node with the two domains of the pair and other domains (denoted DL) located between them, an apomorphic child node with the two domains of the pair whose subtree contains the query sequence, and a plesiomorphic child node with the two domains of the pair whose representative sequence contains the DL domain list.

   The pattern associated with other event candidates is an ancestral node with one of the two domains of the pair, ie, the one for which the tree receives the pattern, an apomorphic child node with the two domains of the pair, and a plesiomorphic child node with one of the two domains of the pair, ie, the same one as in the ancestral node.

   We note that when we refer to a node's domains, we mean the inferred architecture for an internal node or the known architecture for a leaf.
5. The choice of apomorphic and plesiomorphic representative sequences is very important for the

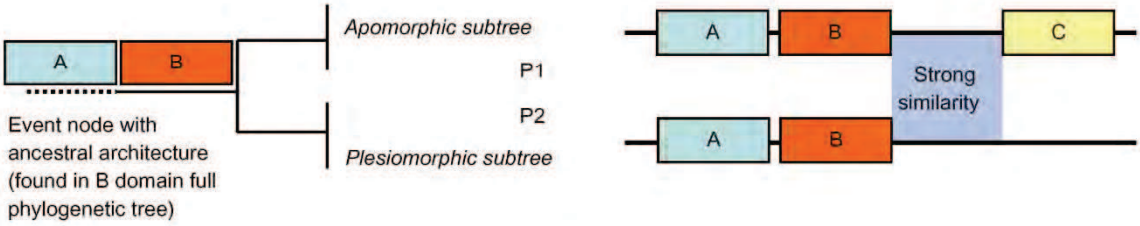


Figure 1. A virtual example of an event leading to a novel domain architecture.
Notes: Here a gain event is confirmed because the genome sequences between domains B and C on the apomorphic sequence and after domain B on the plesiomorphic sequence are strongly conserved. P1 and P2 indicate the two protein sequences chosen as representatives of the apomorphic and plesiomorphic subtrees.

subsequent steps. It will influence the reliability of event type determination (step 6), and also the reliability of event verification at the genomic level (step 8). Thus the process chooses the least remotely derived sequences, ie, the sequences with the domain architecture closest to the ancestral one and with the shortest branch to the ancestral node. These sequences are assumed to have accumulated fewer mutations and recombinations than the others. For the choice of apomorphic sequence, there is an exception to this last rule: when an event candidate's pattern is relevant for the two trees of the domain pair, we choose a sequence that belongs to the two apomorphic subtrees found in the two trees of the domain pair (for deletion candidates, we choose the query). When the criteria are not sufficient to choose between plesiomorphic sequences, the sequence closest to the apomorphic sequence species in the species tree is chosen.

6. The next step is to determine precisely the type of the event and all the domains involved in the transformation. This is done by computing the difference between the representative plesiomorphic and apomorphic sequence domain architectures. Sometimes several event types are similar. We will see at step 7 how we validate the event type. Table 1 summarizes all the different cases.
7. To produce "definitive" conclusions, the process confronts each individual event candidate (produced by the study of all the domain pairs in the query sequence architecture) with all the others, through an expert system, applying logical rules. We cannot give details of all the specific rules here, but their aim is to group some individual events or to remove some ambiguity, whenever possible. Non-grouped events are identically conserved. As an example of grouped event candidates, given an apomorphic domain architecture A B C D, the process could identify two insertion candidates by studying the A, B pair and the C, D pair, but they are probably linked to a single event, the insertion of B and C between A and D.

   In addition, we can see that the "shuffling/insertion" ambiguity in Table 1 could also be removed if, for example, the plesiomorphic architecture was A D when the process studied the A domain tree. In this case, the shuffling hypothesis is eliminated.
8. When event candidates are confirmed, the next step is to try to verify them at the genome level, by trying to find an alignment break position between two DNA segments, one associated with the representative apomorphic protein and the other with the representative plesiomorphic protein. DNA segments are extracted between concerned domains using Ensembl online access, and a Blast (tblastn) search is then performed on the DNA regions associated with the proteins, using the domain's amino acid segments as a query. Overlapping of Blast high similarity pairs is managed to extract the most significant area.

   BlastZ[23] is then used to align the two segments and to detect the alignment break position that should be the recombination point. More details of this process can be found in.[9]

   We note that for many events we find no such position, because the divergence date between the apomorphic and plesiomorphic species is often too far distant, and many other accumulated events have since masked the recombination event. When this information is found, it is supplied to the I.O.D.A user in an "Expert comment" field.

Table 1. In the studied query sequence, the domain pair A B or a pair with a virtual tip A A-after is shown in bold.

Event type	Plesiomorphic architecture part	Apomorphic architecture part	Event description
Gain	A A-after	**A B X**	Gain of domains B and X after domain A
Loss	A X	**A after**	Loss of domains X after domain A
Insertion	A X	**A B Y X**	Insertion of domains B and Y between A and X
Deletion	A X B	**A B**	Deletion of domains X between A and B
Shuffling	A X	**A B Y**	Replacement of domains X by domains B and Y
Shuffling/ insertion	A X	**A B Y X**	Replacement of domains X by domains B, Y and X or Insertion of domains B and Y between domains A and X

Notes: X and Y indicate lists of other domains. The event candidate is detected on the phylogenetic tree of domain A. When the tree of domain B is studied, a symmetric case is obtained.

In "ideal" studies, we identify two close recombination points on the common apomorphic sequence found on the two domain trees that show the event. If the same position is identified based on two different plesiomorphic sequences, then the event hypothesis is very strongly supported. However, these cases are very infrequent in the database.

9. We introduced this final step to detect, *a posteriori*, the most obvious artefacts. An artefact probability is supplied to the I.O.D.A user for all events. Our process detects two kind of artefacts:
   - Wrong propagation of domain architecture in the MACSIMS alignments (used to initiate our studies). The artefact detection agent re-predicts the apomorphic and plesiomorphic sequence domain architectures from the Pfam database to verify them.
   - Sequencing, assembling or gene prediction errors in the genomes used in this study. This agent is able to detect frequent artefacts resulting from the use of a gene isoform as an apomorphic one, although the plesiomorphic variant still exists, or reciprocally as a plesiomorphic one when the apomorphic variant exists.

Phylogroups and gene loss/gain study

The OrthoMCL algorithm was used to create groups of orthologous sequences from the same set of species as used for the phylogeny reconstruction. Phylogroups were created by clustering the groups using ortholog information obtained by the phylogenetic analyses. The "gene loss/gain" module is based on the phylogroup analysis. Gene gain and loss events were identified using the PARS algorithm.[22] As gene transfer in chordates is unlikely, a gene gain was assumed to occur only once. An event that occurred more than once was thus assumed to signal an artefact. The PARS algorithm minimizes the gain and loss events. For example, when an ortholog is frequently absent on a given tree, the algorithm predicts several gains. These cases should be considered as putative artefacts, possibly due to problems with the sequencing/assembly process.

Rules for event validation

If orthologs are recorded absent only on the leaves (except for the well-annotated genome: human and mouse), the expert system (a DAGOBAH agent) will not confirm the loss, which might be due to an annotation artefact or unfinished sequences. If the orthologs are recorded absent higher up the tree, or if all the orthologs are also absent in daughter branches, then DAGOBAH will valid the loss events.

External access

To facilitate access to all the data contained in the chordate database, I.O.D.A entries can be easily linked to and from external pages using the URL: http://ioda.univ-provence.fr/IodaSite/Site.jsp?id=XXXXX, where XXXXX can be replaced by any reference or keyword searchable on the I.O.D.A site (eg, P35125, which is a Uniprot reference). In this way, other databases focused on specific themes can include additional evolutionary information in their data.

Results and Discussion

The data in the chordate proteome history database are divided into two subprojects. The first subproject includes phylogenies, new architecture and duplication events. The second one is dedicated to chordate phylogroups analysis.

Phylogenetic data

As we were interested in the evolution of the human proteome, the scope of the phylogenetic analyses was limited to the chordate, focusing exclusively on well-annotated genome species. The phylogenetic analysis was assumed to be robust for small families, as all the homologous sequences should be present, forming reliable ortholog and paralog groups. Based on the phylogenetic tree, genetic events that affect different protein characteristics were investigated, including orthology/paralogy and domain architecture.

Functional data for the different orthologous groups were collected from the GO,[1] KEGG,[17] ArrayExpress,[18] STRING[19] databases, and links are provided to the Ensembl,[21] Uniprot and Pfam[24] databases and the NCBI taxonomy.[25]

Phylogroups

The phylogroup analysis is used as a filter and provides information about the size of the gene families, about potential gene loss in a given lineage, and finally about the appearance or gain of a novel gene family. Phylogroups are in fact OrthoMCL[7] ortholog groups that we overclusterized using orthology relationships

offered by the automatic analysis of phylogenetic trees produced. OrthoMCL clustering can lead to artefact groups made up of fast-evolving orthologs. We correct this artefact by clustering the groups using ortholog information obtained by the phylogenetic analyses. Thus several OrthoMCL groups can be integrated in a single group, denoted "phylogroup". The phylogroups can then be used to perform functional analyses as described for phylogenetic analyses.

In addition, the phylogroups are exploited in the evolutionary analyses for the detection of events such as gene loss and gene appearance. Gene appearance can be the result of various scenarios,[26] eg, (i) pseudo-appearance due to duplication followed by high rates of mutation, (ii) gene rearrangement leading to different domain architectures in the orthologs, (iii) horizontal gene transfer (only a few examples in the chordates) or (iv) de novo genes.

Our phylogroup approach and the associated gene loss and gain results offer a number of advantages over other published ortholog databases that use clustering: (i) the ortholog group is corrected by the phylogeny and (ii) we include expert rules to give greater confidence to the ortholog loss/gain events.

Database access and web interface features

Browsing and querying

The chordate proteome history database is publicly accessible at http://ioda.univ-provence.fr. The database is organized in two interconnected projects: (i) domain events and phylogenies and (ii) chordate phylogroups and gene loss/gain. The two subprojects are linked: the corresponding phylogroup can be accessed from a gene's phylogeny study page, and conversely, the domain events and phylogeny studies can be accessed from the phylogroup page.

The database can be browsed using the "search" window by entering various queries, eg, (i) the human protein name, using Ensembl or Uniprot identifiers, (ii) the Ensembl identifier for nonhuman species, (iii) key words, (iv) one or more domain names, (v) partial domain names or (vii) a combination of these key words. We note that numbered information and user guidelines are provided in wiki pages.

Phylogenies and domain event searches in the phylogeny subproject

The phylogenetic reconstructions for each gene are available and can be retrieved directly from queries. The phylogeny subproject can be searched for events leading to new domain architectures, ie, caused by the loss or gain of a domain or domain shuffling. Figure 2 shows an example of a search using the UniProtKB/Swiss-Prot accession number P35125 as a query. By clicking on the search window (entry page), two results pages are available; phylogenetic study and events studies (see Fig. 2).

The phylogenetic study results page includes the phylogenetic trees, the ortholog list with the functional

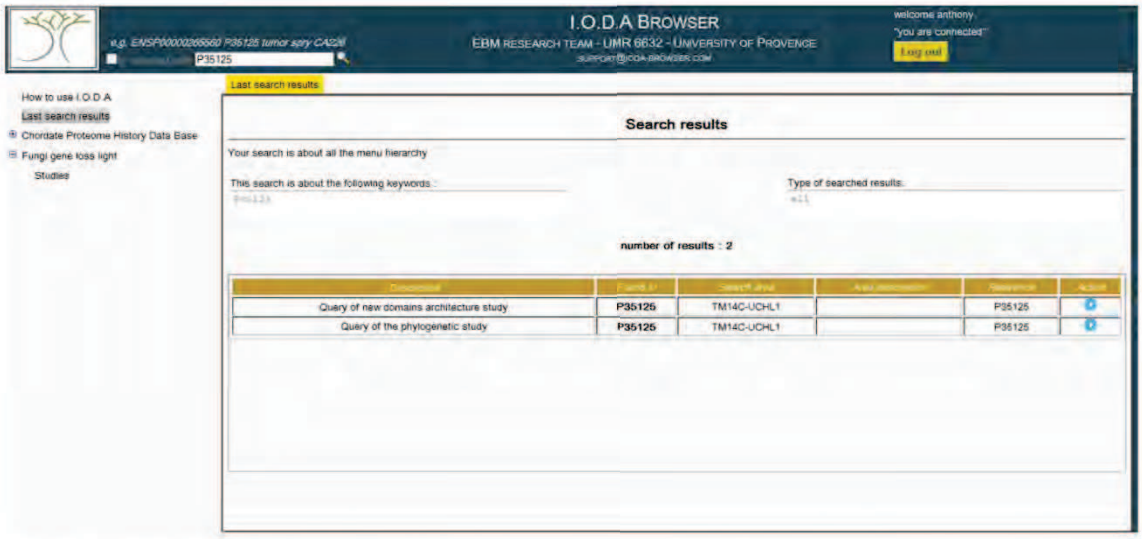


Figure 2. The chordate proteome history database entry page.
Notes: The entry page of a query protein (P35125) includes links to two available results: (i) phylogeny study and (ii) events study.

links, the paralog list and the list of homologs if the phylogenetic analysis results in some weakly supported nodes.

The events study results page includes links to each possible type of domain architecture evolution, ie, domain shuffling, domain insertion or deletion inside the sequence and domain loss or gain at the N- or C-terminus. For example, in the case of P35125, a domain shuffling event was detected (Fig. 3). By clicking on the "Shuffling events" tab and selecting a specific shuffling event, the user gains access to two information pages: "from Tree" and "Event pattern" (Fig. 4). The "from Tree" tab shows the phylogenetic tree used to deduce the event, together with the domain organization of the leaf sequences. In addition, the branch on which the event is assumed to occur is identified. The "Event pattern" tab provides more details about the domain organization of the apomorphic (derived) and the plesiomorphic (similar to the ancestral) representative sequences.

Phylogroup subproject

By clicking on the "Chordate phylogroups and gene loss/gain" and "Studies" menus, the study box shows the ortholog distribution on the different species under investigation. The "Group statistics" menu gives the user an overall view of the group distribution, the sequence number and the number of events, while the "Groups" menu gives the list of all ortholog groups. The tree box shows the species tree where the gene appears and when it is lost. The ortholog box provides the list of all the orthologs, and the functions of the ortholog sequences can be easily retrieved by clicking on the functional request button (see below: *Gulo* gene, for example).

A case study: the example of *Gulo* gene analysis from phylogroup data and phylogenies

The *Gulo* gene encodes an enzyme known to be involved in the pathway of vitamin C biosynthesis. This gene has been lost in primates,[27,28] resulting in the inability of primates to produce vitamin C. Any GULO protein found in our selected species could be used; for example, the user can type the mouse protein reference: ENSMUSP00000060912 in the search toolbar. Several results are available in the phylogroup or phylogeny analyses. Firstly, the phylogroup

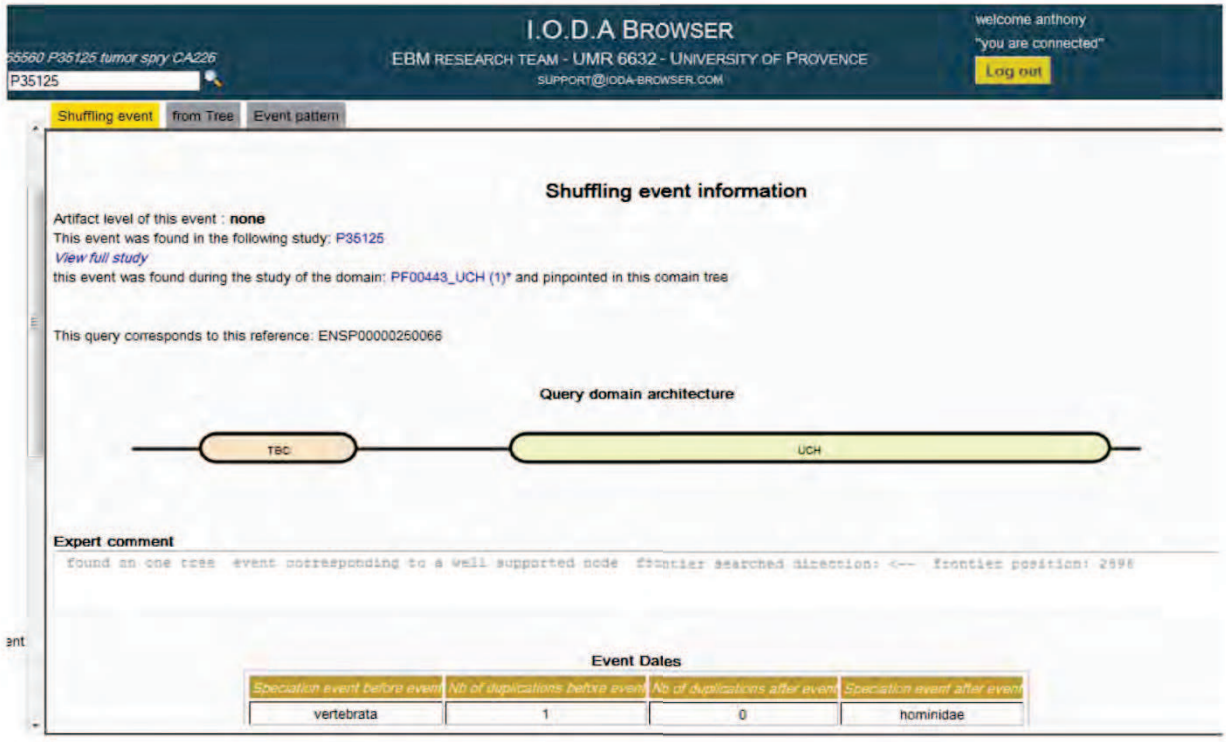


Figure 3. Domain structure organization.
Notes: The events study results page provides links to domain architecture evolution, eg, domain shuffling, domain insertion or deletion, domain loss or gain. In this example (P35125), a shuffling domain exchange was detected.

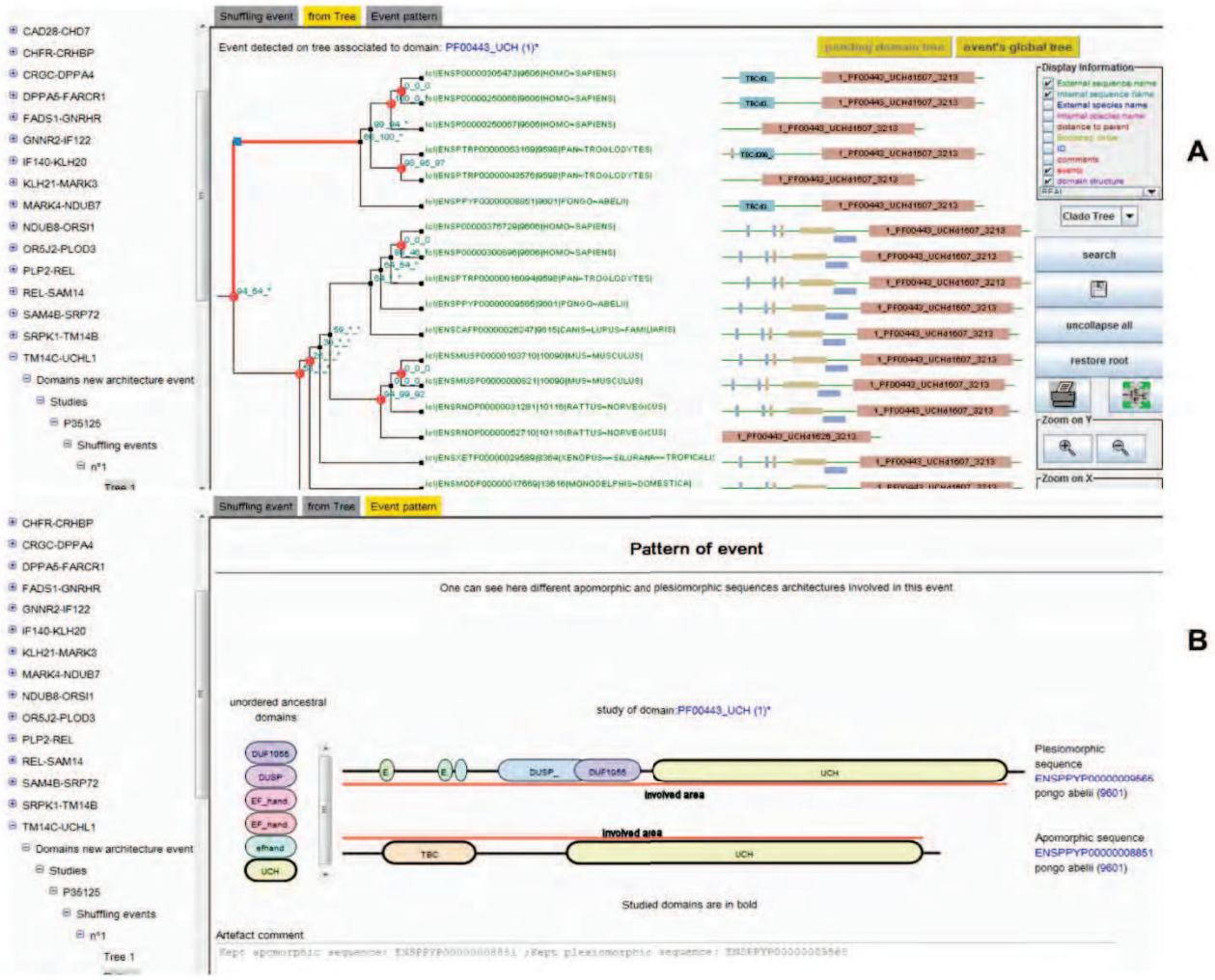


Figure 4. Tree and event pattern pages.
Notes: (**A**) The "from Tree" tab depicts the topology of the tree on the left-hand side and the domain organization for the leaves on the right-hand side. Gene duplications (red circles) and any detected domain architecture events (blue rectangles) are localized on the tree. Bootstrap values for each node are shown as a triplet corresponding to the three algorithms used to construct the tree. (**B**) The "Event pattern" tab shows the domain organization of the apomorphic (derived) and the plesiomorphic (similar to the ancestral) representative sequences.

results (OG_113469) indicate that the protein is found in 8 out of 14 species. In the *tree* tab, the loss events associated with this phylogroup are depicted (Fig. 5). This orthologous group existed before the last chordata ancestor, and subsequently two loss events in primates and actinopterygii ancestors occurred. The gene loss in teleosts has been observed previously[29] and this result agrees with the loss inferred in actynopterygii. These two loss events explain the six missing species and agree with the results already published. Secondly, the user can browse the phylogeny analysis in which ENSMUSP00000060912 is present (ie, Q15392: *All trees* tab) and examine the phylogeny based on the Q15392 entire protein by clicking on *Protein's best tree*. According to the phylogenetic tree, Q15392 is paralogous to Gulo. The Gulo ortholog group (paralogous to the Q15392 ortholog group) is found in this phylogenetic analysis and confirms that the gene is missing in both primates and actinopterygii.

New protein domain architecture: the example of shuffling events

A number of shuffling domain exchanges could be evidenced by using the database (as described above in the case of P35125). To summarize, 1943 shuffling domains were reported in the current version of the database. These 1943 shuffling events exclude all putative artefacts and could be assigned as relevant shuffling events with no ambiguity. In the field of new

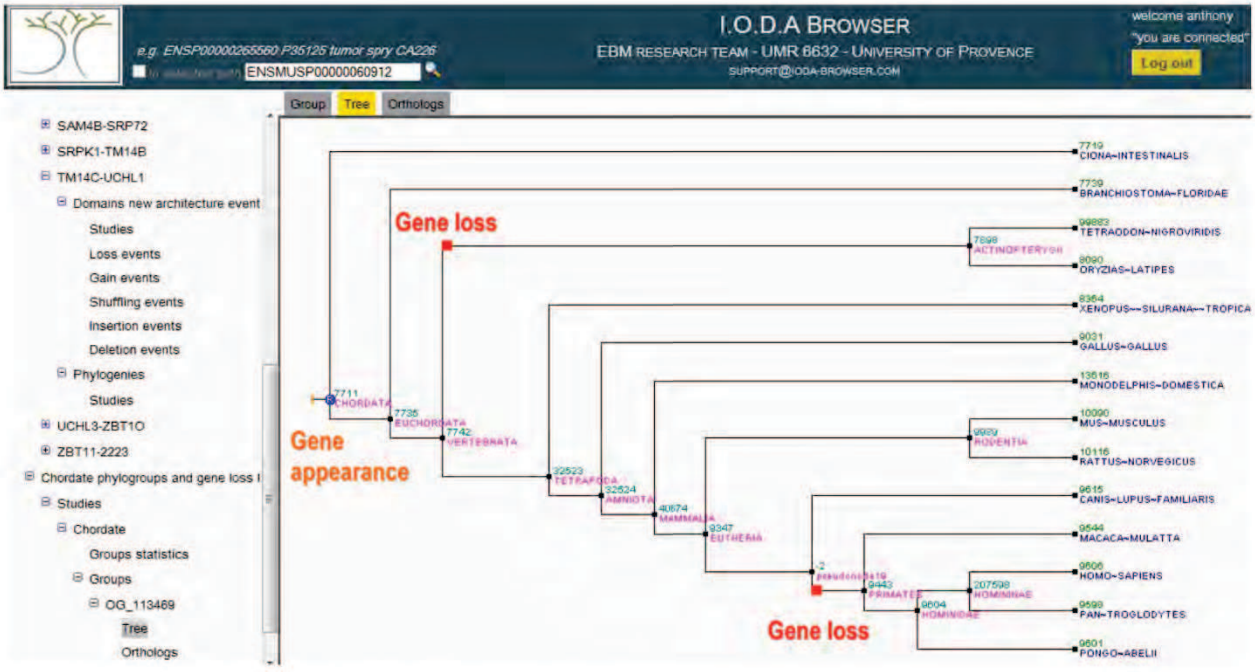


Figure 5. Detection of gene gain and loss in phylogroups.
Note: Example of *Gulo* gene analysis (ENSMUSP00000060912), the gene appearance and loss are directly depicted in the phylogenetic tree.

gene origination, these shuffling events are of prime importance for users, as the creation of new proteins/function could be carried out by bringing different domains together.[2]

Conclusions and Perspectives

In summary, the chordate proteome history database combines ortholog clustering, phylogeny and automatic functional link searches with automatic detection of important genomic events at the gene or protein domain levels. We are focusing on new enhancements for the medium-term including: (i) detection of other evolutionary events to achieve a more overall view of the genomic changes (eg, pseudogenization), (ii) introduction of other chordate genomes thanks to the current growing number of genomes sequenced and improved quality (structural and functional annotation) of the present genomes and (iii) development of other databases focusing on different kingdoms (eg, fungi) under the I.O.D.A. umbrella.

Author Contributions

Conceived and designed the experiments: AL, JP, PP, PG. Analysed the data: AL, JP, JD, JDT, OP, PP, PG. Wrote the first draft of the manuscript: AL, JDT, PP, PG. Contributed to the writing of the manuscript: AL, JDT, PP, PG. All authors reviewed and approved of the final manuscript.

Competing Interests

Author(s) disclose no potential conflicts of interest.

Funding

This research was supported by the ANR EvolHHuPro (ANR-07-BLAN-0054-01).

Disclosures and Ethics

As a requirement of publication author(s) have provided to the publisher signed confirmation of compliance with legal and ethical obligations including but not limited to the following: authorship and contributorship, conflicts of interest, privacy and confidentiality and (where applicable) protection of human and animal research subjects. The authors have read and confirmed their agreement with the ICMJE authorship and conflict of interest criteria. The authors have also confirmed that this article is unique and not under consideration or published in any other publication, and that they have permission from rights holders to reproduce any copyrighted material. Any disclosures are made in this section. The external blind peer reviewers report no conflicts of interest.

References

1. Harris MA, Clark J, Ireland A, et al. The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res*. 2009;32:D258–61.
2. Levasseur A, Pontarotti P. The role of duplications in the evolution of genomes highlights the need for evolutionary-based approaches in comparative genomics. *Biol Direct*. 2011;6:11.
3. Huerta-Cepas J, Capella-Gutierrez S, Pryszcz LP, et al. PhylomeDB v3.0. an expanding repository of genome-wide collections of trees, alignments and phylogeny-based orthology and paralogy predictions. *Nucleic Acids Res*. 2011;39:D556–60.
4. Penel S, Arigon AM, Dufayard JF, et al. Databases of homologous gene families for comparative genomics. *BMC Bioinformatics*. 2009;16:10.
5. Lopez MD, Samuelsson T. eGOB: eukaryotic Gene Order Browser. *Bioinformatics*. 2011;27:1150–1.
6. Wang D, Zhang Y, Fan Z, Liu G, Yu J. LGCbase: A comprehensive database for lineage-based co-regulated genes. *Evol Bioinform Online*. 2012;8:39–46.
7. Li L, Stoeckert CJ Jr, Roos DS. OrthoMCL. Identification of ortholog groups for eukaryotic genomes. *Genome Res*. 2003;13:2178–89.
8. Smith B, Ceusters W, Klagges B, et al. Relations in biomedical ontologies. *Genome Biol*. 2005;6:R46.
9. Gouret P, Paganini J, Dainat J, et al. Integration of evolutionary biology concepts for functional annotation and automation of complex research in evolution: the multi-agent software system DAGOBAH, evolutionary biology—concepts, biodiversity, macroevolution and genome evolution, Chap. 5, 2011 Springer; In press.
10. Gouret P, Danchin EGJ, Gilles A, Vitiello V, Balandraud N, Pontarotti P. FIGENIX: Intelligent automation of genomic annotation: expertise integration in a new software platform. *BMC Bioinformatics*. 2005;6:198.
11. Paganini J, Gouret P. Reliable phylogenetic trees building: a new web interface for FIGENIX. *Evolutionary Bioinformatics*. 2012;In press.
12. Gouret P, Thompson JD, Pontarotti P. PhyloPattern. regular expressions to identify complex patterns in phylogenetic trees. *BMC Bioinformatics*. 2009; 10:298.
13. Flicek P, Aken BL, Ballester B, et al. Ensembl's 10th year. *Nucleic Acids Research*. 2010;38:D557–62.
14. Katoh K, Misawa K, Kuma K, et al. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res*. 2002;30:3059–66.
15. Thompson JD, Prigent V, Poch O. LEON: multiple aLignment Evaluation Of Neighbours. *Nucleic Acids Res*. 2004;32:1298–307.
16. Thompson JD, Muller A, Waterhouse A, et al. MACSIMS: multiple alignment of complete sequences information management system. *BMC Bioinformatics*. 2006;7:318.
17. Kanehisa M, Goto S. KEGG: kyoto encyclopaedia of genes and genomes. *Nucleic Acids Res*. 2000;28:27–30.
18. Parkinson H, Sarkans U, Kolesnikov N, et al. ArrayExpress update—an archive of microarray and high-throughput sequencing-based functional genomics experiments. *Nucleic Acids Res*. 2011;39:D1002–4.
19. Szklarczyk D, Franceschini A, Kuhn M, et al. The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. *Nucleic Acids Res*. 2011;39:D561–8.
20. Binns D, Dimmer E, Huntley R, Barrell D, O'Donovan C, Apweiler R. QuickGO: a web-based tool for Gene Ontology searching. *Bioinformatics*. 2009;25:3045–6.
21. Hubbard TJ, Aken BL, Ayling S, et al. Ensembl. *Nucleic Acids Res*. 2009;37:D690–7.
22. Mirkin BG, Fenner TI, Galperin MY, Koonin EV. Algorithms for computing parsimonious evolutionary scenarios for genome evolution, the last universal common ancestor and dominance of horizontal gene transfer in the evolution of prokaryotes. *BMC Evol Biol*. 2003;3:2.
23. Altschul SF, Madden TL, Schaffer A, et al. Gapped BLAST and PSI-BLAST—a new generation of protein database search programs. *Nucleic Acids Res*. 1997;25:3389–402.
24. Finn RD, Mistry J, Tate J, et al. The Pfam protein families database. *Nucleic Acids Res*. 2010;Database Issue 38:D211–22.
25. Sayers EW, Barrett T, Benson DA, et al. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res*. 2009;37:D5–15.
26. Long M, Betran E, Thornton K, Wang W. The origin of new genes: glimpses from the young and old. *Nat Rev Genet*. 2003;4:865–75.
27. Nishikimi M, Fukuyama R, Minoshima S, Shimizu N, Yagi K. Cloning and chromosomal mapping of the human nonfunctional gene for L-gulono-gamma-lactone oxidase, the enzyme for L-ascorbic acid biosynthesis missing in man. *The Journal of biological chemistry*. 1994;269: 13685–8.
28. Ohta Y, Nishikimi M. Random nucleotid substitution in primate nonfunctionnal gene for L-gulono-gammalactone oxidase, the missing enzyme L-ascorbic acid biosynthesis. *Biochimica et Biophysica Acta* (BBA). 1999;1472:408–4141.
29. Maeland A, Waagbø R. Examination of the qualitative ability of some cold water marine teleost to synthesis ascorbic acid. *Comparative Biochemistry and Physiology*. 1998;121:249–55.

3 Analyses et discussions

3.1 Analyse des résultats d'OrthoMCL

Les analyses qui suivent portent sur les 22 558 GOs créés avec OrthoMCL. Elles ne prennent pas en compte les 336 GOs issus de l'étude comparative des GOs entres les différentes strates. Avec la parcimonie de Dollo, j'ai inféré les gains et les pertes des 22 558 GOs sur la lignée humaine afin d'en faire une analyse phylostratigraphique (Domazet-Lošo, Brajkovic, & Tautz, 2007). La synthèse du résultat se trouve dans l'illustration ci-dessous (Illustration 34). Parmi les 22 558 GOs, 16 657 possèdent une protéine humaine et 5 901 contiennent une perte putative de gène dans la lignée humaine (Illustration 31). J'ai placé les 5 901 GOs avec pertes détectées dans la lignée humaine, sur l'arbre des espèces, en fonction des dates de pertes de gènes respectives. J'ai également inféré les 22 558 GOs qui correspondent à 22 558 gains de gènes sur l'arbre en fonction de leurs dates d'apparition. La différence entre les gains et les pertes permet d'estimer la taille du protéome à chaque nœud ancestral. Les résultats montrent, par exemple, que l'ancêtre des Eucaryotes devait posséder plus de 4 000 gènes et celui des Eutélostomiens (*Euteleostomi*) plus de 15 000 gènes.

Par l'étude de la rétention des GOs je mets en évidence que cette rétention peut varier selon les lignées observées. Ainsi 22,4% des GOs présents chez l'ancêtre des Chordés semblent être perdus dans la lignée de *H. sapiens* tandis que 42% de ces GOs semblent perdus dans la lignée menant à *C. intestinalis*. Ce résultat est en accord avec les pertes importantes de gènes déjà observées chez *C. intestinalis* (Hughes & Friedman, 2005 ; Putnam *et al.*, 2008). Je trouve un résultat similaire avec une perte des GOs présents chez l'ancêtre du groupe *Fungi-Metazoan* à 21,8% dans la lignée de *H. sapiens*, et à 32,4% dans la lignée menant aux *Fungi*. Ces résultats sont en accord avec la forte perte de gènes observée par Koonin dans la lignée des *Fungi* (Koonin *et al.*, 2004). A noter également que les pertes de GOs présents chez l'ancêtre des Eutélostomiens sont plus élevées dans la lignée menant aux Clupéocéphales (*Clupeocephala*) (25,4%) que dans celle menant à *H. sapiens* (17%). Ces pertes plus importantes chez les Clupéocéphales ont précédemment été mises en évidence (Blomme *et al.*, 2006 ; Hughes & Friedman, 2004a, 2004b, 2005).

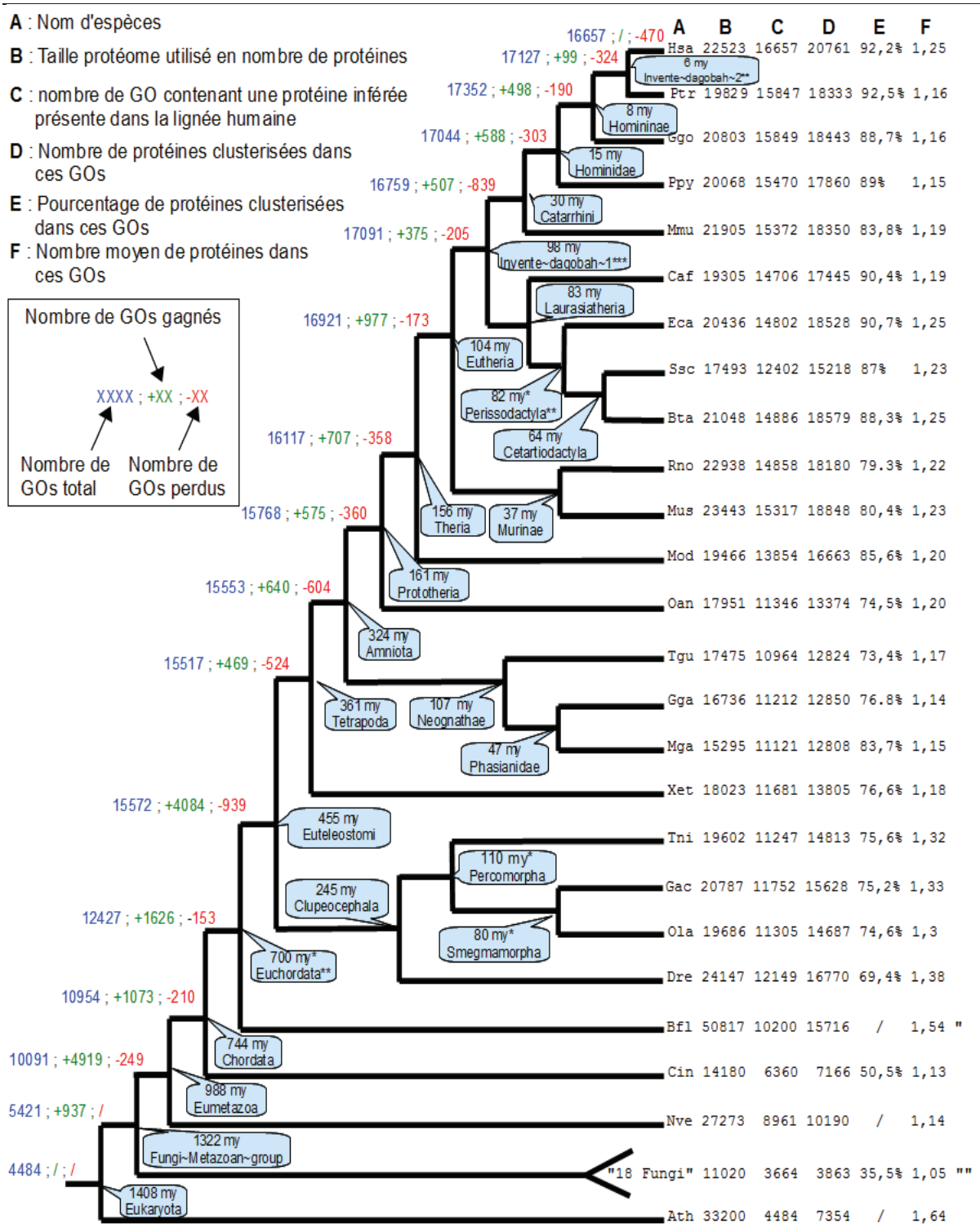


Illustration 34 : Inférence des 22 558 GOs sur l'arbre des espèces

Les protéomes ancestraux, les gains et les pertes sont inférés à partir des 22 558 GOs créés qui sont présents dans la lignée humaine. La topologie est celle utilisée actuellement comme référence. Pour la description de la topologie se référer à l'Illustration 27.

"Les isoformes n'ont pas été filtrés

""Le pan protéome de Fungi a été créé par OrthoMCL à partir de 18 protéomes d'espèces de Fungi. Ce pan protéome a été défini en prenant une protéine de référence par GO construit

J'ai également effectué l'inférence des gains et des pertes en utilisant des arbres d'espèces différents. Cela permet de voir l'influence du choix des topologies et démontre que la topologie choisie dans l'étude pour les nœuds incertains était cohérente. J'ai étudié la dynamique des GOs en inversant le placement de la lignée Laurasathériens (*Laurasatheria*) avec celle des Murinés (*Murinae*) (Annexe 11) et celle de *C. intestinalis* avec celle de *B. floridae* (Annexe 12). Les résultats obtenus sont moins parcimonieux au niveau des pertes que ceux obtenus avec la topologie choisie comme référence (Illustration 34). En effet les spéciations entre la lignée humaine et celle de *C. intestinalis* et de *B. floridae* sont relativement proches. Cela est encore plus vrai avec les spéciations entre la lignée humaine et celle des Laurasathériens et des Murinés. Ces changements de topologies, conduisent à constater que le nombre des pertes est beaucoup plus élevé dans ces courts laps de temps que dans la topologie de référence. En théorie, il est plus facile de gagner des gènes que d'en perdre. En se basant sur cette hypothèse, il semble que la topologie que j'ai choisie est meilleure, car elle présente moins de pertes dans une courte période d'évolution. Mais les deux inversions de topologie, produisent peu d'informations. En effet, le nombre de gains détectés dans le court temps d'évolution borné par ces spéciations, diminue et suit le chemin inverse des pertes. La difficulté à bien différencier ces lignées peut s'expliquer, dans le cas des Murinés et des Laurasathériens, par des possibilités d'hybridation entre ces lignées proches, et dans le cas *C. intestinalis* et *B. floridae* par des pertes massives survenues dans la lignée de *C. intestinalis*.

Dans la topologie phylogénétique choisie (Illustration 34), la progression des gains de familles de gènes représentées par les GOs semble constante dans le temps avec une légère augmentation au cours des derniers 30 millions d'années (Illustration 35). Cette légère augmentation peut s'expliquer par le nombre supérieur d'espèces représentées dans ce laps de temps. Et comme ces espèces ont un faible temps de divergence, la possibilité de retrouver des protéines communes est accrue.

La progression des pertes semble s'accélérer avec une forte inflexion positive au cours des derniers millions d'années d'évolution. L'analyse approfondie avec GLADX pourra confirmer ou infirmer cette accélération en étudiant les artefacts probables et les pertes avec plus de précision.

Comme cela a été déjà observé (Dehal & Boore, 2005), la duplication (1R) (Gu, Wang, & Gu, 2002 ; McLysaght, Hokamp, & Wolfe, 2002) ou les duplications (2R) (Larhammar, Lundin, &

Hallböök, 2002 ; Y. Wang & Gu, 2000) du génome chez l'ancêtre des Vertébrés ne semblent pas générer un accroissement du nombre de GOs observés.

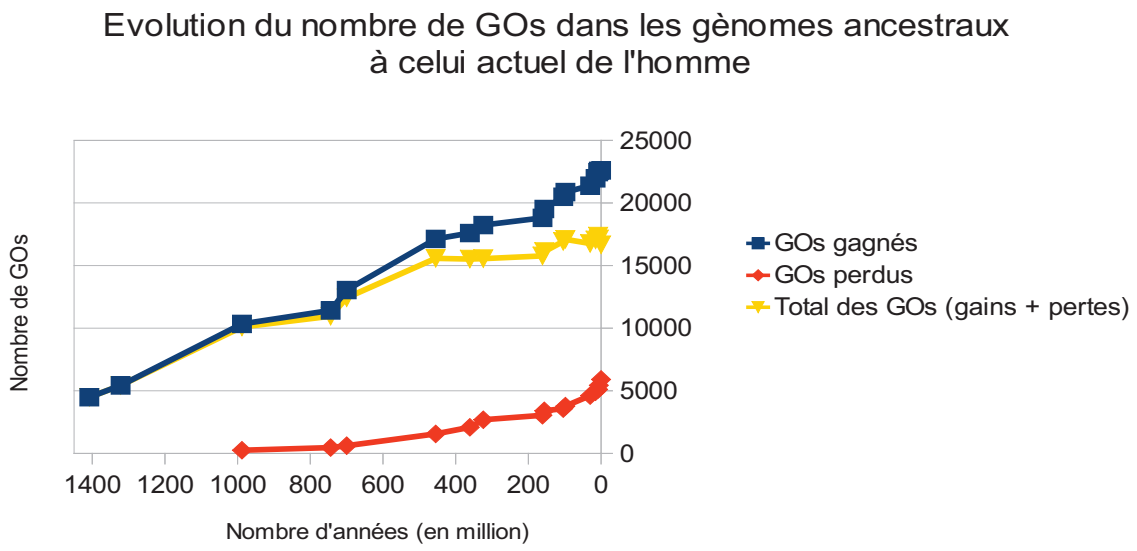


Illustration 35 : Progression des gains, des pertes et du total des GOs au cours du temps jusqu'à nos jours

3.2 Analyse des résultats de GLADX

L'analyse des GOs par OrthoMCL et la parcimonie de Dollo a permis de sélectionner 6 237 GOs contenant des pertes de gènes unitaires dans la lignée humaine (5 901 GOs issus des GOs formés avec l'ensemble des espèces, et 336 issues de l'étude comparative des GOs entres les différentes strates). Parmi les 6 237 GOs sélectionnés GLADX a initié 4 147 études avec la phylogénie de départ (1 553 + 2 594) tandis que cette phylogénie a échoué pour 2 090 GOs (1 179 + 911) soit 33,5% des études qui ne peuvent donc pas être poursuivies (Illustration 31).

3.2.1 Les études infructueuses de GLADX

D'une part, GLADX n'a pu établir une phylogénie de départ pour 2 090 GOs des 6 237 étudiés et n'a pas donc pu donner une idée des événements qui se sont produits dans ces GOs. Ces échecs pourraient provenir de GOs artefactuels prédits par OrthoMCL. De nombreux GOs, en effet, ne possèdent que deux ou trois protéines provenant d'espèces parfois peu proches phylogénétiquement. L'analyse des alignements de ces phylogénies infructueuses montre des alignements de mauvaise qualité contenant peu de séquences (moins d'une dizaine). Certaines phylogénies n'ont pu aboutir à cause des tests internes du pipeline de

phylogénie utilisé. En effet, lorsque le test de composition en acide aminé effectué avec l'outil *Tree-Puzzle*, ou celui de la congruence des domaines détectés dans les séquences est négatif, aucune phylogénie n'est créée.

D'autre part, parmi les 4 147 études initiées par une phylogénie, GLADX ne trouve ni perte ni présence de gènes chez l'homme dans 549 études (soit 19%). J'explique ce résultat comme suit :

- Dans 91% de ces 549 études, GLADX s'arrête après la première phylogénie. L'abandon de l'étude est causé par un paramètre fixé pour l'étude de la première phylogénie. En effet, j'ai fixé à 3 le nombre minimal de protéines orthologues requises pour continuer l'analyse.

- Dans les 9% restants, bien que les phylogénies contiennent des groupes d'orthologues égaux ou supérieurs à 3 gènes, la reconstruction des états ancestraux fait apparaître que le gène de référence est apparu dans une lignée différente de celle de *H. sapiens*. Par exemple, si GLADX trouve qu'une famille de gènes semble présente exclusivement chez les Clupéocéphales, il en conclut que cette famille de gènes est apparue après la séparation des Clupéocéphales et des Tétrapodes du LCA des Eutélostomiens, donc uniquement dans la lignée amenant aux Clupéocéphales. Dans ce cas de figure, cette famille de gènes est notée absente dans toutes les autres lignées.

3.2.2 Les études réussies de GLADX

3.2.2.1 Les gènes humains trouvés présents

Sur les 4 147 études réussies, GLADX trouve des gènes de *H. sapiens* présents dans le phylum le plus large (phylum le plus ancien où le gène apparaît dans la lignée humaine) dans 2 191 études (1 631 + 560 voir Illustration 31) soit 52,8% des GOs étudiés.

Ces 2 191 études ne présentent pas *a priori* de perte de gènes unitaires dans la lignée humaine. Pour s'en assurer, GLADX a approfondi la recherche sur ces 2191 études, et analysé les sous-familles de gènes apparues par duplications dans la lignée observée (celle de l'espèce de la protéine donnée en entrée de GLADX). GLADX met en évidence des pertes de gènes unitaires après duplications identifiées pour 309 études (102 + 110 + 97 voir Illustration 31) et montre que, pour les 1 882 autres (45% des études réussies), *H. sapiens* est toujours présent.

Sur les 4 147 GOs étudiés, 1 882 GOs qui semblaient comporter une perte de gènes unitaires

dans la lignée humaine lors de l'analyse des GOs, sont invalidés par les analyses phylogénétiques de GLADX. L'analyse manuelle des GOs faux positifs a permis de mettre en évidence de nombreux artefacts liés à la création des GOs. Par exemple dans le GO OG_115946 (Illustration 36) la protéine humaine est absente.

OG_115946:	9598|ENSPTRP00000035350	9601|ENSPPYP00000021257
9796|ENSECAP00000007956	10090|ENSMUSP00000019516	9258|ENSOANP00000020133
9103|ENSMGAP00000013787	99883|ENSTNIP00000016676	

Illustration 36 : Groupe de protéines obtenu dans le groupe d'orthologues OG_115946

J'ai initié l'étude de ce GO par GLADX en prenant pour référence la protéine ENSPTRP00000035350 et il apparaît que l'homme est présent dans la phylogénie de départ (Illustration 37).

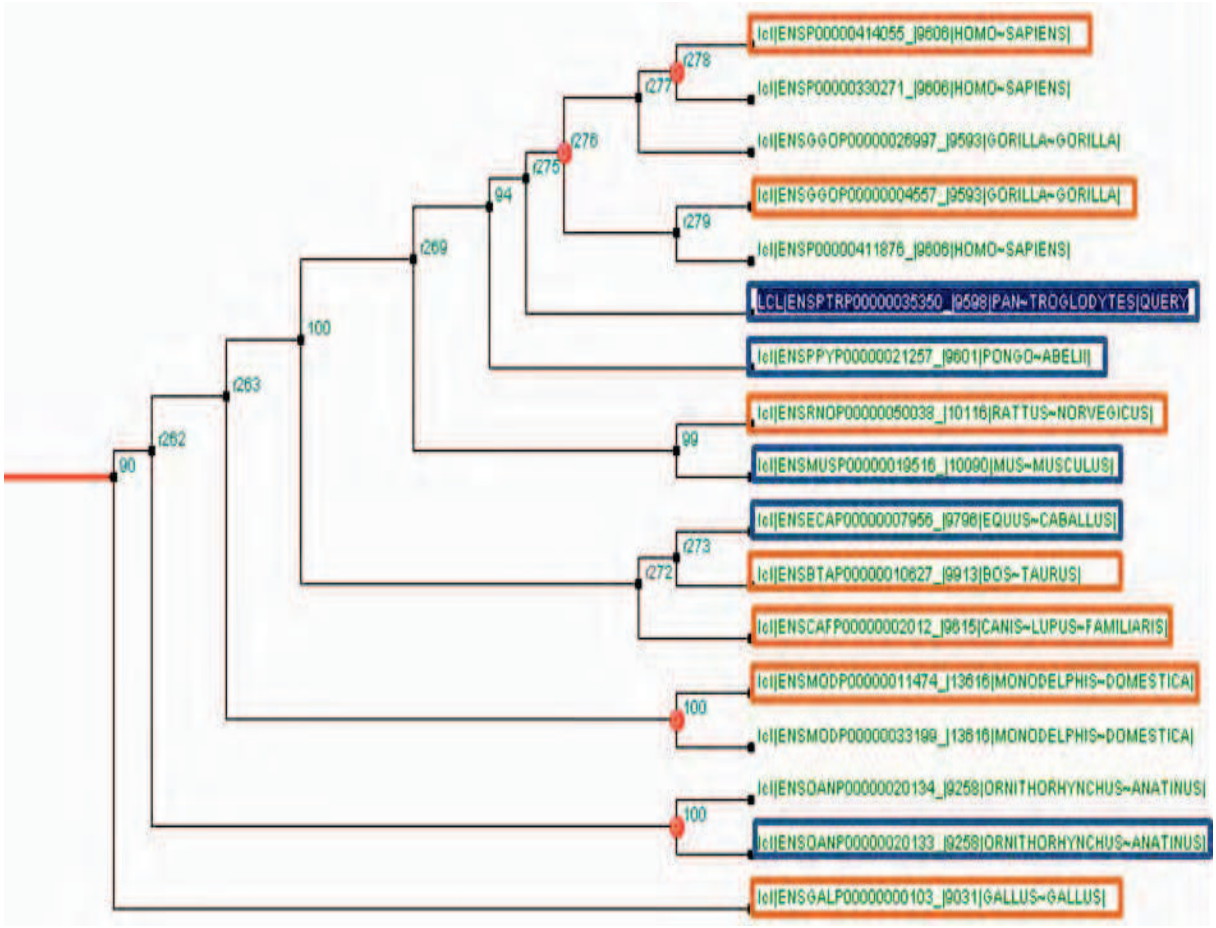


Illustration 37 : Phylogénie au départ d'une étude par GLADX

Les groupes d'orthologues OG_115946 (bleu) et OG_115946 (orange) créés par OrthoMCL sont mis en évidence.

A partir des orthologues à ENSPTRP00000035350 trouvés dans la phylogénie et qui ne sont

pas représentés dans le GO OG_115946, j'ai fait une recherche dans les GOs qui n'ont pas été sélectionnés pour une perte putative. Cette recherche a permis de constater que les espèces présentes dans le GO OG_113949 sont complémentaires des espèces présentes dans le GO OG_115946 (Illustration 38), et que tous deux forment, selon la phylogénie, un seul GO (Illustration 37).

```
OG_113949:          10116|ENSRNOP00000050038          9615|ENSCAFP00000002012
9606|ENSP00000414055      9593|ENSGGOP00000004557      9913|ENSBTAP00000010627
9544|ENSMMUP00000031870 13616|ENSMODP00000011474 9031|ENSGALP00000000103
7955|ENSDARP00000041145  7955|ENSDARP00000003329  69293|ENSGACP00000006206
8090|ENSORLP00000015272
```

Illustration 38 : Groupe de protéines obtenu dans le groupe d'orthologues OG_113949

En raison de la taille des protéines, ce GO a été divisé en deux GOs par OrthoMCL (OG_115946 + OG_113949). En effet dans le GO OG_115946 les protéines ont une longueur avoisinant 250 aa tandis que dans le GO OG_113949 les protéines clustérisées ont une longueur d'environ 500 aa. Certaines espèces comme *H. sapiens* possèdent les deux isoformes tandis que d'autres espèces présentent qu'un seul isoforme, court ou long. Pour la création des GOs j'ai utilisé des protéomes où une seule protéine par gène est conservée, en choisissant toujours l'isoforme le plus long.

3.2.2.2 **Les gènes humains nouvellement annotés**

Sur ces 4 147 études, GLADX a relevé chez l'homme 398 nouveaux gènes putatifs (373 + 25) qui semblent non annotés dans la base de données Ensembl V57 qu'il utilise. La détermination des gènes intacts est faite, soit au niveau protéique, en se basant sur la qualité des protéines prédites par rapport aux orthologues présents dans l'étude, soit au niveau nucléotidique, par l'étude des mutations par le scanner. Les séquences scannées au niveau nucléotidique sont considérées comme des gènes fonctionnels putatifs lorsqu'aucun codon stop prématuré n'est présent.

Pour vérifier l'efficacité de GLADX pour nouvellement annoter les gènes fonctionnels putatifs, j'ai pris au hasard un échantillon de 51 gènes annotés par GLADX sur lesquelles j'ai effectué une expertise manuelle (Annexe 13).

Une première vérification étudie les annotations disponibles dans les bases de données. En premier lieu, j'ai analysé les séquences annotées qui se trouvent dans la base de données Ensembl, aux positions où GLADX a sauvé un gène. Pour cela j'ai utilisé la base Ensembl

V57 (Mars 2010) utilisé par GLADX, et la base plus récente Ensembl V61 (Février 2011).

Dans la base de données V57 d'Ensembl je trouve 4 des 51 gènes annotés par GLADX déjà annotés, mais, les gènes ne se recouvrent pas entièrement. Malgré la vérification de GLADX de l'existence préalable de ses prédictions de gènes sur Ensembl V57, ces 4 gènes semblent avoir été manqués. L'erreur est imputable à l'overlap des gènes sauvés et ceux déjà annotés sur Ensembl, ou au score de similarité avec la protéine prédite, qui ne suffit pas à GLADX pour déterminer que ces gènes sont semblables. Pour éviter cet écueil, il est possible de modifier les seuils appliqués dans GLADX pour permettre une reconnaissance plus souple, au risque d'engendrer des faux positifs. Enfin, les quatre gènes présents dans cette base de données étaient absents des phylogénies de départ des études de GLADX. La raison de leurs absences peut être comprise en étudiant les différentes étapes du processus de phylogénie. En définitive, les 47 gènes « sauvés », non annotés dans Ensembl V57, correspondent dans la majorité des cas à des séquences annotées comme pseudogènes.

Dans la base de données plus récente V61, je trouve que 20 des 47 gènes annotés par GLADX sont désormais annotés comme codants.

J'ai réalisé une seconde vérification en faisant un BLAST (BLASTN et/ou TBLASTN) des séquences des 47 gènes « sauvés », sur les bases de données d'ARNm et d'ESTs du NCBI. Par cette méthode, j'ai trouvé dans certains cas, qu'il existe une transcription des séquences ainsi qu'une probable traduction en protéine. Cette vérification montre que, parmi ces 47 gènes sauvés par GLADX, 33 semblent coder une protéine, 7 sont au moins transcrits et 7 ne sont pas confirmés par les informations contenues dans les bases de données.

J'ai réalisé une troisième vérification, en estimant les pressions de sélection auxquelles les séquences sauvées sont soumises. Pour cela j'ai étudié le ratio du taux de mutations non synonymes sur le taux de mutations synonymes (Annexe 14). Un ratio neutre proche de 1 définit une évolution sous la neutralité qui peut toucher les séquences non codantes telles que les pseudogènes. Un ratio <1 ou >1 permet de déduire une sélection purifiante ou positive qui touche essentiellement les séquences codantes. Par cette étude, j'ai mis en évidence seulement 3 séquences qui semblent évoluer proches de la neutralité. Par exemple, dans l'étude qui a comme référence la séquence ENSPTRP00000052322, la séquence génique putative annotée chez l'homme par GLADX a un ratio de 0,97. Ce ratio indique donc une pseudogénisation. La séquence semble encore bien conservée car GLADX n'a trouvé qu'une mutation dans un site d'épissage. Cette mutation peut être à l'origine de la pseudogénisation, qui semble *a priori*

très récente.

L'étude de l'échantillon de 51 gènes sauvés par GLADX démontre la bonne efficacité de GLADX pour annoter les nouveaux gènes. Néanmoins, dans certains cas, un approfondissement expérimental est nécessaire pour vraiment trancher et savoir si la séquence est traduite en protéine ou non.

3.2.2.3 Les pertes de gènes unitaires

Au total j'ai trouvé 1 318 événements de pertes de gènes dont 311 possèdent encore leurs pseudogènes unitaires dans le génome humain (Illustration 31 et Tableau 2). Le tableau suivant indique à quelles périodes les pertes détectées ont eu lieu (Tableau 4).

Taxid ancêtres pertes observées	6072	7711	7735	117571	32523	32524	9254	32525	9347	-2	9526	9604	207598	-3	9606
Fenêtre de temps des pertes (my)	1322 988	988 744	744 700	700 455	455 361	361 324	324 161	161 156	156 104	104 98	98 30	30 15	15 8	8 6	6 0
Pseudogènes	0	0	0	0	0	0	6	3	7	3	82	43	32	39	96
Pertes (sans pseudogènes)	95	14	17	126	58	174	68	36	48	9	144	34	47	49	88
Total (Pseudogènes + Pertes)	95	14	17	126	58	174	74	39	55	12	226	77	79	88	184
Total cumulé	95	109	126	252	310	484	558	597	652	664	890	967	1046	1134	1318

Tableau 4 : Pertes détectées (pseudogènes et sans signal) en fonction des ancêtres observés

La ligne « total cumulé » correspond à l'accumulation du total des pertes de gènes observés au cours de l'évolution (95 + 14 = 109 ; 95 + 14 + 17 = 126 ; etc.). Cette ligne est reprise dans l'illustration suivante (Illustration 39) pour souligner l'évolution du nombre de pertes détectées au cours du temps évolutif analysé. L'illustration révèle une accélération du nombre de pertes détectées au cours du temps. Cette accélération peut refléter une réalité, mais peut également résulter d'une sur-prédiction de pertes détectées. Malgré une sur-prédiction possible de pertes détectées, la tendance va dans le sens des observations faites par l'analyse des pertes au travers des GOs. L'épaulement des pertes vers 300 millions d'années correspond à la transition Tétrapodes/Amniotes.

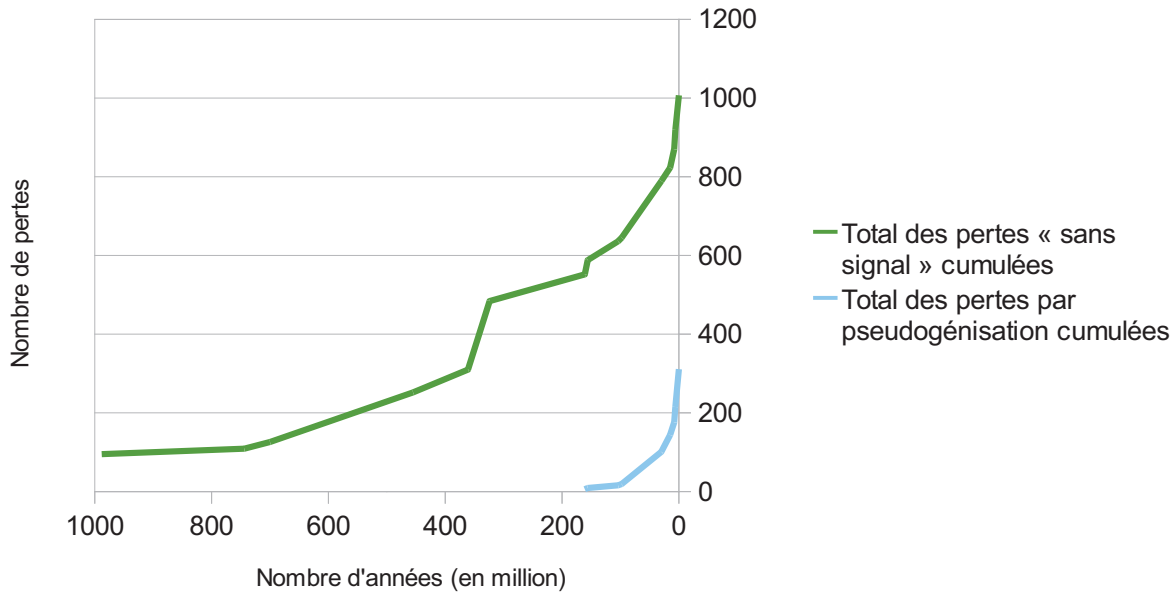


Illustration 39 : Nombres de pertes (pseudogènes et sans signal) cumulées au cours du temps

3.2.2.3.1 Analyse des pseudogènes unitaires

3.2.2.3.1.1 Comparaison avec des études publiées

J'ai abordé l'analyse des pseudogènes unitaires par la comparaison du résultat (1 318 pertes de gènes unitaires de la lignée humaine détectées par GLADX) avec les résultats présentés par les deux études considérées comme les plus complètes sur le sujet (Z. D. Zhang *et al.*, 2010 ; Zhu *et al.*, 2007) (Annexe 15). En ne tenant pas compte des pertes touchant les récepteurs olfactifs, elles identifient au total 107 pertes de gènes unitaires, dont 14 communes aux deux études. Ces 107 pertes de gènes unitaires sont caractérisées par la présence de pseudogènes unitaires.

Parmi les 1 318 pertes de gènes unitaires détectées par GLADX, je retrouve 47 des 107 pertes décrites dans ces deux études. L'étude au hasard d'un échantillon de 1 318 gènes d'un génome d'environ 22 000 gènes, aurait conduit à détecter seulement 6 des 107 pertes. Les 47 pertes se répartissent en 18 des 45 pertes décrites dans l'étude de Zhu et 42 des 76 décrites dans l'étude de Zhang (sachant que, sur les 14 pertes communes aux deux études, j'en retrouve 13). Les études publiées par ces deux auteurs se basent sur la présence de pseudogènes. Il est important de remarquer que sur les 47 pertes communes décrites par GLADX, 7 sont sans pseudogène. L'annotation manuelle effectuée dans leurs études pour détecter la présence de pseudogènes fait preuve d'une bonne précision. En effet, ces auteurs détectent la présence de pseudogènes

même lorsque manquent plus de 80% du signal (Zhu *et al.*, 2007).

GLADX utilise des phylogénies pour reconnaître les séquences orthologues pseudogénisées. Les pseudogènes doivent donc encore posséder une séquence de bonne qualité qui permette la création d'une phylogénie. En effet, pour 5 des 7 pertes sans pseudogène, GLADX semble détecter la séquence orthologue qui correspond au pseudogène, mais la phylogénie échoue. Il est possible que de nombreuses mutations au sein des pseudogènes empêchent la réalisation des phylogénies. Des pseudogènes peuvent donc ne pas être détectés par GLADX. Il semble que, pour 10% des études, GLADX trouve correctement des pertes de gènes unitaires mais n'arrive pas à trouver les pseudogènes présents, la phylogénie jouant le rôle de facteur limitant.

Parmi les 107 pseudogènes unitaires humains décrits par les deux équipes citées, 60 ne sont pas retrouvés dans les résultats de pertes de gènes unitaires produits par GLADX. Ils se subdivisent en :

- 32 orthologues humains présents directement dans l'analyse des GOs créés par clustérisation et qui contiennent le gène utilisé comme référence dans leurs études.

- 12 orthologues humains absents dans les GOs mais sauvés par GLADX. Sur ces 12 orthologues retrouvés à l'aide de GLADX, 4 sont nouveaux, les autres ont déjà été décrits dans la base de données utilisée (Ensembl V57).

- 16 pour lesquels je n'ai pas de résultats. Toutefois, pour 8 des 16 cas, le gène pris en référence dans leurs études est absent des GOs. Parmi eux, 2 sont absents à cause des changements d'annotations dans les différentes versions des bases de données utilisées (élimination de la séquence), et 6 sont éliminés lors de la création des GOs par OrthoMCL. Sur les 16 cas sans résultats, je trouve dans 2 cas le gène pris en référence pour leur étude dans des GOs, mais ces GOs ne concernent pas la lignée humaine. Dans 6 cas les gènes de références sont bien dans des GOs analysés par GLADX, mais dans 5 cas aucune phylogénie au début des études n'est obtenue, et dans 1 cas la phylogénie existe mais GLADX identifie un groupe d'orthologues trop petit pour lancer une analyse (< 3 protéines).

Ces résultats conduisent à considérer que 46 des 107 gènes identifiés par ces auteurs (Z. D. Zhang *et al.*, 2010 ; Zhu *et al.*, 2007) semblent être des faux positifs. Parmi ces 46 pertes, 34 (32 + 2) sont infirmées par l'analyse de GOs, et 12 par l'analyse effectuée par GLADX. Les 14 cas restants demandent à être approfondis pour comprendre l'impossibilité de les analyser par la stratégie utilisée.

3.2.2.3.1.2 **Vérification d'un échantillon de pseudogènes unitaires**

Parmi les 311 pseudogènes détectés par GLADX (Illustration 31), j'ai étudié en détail un échantillon aléatoire de 41 pseudogènes. J'ai recherché, pour ces pseudogènes, des ESTs dans la base de données UniGene (ressource NCBI). Sur l'échantillon de 41 pseudogènes, je ne retrouve aucun ESTs dans 18 pseudogènes, dont 6 sont décrits dans la littérature (Annexe 16). Pour 23 pseudogènes je trouve des ESTs (partiels ou complets), dont 11 pseudogènes sont décrits dans la littérature (Annexe 17). J'ai utilisé les ESTs pour détecter des mutations délétères et étayer les résultats obtenus par GLADX. Pour 17 pseudogènes, les ESTs permettent de confirmer des mutations observées. Pour 2 pseudogènes les ESTs partiels ne me permettent pas d'observer les mutations détectées par GLADX. Et 4 pseudogènes détectés par GLADX semblent des faux positifs. L'analyse de ces 4 faux positifs putatifs montre que 2 sont dus à des erreurs de séquences dans la base de données Ensembl. Un possède un ARNm complet mais n'est pas annoté comme codant dans les bases de données Ensembl et du NCBI. Dans un seul cas (ENSMUSP00000074553) GLADX est incapable d'annoter le gène comme intact. Cette erreur est due aux limites de la méthode et des outils implémentés dans GLADX. Ces limites se retrouvent dans la reconstruction des séquences ancestrales. L'étude manuelle d'un échantillon de 41 pseudogènes trouvés par GLADX en utilisant les séquences ESTs disponibles dans les bases de données, confirme la rareté des faux positifs dans la détection des pseudogènes par GLADX.

Dans la base de données Ensembl V57, la présence d'un pseudogène est souvent annotée aux positions dans le génome où GLADX découvre des pseudogènes. J'ai constaté que seuls des pseudogènes déjà décrits dans des études de pertes de gènes unitaires, sont annotés « pseudogènes unitaires » dans Ensembl. Pour les pseudogènes unitaires trouvés par GLADX, non décrits dans les publications, la base de données Ensembl peut les contenir et les annoter « pseudogènes », « pseudogènes processés », etc. Les analyses de GLADX permettent de spécifier les types de pseudogènes et de conclure que ce sont des pseudogènes unitaires.

3.2.2.3.2 **Analyse des pertes de gènes unitaires détectées sans pseudogènes**

Sur 1 318 pertes de gènes unitaires, GLADX en trouve 1 007 pour lesquelles il ne peut reconnaître de séquences orthologues pseudogénisées. Ce résultat est logique si l'on tient compte de la grande période évolutive observée. En effet, les pseudogènes sous évolution neutre, tendent à disparaître rapidement. De manière générale, lorsque GLADX détecte un événement de perte, plus le nombre d'espèces monophylétiques sans séquences orthologues retrouvées est élevé, plus l'événement de perte est crédible.

3.2.2.3.2.1 **Analyse de pertes anciennes**

Pour cette analyse, j'ai défini les pertes anciennes de gènes unitaires comme les pertes survenues avant le dernier ancêtre commun des Catarrhiniens c'est-à-dire correspondant à des pertes survenues il y a plus de 30 millions d'années. Ces pertes anciennes de gènes unitaires, dues à leur âge, sont théoriquement moins susceptibles d'être liées à la présence d'un pseudogène. Parmi les 1 007 pertes de gènes unitaires de la lignée humaine sans séquences orthologues pseudogénisées, 789 sont des pertes anciennes. Parmi ces 789 pertes, j'ai examiné les résultats d'un échantillon de 36 pertes pris au hasard (Annexe 18).

Pour 16 pertes, GLADX ne trouve aucun signal du gène de référence par blast dans le génome humain. Pour les 20 autres, GLADX a cherché sans succès un orthologue parmi les séquences retrouvées par blast à l'aide de la phylogénie. Mais dans l'étude de 7 des 20 pertes, une ou plusieurs phylogénies ont échoué, ce qui signifie que des séquences orthologues chez l'homme ont pu être manquées par GLADX. Pour ces 7 études, j'ai vérifié manuellement si des séquences orthologues ont pu être manquées à cause des phylogénies échouées. Pour 2 études j'ai trouvé un pseudogène orthologue putatif raté par les phylogénies infructueuses. L'échec des phylogénies peut s'expliquer par la mauvaise qualité des séquences de ces pseudogènes, liée à leurs pseudogénisations.

Pour compléter l'analyse, j'ai mené une recherche d'ESTs. La recherche d'ESTs s'est faite en utilisant l'orthologue présent de l'espèce phylogénétiquement la plus proche de l'homme. Un seul EST orthologue putatif a été mis en évidence chez l'homme. La séquence de cet EST présente dans le génome humain n'a pas été détectée par le TBLASTN utilisé par GLADX. Ce qui semble indiquer que cette séquence est vraiment abimée par le processus de pseudogénisation et que l'algorithme utilisé pour la recherche d'ESTs est plus sensible que celui implémenté dans GLADX.

En résumé, l'analyse manuelle de l'échantillon de 36 pertes anciennes de gènes unitaires parmi les 789 données par GLADX, confirme les événements de pertes de gènes unitaires. Néanmoins, 3 pseudogènes semblent ne pas avoir été pris en compte.

3.2.2.3.2.2 **Analyse de pertes récentes**

Parmi les 1 007 gènes unitaires sans séquences orthologues pseudogénisées, 218 sont des pertes récentes détectées dans la lignée humaine. Ces 218 pertes récentes sont survenues lors des 30 derniers millions d'année, et GLADX ne détecte aucun pseudogène. Ce résultat est surprenant car une étude montre que les pertes arrivent quasi exclusivement par

pseudogénisation (Costello *et al.*, 2008). A raison d'un taux de divergence de 1% à 2% par million d'années sous évolution neutre, les pseudogènes sont théoriquement encore détectables dans les génomes.

Ces pertes dites récentes sont-elles réelles ? Sont-elles associées à des pseudogènes non pris en compte par l'étude ?

Nous avons vu précédemment que l'étude comparative entre mes résultats et ceux de Zhu et de Zhang, a montré que près de 10% des pertes récentes sans signal pouvaient avoir un pseudogène présent dans le génome étudié mais non identifié par GLADX. Je suppose donc qu'il existe parmi les pertes récentes, des cas de pseudogènes ratés par GLADX.

Pour essayer d'en savoir plus, j'ai étudié un échantillon de 20 pertes pris au hasard parmi les 218 pertes récentes sans signal de pseudogènes (Annexe 19). Les résultats sont les suivants :

- Pour la majorité des études de ces 20 pertes, le nombre maximal de hits de blast analysé par GLADX est atteint, sans reconnaissance d'une séquence orthologue. Tous les hits étudiés sont identifiés comme paralogues. Il est légitime de penser que la limitation du nombre de hits analysés influe sur le résultat. En effet, GLADX arrête la recherche d'orthologues parmi les résultats du blast, lorsque la limite du nombre de hits à étudier fixée par l'utilisateur est atteinte. Il est alors possible d'envisager qu'une séquence pseudogène orthologue soit présente dans un des hits retrouvés par le blast, mais qu'un mauvais score de ce hit le place loin dans le classement des hits à analyser. GLADX passe alors à côté de la présence de pseudogènes orthologues se trouvant dans les hits qui ne sont pas analysés. Cette hypothèse a une plus grande probabilité d'être vraie dans l'étude de familles multigéniques, car le nombre de paralogues proches est élevé. Sur 20 pertes récentes étudiées, 12 concernent des gènes codants pour des récepteurs olfactifs, connus pour être des familles multigéniques. Il est alors fort probable que l'on puisse ne pas prendre en compte leurs pseudogènes. Pour affirmer avec certitude l'absence des pseudogènes il est nécessaire de réaliser une expertise des hits de blast. A défaut, l'absence de pseudogènes ne peut être assurée.
- Pour 5 études de pertes, je relève un ou plusieurs échecs de phylogénies nécessaires pour tester l'orthologie de l'ensemble des hits de blast. L'expertise montre que pour 4 études, GLADX semble avoir raté une séquence orthologue existante qui pourrait être un pseudogène.
- Dans 4 études de pertes, il semble que les pertes soient plus anciennes que ce que

décrit GLADX. Dans ces études le profil phylogénétique du résultat final est instructif. En effet, dans un grand phylum de la lignée humaine le gène est présent chez une seule espèce de Primates et entraîne la déduction de plusieurs pertes indépendantes (Illustration 40).

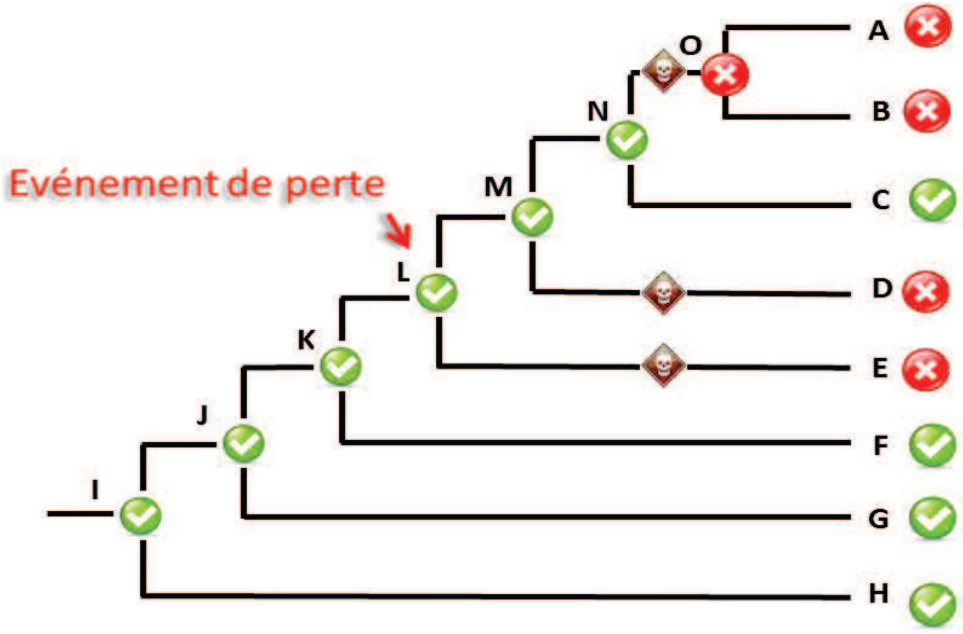


Illustration 40 : Effet de la sur-prédiction d'un gène sur la détection des événements de pertes
Dans ce phylum, la sur-prédiction du gène dans l'espèce C par GLADX entraîne la détection de 3 pertes indépendantes. Sans ce gène, une seule perte serait détectée à l'ancêtre L (flèche rouge).

Sans la détection de ces gènes, GLADX aurait prédit des pertes beaucoup plus anciennes. Lorsque l'on analyse la phylogénie qui a permis de trouver le gène présent dans l'espèce en question, il apparaît que les séquences de la phylogénie correspondent à des paralogues proches qui sont pris pour orthologues. Ces sur-prédictions d'orthologie se produisent lorsque les nœuds qui lient ces séquences dans la phylogénie ont un mauvais *bootstrap*. Dans les phylogénies, l'outil NOTUNG arrange ces nœuds peu soutenus pour les placer de manière parcimonieuse et limiter le nombre de duplications. Il se trouve que dans les 4 études analysées, NOTUNG semble engendrer des artefacts de duplications. Ainsi ces pertes définies comme récentes par GLADX sont en fait des pertes survenues il y a plus de 30 millions d'années.

L'analyse de deux échantillons aléatoire de pertes de gènes unitaires trouvés par GLADX, montre qu'il existe des artefacts de datation des événements qui provoque la détection de pertes anciennes à des périodes plus récentes. Il est donc normal de ne pas trouver de

pseudogènes pour ces pertes anciennes. Les analyses montrent également que GLADX présente des limites dans la détection des pseudogènes. En effet, GLADX peut ne pas détecter des pseudogènes à cause du nombre de hits qu'il étudie, ou parce que la phylogénie de certains hits échoue. Néanmoins, les résultats de GLADX sont encourageants car ils ne sont pas contradictoires. Les événements de pertes de gènes unitaires chez l'homme sont avérés. Dans les pertes récentes de gènes unitaires, les événements de délétions ne peuvent pas être définitivement écartés.

3.2.2.3.2.3 **La délétion dans les événements de pertes de gènes unitaires**

Pour Costello, la quasi-totalité des pertes de gènes unitaires ne semble pas liée à un événement de délétion (Costello *et al.*, 2008). Le phénomène de perte de gènes unitaires par délétion semble rare mais non inexistant.

Parmi les études des 20 pertes récentes analysées ci-dessus, l'étude (ENSMUSP00000065456) a particulièrement attiré mon attention. La perte détectée dans cette étude pourrait être liée à un événement de délétion. En analysant les résultats de cette étude, j'ai observé que GLADX a trouvé dans le génome de *P. abelii*, *G. gorilla*, *P. troglodytes*, une séquence orthologue intacte dans le premier hit retrouvé par blast. Pour ces espèces, le premier hit est fortement différent des hits suivants. L'analyse des hits orthologues de ces espèces montre qu'aucun équivalent n'est trouvé chez *H. sapiens*. Les autres hits retrouvés chez les Primates sont également trouvés chez *H. sapiens* et leurs analyses par GLADX montrent que ce sont des séquences paralogues. Il semble peu probable qu'une pseudogénisation rapide au cours des 6 millions d'années d'évolution engendre une si grande divergence de séquence et empêche de la retrouver. Peut-on parler alors de délétion ? Dans l'étude ENSMUSP00000065456, chez les Primates où l'orthologue a été trouvé, il existe vraiment un seuil de différence important entre le hit de cet orthologue et les autres hits. Ainsi, l'absence de ce hit chez *H. sapiens* fait penser que cette perte est un candidat idéal pour être la première perte par délétion décrite dans cette espèce.

Si j'ai trouvé une délétion dans un échantillon de 20 études touchant des pertes récentes vérifiées manuellement, il est donc possible que d'autres cas de délétion existent parmi les 218 pertes récentes détectées par GLADX sans signal de pseudogènes. C'est un travail à mener, qui demande des investigations supplémentaires aux résultats de GLADX.

3.2.2.3.3 Etude du temps de fixation des gènes avant d'être perdus

Grâce à la connaissance des dates d'apparition des gènes unitaires et celles de leur perte, j'ai pu analyser leur rétention au cours de l'évolution. Dans le tableau ci-dessous (Tableau 5), j'ai classé les pertes selon les dates d'apparition des gènes et les dates de leur disparition. Pour chercher les pertes de gènes unitaires, j'ai paramétré GLADX afin qu'au moins trois orthologues soient présents dans la première phylogénie pour continuer l'étude. De par ce fait, les gènes les plus récents pouvant être observés par GLADX, et qui sont perdus chez *H. sapiens*, sont apparus chez l'ancêtre des Hominidés (*Hominidae*, taxid 9604), et sont présents chez *P. abelii*, *G. gorilla* et *P. troglodytes*. Un seul pseudogène de ce type est trouvé. Au total seulement 4 pertes concernent des gènes qui se sont établis il y a seulement 15 ou 30 millions d'années. Les autres pertes concernent des gènes qui se sont établis il y a plus de 60 millions d'années.

	Age en My >	1408	1322	988	744	700	455	361	324	161	156	104	98	30	15	8	6	0	< disparition			
Age en My v	Taxid lignée Humaine v>	2759	33154	6072	7711	7735	117571	32523	32524	9254	32525	9347	-2	9526	9604	207598	-3	9606	< disparition	Total Pertes	Total Pseudogènes	Total
1408	2759	4484		L95	L7	L12	L66	L17	L31	L6 P2	L5	L10 P1		L5 P4	P4	P2	L2	P2		256	15	271
1322	33154	937			L7	L5	L27	L3	L11	L3		L5	P1	L2 P1				P1		63	3	66
988	6072	4919					L33	L24	L64	L11 P1	L3 P1	L6 P3	L1 P1	L11 P6	P7	L4 P4	L1 P3	L10 P9		176	35	211
744	7711	1073						L7	L15	L1	L1	L3		L1 P3	L1	L1		L2 P3		32	6	38
700	7735	1626						L7	L12	L3	L3	L7	L1	L6 P6	L3	L3 P2		L3 P4		48	12	60
455	117571	4084							L41	L38 P3	L22 P2	L12 P3	L5 P1	L23 P21	L7 P10	L7 P7	L18 P5	L20 P26		193	78	271
361	32523	469								L6		L3	L1	L10 P6	L2 P2	L3	L2 P2	L10 P7		37	17	54
324	32524	640									L2	L2	L1	L7 P3	L2 P4	L3 P1	L1 P3	L3 P4		21	15	36
161	9254	575												L19 P7	L4 P3	L7 P4	L4 P2	L9 P4		43	20	63
156	32525	707												L23 P8	L6 P4	L11 P3	L7 P8	L9 P14		56	37	93
104	9347	977												L37 P11	L8 P7	L6 P5	L6 P6	L8 P11		64	40	104
98	-2	375												P6	L1 P2	L3 P4	L8 P9	L5 P9		17	30	47
30	9526	507															P1	L1 P1		1	2	3
15	9604	568																P1		0	1	1
8	207598	498																				
6	-3	99															Ancêtre à partir duquel les pertes sont détectables					
0	9606	34																				
^ Apparition	^ Apparition																			1007	311	1318
	Total pseudogènes >			0	0	0	0	0	0	6	3	7	3	82	43	32	39	96				311
	Total pertes >			95	14	17	126	58	174	68	36	48	9	144	34	47	49	88				1007
	Total >			95	14	17	126	58	174	74	39	55	12	226	77	79	88	184				1318

Tableau 5 : Pertes rangées selon les dates d'apparition des gènes et de leurs dates de pertes

En rouge sont indiquées les pertes sans signal, en orange les pseudogènes. L correspond à perte (Loss) et P à pseudogènes. Le nombre associé définit le nombre de pertes. Les deux colonnes de gauche indiquent la date à laquelle les gènes semblent apparaître et le taxid des ancêtres concernés. La colonne en jaune indique le nombre de familles de protéines correspondant aux GOs apparus à chaque ancêtre. Les deux lignes du haut indiquent l'âge où les gènes semblent perdus et le taxid des ancêtres concernés.

L'illustration suivante (Illustration 41) présente les pertes de gènes unitaires trouvées par GLADX (en ordonnée) selon le temps de rétention des gènes concernés (en abscisse). Ces résultats montrent clairement que les gènes anciens tendent à être plus stables, ce qui est une tendance attendue (Albà & Castresana, 2005). En effet, ces résultats montrent que plus un gène est établi depuis longtemps moins il a tendance à être perdu.

Lorsque l'on réduit la durée des fourchettes de temps de fixation on perçoit des variations. Elles sont dues au fait que les ancêtres reconstruits, utilisés pour les datations, ne sont pas répartis de façon homogène sur le temps d'évolution observé. Ainsi, en fonction de la fourchette de temps fixée, une quantité d'observations plus ou moins importante est disponible.

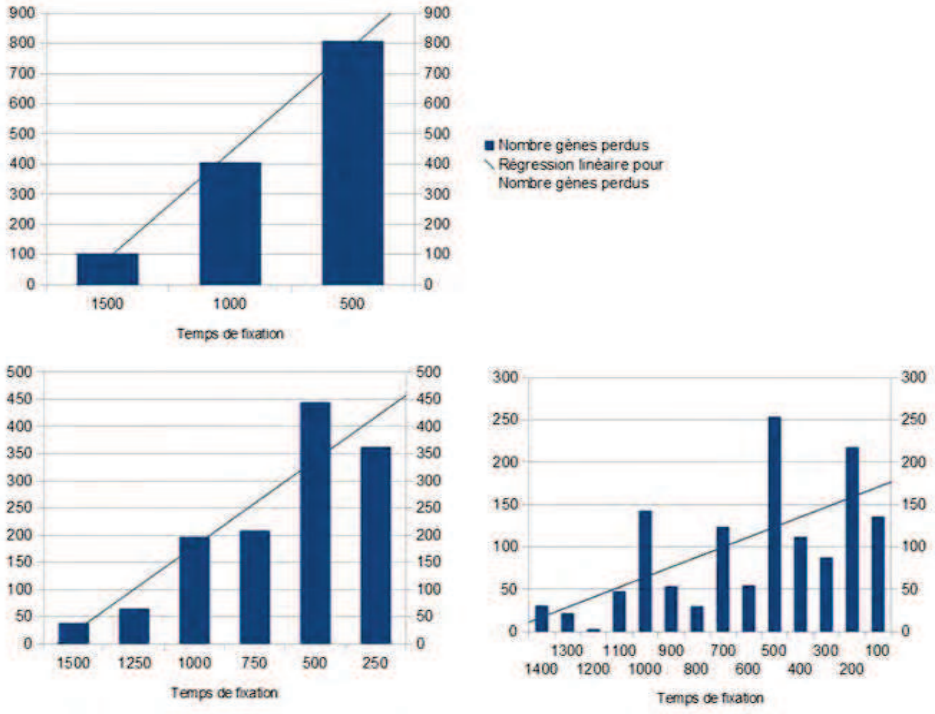


Illustration 41 : Nombre de pertes selon le temps de fixation des gènes

A) Nombre de gènes perdus en fonction du temps de fixation par fourchette de 500 millions d'années.

B) Nombre de gènes perdus en fonction du temps de fixation par fourchette de 250 millions d'années.

C) Nombre de gènes perdus en fonction du temps de fixation par fourchette de 100 millions d'années.

J'ai ensuite analysé les pertes en fonction de l'ancêtre dans lequel les gènes sont apparus. Tandis que les gènes des ancêtres observés ont tendance à être de moins en moins nombreux à disparaître avec le temps (Annexe 20, Annexe 21), les gènes apparus chez les ancêtres des Amniotes et des Tétrapodes ne semblent pas suivre cette tendance, bien au contraire (Annexe 22). Il est difficile de s'avancer sur ces observations car elles manquent de précision. Il

faudrait passer en revue toutes les pertes détectées pour écarter les artefacts, qui pourraient avoir un impact important sur ces résultats.

3.2.2.3.4 **Etude de la fonction des gènes perdus**

Toutes les 1 318 pertes de gènes unitaires détectées par GLADX (Illustration 31) peuvent apporter des informations précieuses sur l'histoire de l'évolution des génomes. Ces pertes sont théoriquement liées à des pertes de fonctions. L'apport de connaissance sur ces fonctions est donc primordial. Pour déterminer les fonctions des gènes perdus, j'ai utilisé un module spécifique de DAGOBAH afin de chercher les annotations fonctionnelles connues des orthologues des gènes perdus trouvés dans les phylogénies. Ce module récupère les annotations fonctionnelles auprès de quatre bases de données (KEGG, ArrayExpress, String et QuickGO). J'ai étudié pour les 1 318 pertes :

- d'abord les fonctions définies expérimentalement. J'ai trouvé des informations fonctionnelles pour 260 pertes sur les 1 318.
- ensuite toutes les informations fonctionnelles disponibles (inférées expérimentalement et par analyses informatiques). J'ai alors trouvé des informations fonctionnelles pour 1 064 pertes. Je n'ai trouvé aucune information fonctionnelle pour 254 pertes.

Une observation générale des résultats montre que les fonctions perdues sont hétérogènes, avec de nombreuses pertes de récepteurs olfactifs depuis l'ancêtre des Euthériens.

Les interprétations évolutives des fonctions perdues ne rentrent pas dans le cadre de cette thèse et font l'objet de nouvelles recherches.

Conclusions et perspectives

L'étude des mécanismes de l'évolution implique l'analyse des événements génétiques et leurs conséquences phénotypiques à différents niveaux. Les résultats de cette analyse mis dans un référentiel historique, permettent de comprendre les processus évolutifs. La description du processus évolutif de plusieurs espèces, plusieurs lignées, permet de faire des corrélations afin de répondre à la question des forces évolutives. Des modèles sont utilisés pour décrire les mécanismes de l'évolution impliqués dans l'histoire d'un trait particulier. Il est possible, par exemple, de mettre en évidence, au niveau génétique, des réponses à des changements environnementaux telles des convergences évolutives entre des Mammifères retournés au milieu marin. **L'étude de l'évolution nécessite donc de procéder à la description des différents événements qui se sont manifestés dans le temps, et d'autre part à leurs interprétations évolutives pour comprendre les mécanismes mis en jeu**. Cette thèse s'intéresse à la description d'un type d'événement particulier : la perte de gènes unitaires. Ce type de perte, qui concerne des gènes bien établis dans les génomes ancestraux, participe à la divergence des génomes au cours de l'évolution (Hughes & Friedman, 2004a). De nombreuses pertes de gènes bien établis, liées à des pertes de fonctions, ont été décrites dans la lignée humaine. Voici plusieurs exemples. Les Mammifères ont perdu des gènes vitellogénine (Brawand, Wahli, & Kaessmann, 2008) qui, chez les non-Mammifères, participent à la création de réserves nutritionnelles dans le jaune d'œuf. Le recours à la lactation chez les Mammifères a peut-être participé au relâchement de la sélection de ces gènes, et induit leur perte. Les Euthériens ont perdu le gène codant la photolyase (Lucas-Lledó & Lynch, 2009 ; Yasui *et al.*, 1994), ce qui a induit un mécanisme moins efficace de réparation de l'ADN par excision de nucléotides. Chez les Catarrhiniens nous pouvons citer la perte du gène Uox qui oxyde l'acide urique en allantoïne à la dernière étape du métabolisme des purines. Le produit final du métabolisme des purines chez les espèces du phylum des Catarrhiniens est donc l'acide urique, alors que c'est l'allantoïne chez d'autres Mammifères. Chez les Hominidés on peut citer la perte du gène Gulo qui induit la perte de la capacité à produire la vitamine C.

Comme le montrent les exemples précédents, les pertes de gènes unitaires ont participé à l'évolution phénotypique qui caractérise chaque phylum. Ces exemples montrent l'intérêt de l'identification des pertes de gènes unitaires pour comprendre l'évolution des espèces.

A l'origine, les pertes de gènes unitaires étaient trouvées et analysées une à une. Avec l'arrivée de génomes complets et l'engouement que suscite l'étude de ce type de perte de gènes, des études à grande échelle ont été effectuées, identifiant des centaines de pertes dans

la lignée humaine (Kuraku & Kuratani, 2011 ; Z. D. Zhang *et al.*, 2010 ; Zhu *et al.*, 2007). Mais ces études à grande échelle, peu automatisées, sont fastidieuses.

Le travail de cette thèse présente d'abord la création d'un outil automatisé dédié à l'étude des événements de pertes de gènes unitaires au sein de nombreux génomes, et, dans une seconde partie, la mise au point d'une stratégie utilisant cet outil pour effectuer des analyses à grande échelle. Grâce à cette stratégie, j'ai étudié l'ensemble des événements de pertes de gènes unitaires survenus dans l'évolution de la lignée humaine. La description de ces événements de pertes s'insère dans une démarche qui vise, à terme, à définir des interprétations évolutives pour comprendre les mécanismes en jeu.

1 L'outil GLADX

Dans un premier temps, j'ai développé, grâce à l'environnement informatique stimulant existant au laboratoire, un outil automatisé original appelé GLADX, dédié à l'étude des événements de pertes de gènes unitaires, de la manière la plus complète (Chapitre II). Cet outil automatique mime l'expertise manuelle pour mener à bien les études. Grâce à l'utilisation de phylogénies, GLADX permet de déterminer de façon précise les relations entre les gènes. De plus, des étapes de ré-annotation systématiques, permettent d'annoter des gènes putatifs, et des pseudogènes. Il permet également d'observer le processus de pseudogénisation au niveau génétique. J'ai testé la méthode automatisée dans GLADX (Article 2) sur 14 pertes par pseudogénisation bien décrites dans la littérature, et démontré que GLADX donne des résultats concluants. J'ai ensuite utilisé GLADX à grande échelle (Chapitre III) pour analyser les pertes de gènes unitaires au cours de l'évolution de la lignée humaine. J'ai vérifié les pertes détectées par GLADX dans la lignée humaine, par l'analyse d'échantillons pris au hasard et montré que les pertes détectées sont avérées.

Cependant, j'ai observé que GLADX, est parfois imprécis dans la détermination de la date de certaines pertes, et dans la détection de certains pseudogènes. Ces imprécisions peuvent avoir un impact en fonction de la question posée. Si, par exemple, le but est de rechercher à tout prix la trace d'un pseudogène, l'approche manuelle est mieux adaptée. En effet, l'utilisation de la phylogénie par GLADX pour identifier les pseudogènes n'est pas efficace quand les séquences sont très dégénérées.

Des points ont été identifiés pour améliorer de façon notable l'outil GLADX et résoudre, entre autre, les imprécisions évoquées ci-dessus. En effet, certaines informations utiles pour

affiner les résultats de GLADX nécessitent des expertises manuelles qui peuvent être automatisées.

1.1 Impact des gaps de séquençages

L'existence de gaps dans les génomes séquencés induit la possibilité de sur-prédire des événements de pertes. En effet, il est possible que des orthologues recherchés se trouvent dans des zones non encore séquencées. Dans le projet de recherche sur la lignée humaine, j'ai fait l'hypothèse que les génomes sont entièrement séquencés et ne contiennent pas de gaps. En réalité, peu de génomes sont complètement séquencés. Chez les Mammifères seuls les génomes de *M. musculus* et *H. sapiens* sont considérés comme entièrement séquencés (Church *et al.*, 2009 ; IHGSC, 2004). Mais, même lorsqu'un génome est considéré comme entièrement séquencé, il existe encore des gaps comme cela a été montré récemment pour le chromosome 20 de *H. sapiens* (Minocherhomji *et al.*, 2012). Les gaps étant rares chez *H. sapiens*, leur impact sur les résultats de pertes de gènes unitaires chez *H. sapiens* peut être considéré comme quasiment nul. En revanche, il est légitime de considérer que dans les génomes « moins bien » séquencés, des gaps puissent engendrer la détection de pertes qui sont de faux positifs. Des méthodes d'expertise existent pour répondre à ce type de problème, mais elles ne sont pas encore implémentées au sein de GLADX. Par exemple, il est possible d'exploiter les bases de données d'ESTs à l'aide de blasts. Ces bases de données peuvent contenir des séquences codantes qui ne sont pas présentes au sein des génomes utilisés (pseudogènes et gènes). S'il s'agit d'analyser des génomes mal séquencés, comme chez les oiseaux, l'utilisation des bases de données ESTs par GLADX, permettrait de distinguer des séquences enfouies dans les zones encore non séquencées des génomes.

1.2 Amélioration de GLADX

Pour affiner les résultats de GLADX, certaines informations utiles qui sont obtenues par expertise manuelle, peuvent être automatisées.

Il serait utile d'ajouter au sein du module GLADX, un agent spécialisé dans la recherche de transcrits (ESTs). Les informations données par un tel agent seraient utiles pour confirmer les gènes et les pseudogènes dans les génomes mal séquencés (1.1). L'analyse des ESTs peut améliorer les analyses de pertes, mais aussi donner des informations sur la transcription de certains pseudogènes. De plus, l'analyse des transcrits pourrait permettre de confirmer ou

infirmer les mutations observées dans les séquences génomiques.

Automatiser la détection de la délétion de gènes pour les pertes de gènes récentes est envisageable. La délétion d'un gène a de grandes chances d'engendrer la délétion des gènes avoisinants. Ainsi une délétion peut être mise en évidence, si les gènes avoisinants sont également absents. Pour reconnaître ces absences, GLADX peut interroger les bases de données génomiques afin de comparer la synténie autour du gène étudié chez les espèces qui ont le gène, et chez les espèces qui ont perdu le gène. Et pour vérifier que les gènes avoisinants sont réellement absents, GLADX doit vérifier que leurs transcrits sont absents dans les bases de données d'ESTs. En effet, si pour le gène étudié il n'y a aucun EST détecté par TBLASTN, alors les gènes environnants ne devraient pas non plus avoir d'ESTs.

L'automatisation d'une recherche dans les bases de données de polymorphisme nucléotidique (SNP) peut également apporter des informations intéressantes sur le polymorphisme des pseudogènes récents comme cela a été démontré pour le gène/pseudogène CASP12 (Saleh *et al.*, 2004).

Un des éléments qu'il faut développer en priorité est un agent dédié à l'estimation du taux d'évolution des séquences retrouvées par GLADX (pseudogènes ou gènes). L'outil **codeml** présent dans le paquet PAML (Yang, 2007) permet ce type d'étude. L'intégration de cet outil serait utile pour obtenir des preuves supplémentaires afin de déterminer si une séquence orthologue donnée est pseudogénisée.

Pour déterminer l'orthologie de séquences, l'outil de phylogénie n'est pas infaillible. En effet j'ai observé certains artefacts de pertes. Des séquences, apparemment orthologues après expertise, ne sont pas détectées par GLADX car la phylogénie nécessaire pour déterminer leur orthologie n'a pas fonctionné. Ainsi, lorsqu'une phylogénie échoue, GLADX devrait expertiser l'alignement produit en amont de la phylogénie. Combiné à une analyse de synténie conservée, des séquences pourraient être déterminées comme orthologues.

Comme souligné dans l'article 2 et dans le chapitre III (3.2.2.3.2.2), des gènes peuvent être sur-prédits dans les bases de données utilisées, ou sur-prédits par GLADX. Ces sur-prédictions font avorter la détection de l'absence du gène et peuvent entraîner la détection de pseudogénisations plus récentes qu'il ne semble. Par exemple, quand un gène perdu chez les Catarrhiniens est sur-prédit comme existant chez le *G. gorilla*, GLADX en déduit que le gène était fonctionnel chez le dernier ancêtre commun à *H. sapiens* et *G. gorilla*. Ce genre de cas peut être détecté *a posteriori* par une analyse des profils phylogénétiques des événements

trouvés par GLADX. Les gènes induisant des profils phylogénétiques douteux peuvent être ré-annotés par GLADX. En effet, un paramètre de GLADX permet de lancer des analyses en omettant les orthologues de certaines espèces afin que leurs séquences soient ré-annotées par GLADX. Cette analyse *a posteriori* doit être automatisée dans GLADX.

Avec GLADX, l'étape de reconstruction de séquences ancestrales permet d'observer les mutations génétiques apparues au cours de l'évolution. Mais, l'étude des mutations par le scanner d'une séquence ancestrale par rapport à une autre séquence ancestrale est biaisée. En effet, l'analyse des mutations entre deux séquences ancestrales se base obligatoirement sur la transposition des exons d'un gène contemporain connu sur ces séquences ancestrales. Il est alors légitime de se poser la question de la préservation des exons au cours de l'évolution (Keren, Lev-Maor, & Ast, 2010). Il serait intéressant de développer une approche qui fournirait une prédiction protéique dans la séquence ancestrale reconstruite. Cette prédiction pourrait être utilisée afin de faire une comparaison de la structure exonique d'un gène ancestral avec celle de son orthologue plus moderne. Ainsi les fenêtres exoniques à observer sur les séquences ancestrales pourraient être mieux définies.

2 L'étude à grande échelle des pertes de gènes dans la lignée humaine

Dans un deuxième temps, j'ai développé une stratégie pour effectuer une analyse à grande échelle des pertes de gènes unitaires. J'ai appliqué cette stratégie pour étudier l'ensemble des pertes de gènes unitaires qui caractérisent l'évolution du génome humain. Le choix de la lignée humaine a été fait en raison de l'intérêt que suscitent les études anthropocentriques, pour compléter un projet du laboratoire sur l'histoire du protéome humain, et confronter les résultats à ceux de la littérature portant sur la lignée humaine.

La stratégie se décompose en deux étapes. Dans une première étape j'ai créé des GOs, que j'ai ensuite analysés pour détecter les pertes de gènes unitaires putatives. Dans une deuxième étape, j'ai sélectionné et étudié avec GLADX les GOs présentant des pertes dans la lignée humaine.

2.1 La sélection des pertes putatives par l'étude des groupes d'orthologues

J'ai utilisé l'outil OrthoMCL pour former des groupes d'orthologues putatifs (GOs). L'analyse des GOs ainsi créés a permis de détecter 6 237 GOs contenant des pertes putatives de gènes unitaires dans la lignée humaine.

D'une part, l'analyse de ces 6 237 GOs par GLADX a montré que 45% des pertes sont de faux positifs. En effet, dans 45% des études, les orthologues humains trouvés absents dans les GOs sont, en revanche, trouvés présents par GLADX.

D'autre part, j'ai comparé les résultats (1 318 pertes) de l'analyse des 6 237 GOs aux 107 pseudogènes unitaires décrits dans les publications de Zhu (Zhu *et al.*, 2007) et de Zhang (Z. D. Zhang *et al.*, 2010) et constaté que 40 (32 + 6 + 2) ne sont pas détectés à l'étape de l'étude des GOs (Chapitre III, 3.2.2.3.1.1). En ce qui concerne ces 40 pseudogènes, les gènes de référence utilisés dans leurs travaux de recherches sont éliminés (6 cas) par l'outil qui créé les GOs, ou font partie de GOs qui ne satisfont pas les critères d'études (32 GOs possèdent une protéine humaine présente ; 2 GOs semblent spécifiques à une lignée différente de la lignée humaine). Sachant que les pertes de gènes unitaires présents dans de nombreux GOs sont des faux positifs, il est probable que de vrais négatifs existent parmi les 40 pertes non détectées par l'analyse des GOs. Autrement dit, ils peuvent correspondre à des pertes omises par le filtre OrthoMCL utilisé.

La création de GOs et leur analyse permettent d'obtenir rapidement de nombreuses pertes potentielles dans les lignées étudiées. Néanmoins, mes résultats montrent que cette méthode manque de précision. Ce constat prouve l'importance de GLADX qui permet de pousser la recherche sur les GOs obtenus afin de les vérifier, et ainsi apporter un résultat robuste. L'approche des GOs représente une méthode adaptée pour filtrer rapidement les pertes de gènes putatives et le module GLADX permet de les étudier avec plus de précision.

L'analyse des GOs est une méthode rapide et intéressante qui peut être utilisée à d'autres fins que celles de l'analyse des pertes de gènes unitaires, comme par exemple l'étude des gains de gènes, des THG, etc.

2.2 Les événements détectés par GLADX

Les 6 237 pertes de gènes unitaires putatives de la lignée humaine sélectionnées par l'étude des GOs créés avec OrthoMCL, ont été étudiées par GLADX. L'analyse phylogénomique de ces 6 237 gènes par GLADX confirme 1 318 pertes de gènes unitaires dans la lignée humaine, et annote de nouveaux gènes orthologues chez l'homme dans 398 études.

2.2.1 Les nouveaux gènes annotés par GLADX

Malgré la qualité de l'annotation du génome humain dans les banques de données, les résultats obtenus montrent que 398 gènes fonctionnels semblent être absents de la base de données Ensembl V.57 utilisée. Avec la base Ensembl V.61, plus récente, je montre qu'une partie de ces 398 gènes annotés pseudogènes dans Ensembl V.57, sont dorénavant annotés comme fonctionnels. Ces résultats montrent que GLADX est efficace pour annoter de nouveaux gènes, et que, malgré l'utilisation d'un génome très bien annoté, de nombreux gènes sont encore manquants. Bien que l'étude soit centrée sur la lignée humaine, j'ai nouvellement annoté 1 569 gènes dans les autres génomes étudiés. Ces résultats permettent d'entrevoir le gros travail à faire pour compléter l'annotation des génomes.

2.2.2 Les pertes de gènes unitaires détectées

Ce travail de recherche a permis d'identifier 1 318 pertes de gènes unitaires dans la lignée humaine.

La comparaison des résultats avec deux études (Z. D. Zhang *et al.*, 2010 ; Zhu *et al.*, 2007), montre que 47 pertes décrites sont retrouvées par mon approche. En rajoutant les cas des pertes des récepteurs olfactifs de ces études, et les recherches complémentaires que j'ai effectuées dans la littérature, je trouve au total 79 pertes déjà identifiées. Une étude récente (Kuraku & Kuratani, 2011) décrit 141 pertes dans la lignée humaine après le dernier ancêtre commun des Amniotes. Si je compare mes résultats à ceux de cette étude, je retrouve 111 de ces pertes. Les autres pertes ne sont pas étudiées avec GLADX à cause du filtre OrthoMCL utilisé, car les GOs contenant les gènes pris en référence ne satisfont pas aux critères d'études. Ainsi, parmi les 1 318 pertes trouvées, 190 ont déjà été identifiées dans la littérature. Il semble donc que 1 128 pertes de gènes unitaires détectées par GLADX dans la lignée humaine n'aient jamais été décrites précédemment.

Dans la recherche exhaustive des événements de pertes de gènes unitaires dans la lignée

humaine j'ai mis à jour 9 942 événements de pertes dans d'autres lignées. Au total, au cours des analyses effectuées avec GLADX, 11 260 événements de pertes de gènes unitaires (2 878 pseudogénisation + 8 382 pertes sans pseudogène) ont été détectés, répartis dans les lignées définies par les 43 espèces utilisées. A cela, je dois ajouter l'annotation de 1 967 nouveaux gènes parmi ces 43 espèces. L'analyse de l'ensemble des événements est disponible pour la communauté scientifique, à travers l'interface web IODA (Chapitre III, 2.3).

L'utilisation de nombreuses espèces par GLADX (44 pour l'analyse des pertes qui semblent survenues avant le LCA des Amniotes et 24 espèces pour celles survenues après le LCA des Amniotes) engendre la détection de nombreuses pertes. De plus, la divergence entre la plupart des espèces est importante. Plus le temps d'évolution indépendante des espèces est important, plus des mutations ont pu apparaître ; les pertes de gènes unitaires en font partie. De manière générale, peu de pseudogènes ont été détectés parmi les pertes de gènes unitaires (311 sur 1 318). Ce résultat est logique car au cours de l'évolution, les pseudogènes ne restent pas longtemps détectables au sein des génomes, à cause de la dérive génétique sous évolution neutre. Cette dérive s'explique par le taux de divergence, traditionnellement estimé entre 1% et 2% de divergence par million d'années, ce qui rend difficile la détection des pseudogènes de plus de 60 millions d'années.

L'étude des pseudogènes a permis de détecter 26 109 mutations génétiques (mutation de sites d'épissages, apparition de codons stop, etc.). L'ensemble de ces mutations est décrit dans un contexte évolutif qui permet d'observer les processus de pseudogénisation. L'analyse des pseudogénisations peut apporter un éclairage sur les modalités de ce processus. Cette analyse reste à faire.

2.3 **Les résultats**

Nous avons vu ci-dessus que l'étude des pertes de gènes unitaires dans les GOs, et leur étude approfondie par l'outil GLADX, a permis de mettre en évidence 1 318 pertes de gènes unitaires dont 311 pseudogènes unitaires dans la lignée humaine. Parmi ces 1 318 pertes, 1 128 semblent n'avoir jamais été décrites. Les 1 318 pertes de gènes unitaires détectées dans cette étude exhaustive montrent l'importance du phénomène. Rapportées aux 22 523 gènes du génome humain, elles représentent plus de 5 % des gènes. Les pertes de ces gènes au cours de l'évolution ont participé à la diversification des génomes.

Malgré les imprécisions que j'ai relevées dans la détermination de la date de certaines pertes,

et la détection de certains pseudogènes, l'analyse d'échantillons de résultats montre que les pertes trouvées dans la lignée humaine sont réelles. Pour obtenir des résultats plus précis, j'ai identifié et propose les améliorations à apporter à GLADX (Conclusions et perspectives, 1.2). L'amélioration des résultats est également possible en confrontant les événements de pertes de gènes unitaires détectés avec d'autres types d'événements (Conclusions et perspectives, 2.4.2).

Les 1 318 pertes de gènes unitaires dans la lignée humaine et les 9 942 autres pertes trouvées dans les autres lignées, qu'il s'agisse de pseudogènes ou non, sont des données intéressantes pour étudier les pertes de fonctions. Chacune de ces pertes est une information importante pour comprendre l'histoire évolutive des espèces.

2.4 Perspectives

A la suite du travail de thèse sur les pertes de gènes unitaires dans la lignée humaine, un article, en cours de rédaction, présentera la stratégie et l'ensemble des résultats obtenus. Outre l'histoire complète des pertes de gènes unitaires dans la lignée humaine, et la mise en exergue des pertes encore jamais décrites, je souhaite pour cet article, mener des analyses sur le type de fonctions perdues en fonction des périodes de l'évolution (Conclusions et perspectives, 2.4.1). Ces analyses fonctionnelles permettront peut-être de dégager des caractéristiques évolutives en lien avec des modifications de l'environnement ou des organismes (éthologique, physiologique, etc).

L'étude de l'évolution nécessite la description des événements affectant les génomes. Le travail mené sur la perte de gènes unitaires dans la lignée humaine a permis d'obtenir de nombreux résultats. Les premières analyses effectuées dans le cadre de cette thèse ont permis de vérifier la stratégie utilisée et la qualité des résultats. L'analyse biologique commence à peine. Il faut maintenant interpréter les résultats des événements de pertes pour appréhender l'histoire de l'évolution des génomes ancestraux ayant abouti au génome humain. Mais, l'interprétation évolutive de chaque perte exige une étude approfondie, notamment au niveau fonctionnel, pour comprendre les tenants et les aboutissants de l'événement. Cette analyse importante et intéressante ne fait pas partie de cette thèse. Elle fera l'objet de nouvelles recherches.

Dans cette perspective, j'ai décrit de nombreux événements qui seront utilisés prochainement

pour appréhender les mécanismes évolutifs impliqués dans les lignées étudiées. Par exemple, les mécanismes évolutifs peuvent être appréhendés par l'étude de la convergence évolutive qui permet d'observer des adaptations semblables, en réponse à des contraintes environnementales semblables. Comme les convergences évolutives résultent de deux évolutions indépendantes, leurs études nécessitent l'utilisation d'approches phylogénétiques. En effet, les corrélations utilisant l'information phylogénétique permettent de mettre en évidence les phénomènes de convergences et de co-convergences évolutives (Conclusions et perspectives, 2.4.3).

Les orientations du laboratoire ont pour but d'analyser l'histoire évolutive de l'ensemble des événements génomiques pouvant être observés dans une lignée. Pour y parvenir, le laboratoire dispose de nombreux modules dans DAGOBAH, permettant une riche description des différents types d'événements (changements d'architecture en domaines de protéines, duplication, THG, pertes, etc), et d'un module de DAGOBAH dédié à l'étude de la convergence à partir des données recueillies dans l'ontologie.

A partir de ces outils, il s'agira dans une première phase d'analyser la convergence évolutive des pertes, ce qui peut donner des résultats très instructifs. Etant donné le nombre de pertes détectées dans ce travail de thèse, et le temps nécessaire pour effectuer l'analyse évolutive complète d'une perte, l'étude de la convergence permettra de sélectionner les cas intéressants de pertes adaptatives, pour les étudier en premier lieu. Toutes les pertes de gènes unitaires mises en évidences sont des clés importantes pour comprendre l'évolution des espèces. Pour décrypter les mécanismes évolutifs il est important de prendre en compte tous les types d'événements. En ce sens, dans une deuxième phase il sera très instructif d'étudier la convergence avec l'ensemble des événements étudiés au laboratoire.

2.4.1 Analyse fonctionnelle

Une des perspectives à exploiter à court terme est une analyse globale de l'aspect fonctionnel des gènes ayant été détectés comme perdus. On peut ainsi étudier la fonction des gènes perdus selon le temps de fixation des gènes dans les génomes, ou selon les périodes où surviennent les pertes. Pour ce faire, les fonctions récupérées dans QuickGO sont adaptées. Elles ont la particularité d'être classées selon une ontologie biologique exhaustive. Il existe de nombreux niveaux de description : la description est d'autant plus fine que le niveau est profond. Les niveaux les plus élevés se classent en trois catégories : processus biologique, composant cellulaire et fonction moléculaire. Ces trois classes se divisent respectivement en 31, 14, et 20

sous-catégories (Annexe 23). Ces différents groupes descriptifs permettent d'avoir une idée précise et diversifiée du type de fonctions associé aux pertes de gènes unitaires détectées.

Ainsi, pour connaître les types de fonctions perdues en fonction de l'époque où les pertes sont survenues, il faut inférer ces informations sur un arbre des espèces (Illustration 42). Ces résultats peuvent mettre en évidence des familles de fonctions perdues en lien avec des spécificités caractérisant l'évolution des différents phyla observés.

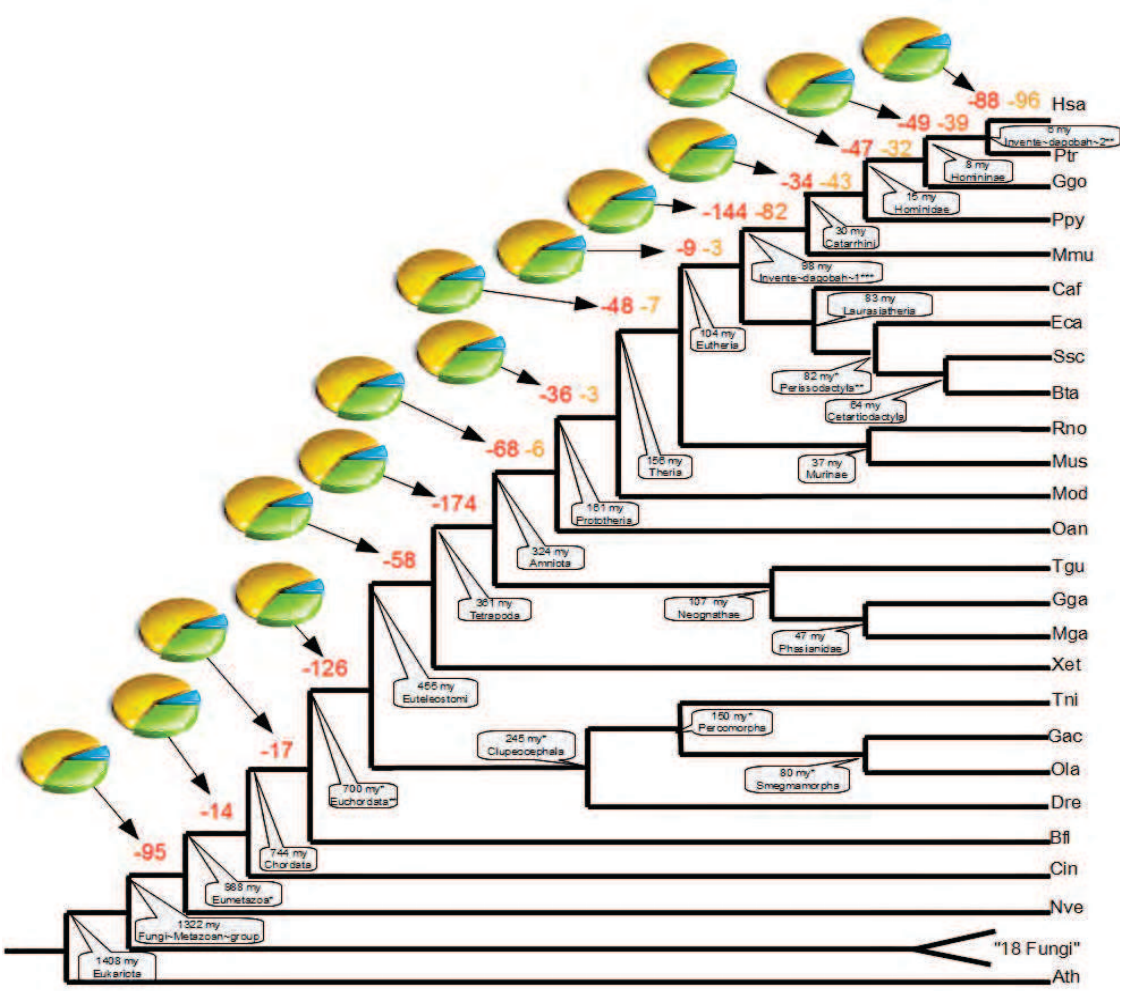


Illustration 42 : Représentation de la recherche des types de fonctions perdues au cours de l'évolution
Les nombres en rouge indiquent les pertes et les nombres en orange indiquent les pseudogènes. Pour la description de la topologie se référer à l'Illustration 27.

La connaissance de la fonction des gènes perdus est importante pour interpréter l'histoire évolutive des pertes. L'étude des fonctions des 1 318 gènes détectés comme perdus a mis en évidence 254 pertes sans aucune information fonctionnelle parmi les orthologues présents. Lorsqu'aucune information n'est disponible, il est possible de faire une annotation fonctionnelle in-silico en comparant la similarité des séquences avec des séquences dont la

fonction est connue, ou à partir de la fonction connue des domaines contenus dans les séquences. Les outils associés aux bases de données Pfam, Prosite ou encore InterPro permettent d'effectuer ce travail. Une deuxième possibilité, plus fastidieuse, consiste à mener des recherches expérimentales (knock-out, transgénèse).

2.4.2 Confrontations avec différents événements

L'étude que j'ai menée sur la perte de gènes unitaires dans la lignée humaine, fait partie d'un projet plus large du laboratoire cherchant à analyser l'ensemble des événements génétiques apparus dans l'histoire du protéome des Chordés. Ainsi, d'autres événements sont disponibles dans la base de données ontologique du laboratoire (changements d'architecture en domaines de protéines, duplications, THGs, etc.). Ces données peuvent être utilisées pour améliorer les résultats de GLADX, et ceux des autres modules de DAGOBAH. En effet, la confrontation des données peut permettre de mieux déterminer les événements qui ont réellement eu lieu et corriger les éventuels artefacts, mais également d'en tirer de nouvelles observations

Les THGs

L'étude de pertes de gènes unitaires que j'ai menée chez les Eucaryotes ne prend pas en compte les événements potentiels des THGs. Une récente expertise manuelle montre que certains résultats obtenus par GLADX semblent être associés à des événements de THGs. Au sein de DAGOBAH, un module existe pour déterminer des événements de THGs à partir de l'analyse de phylogénies. L'analyse de toutes les phylogénies obtenue au départ des études de GLADX avec cet agent, permettrait de détecter des événements de THGs.

Les changements d'architecture en domaines de protéines

Il semble probable que certains pseudogènes détectées par GLADX soient des gènes liés à des événements de changements d'architecture en domaines (échanges, gains et pertes de domaines). En effet parmi les pseudogènes détectés par GLADX, certains sont analysés uniquement au niveau protéique. Les séquences orthologues protéiques retrouvées ne semblent pas assez conservées pour une analyse au niveau nucléotidique. La divergence au niveau protéique peut être liée à des événements de changements d'architecture rendant la protéine peu semblable à la protéine d'origine. Ainsi, il serait intéressant de confronter les événements de pseudogénisation aux événements de changements d'architecture étudiés dans le phylum des Chordés.

2.4.3 Etude de la convergence évolutive

Les corrélations entre espèces confondent les événements partagés au cours de l'évolution et ceux qui sont indépendants. Par contre l'utilisation de phylogénie permet de faire cette différence. Ainsi dans l'exemple illustré ci-dessous (Illustration 43), les quatre gènes représentés (A et B bleus, A et B rouge) sont tous présents dans l'ancêtre commun de ces espèces. Seuls les événements associés à la paire de gènes en bleu sont statistiquement significatifs en faveur d'une corrélation évolutive. En effet la co-perte des deux gènes en bleu survient indépendamment 4 fois. Dans l'histoire évolutive de la paire de gènes en rouge, aucune corrélation évolutive ne peut être faite, car un seul événement de perte est survenu.

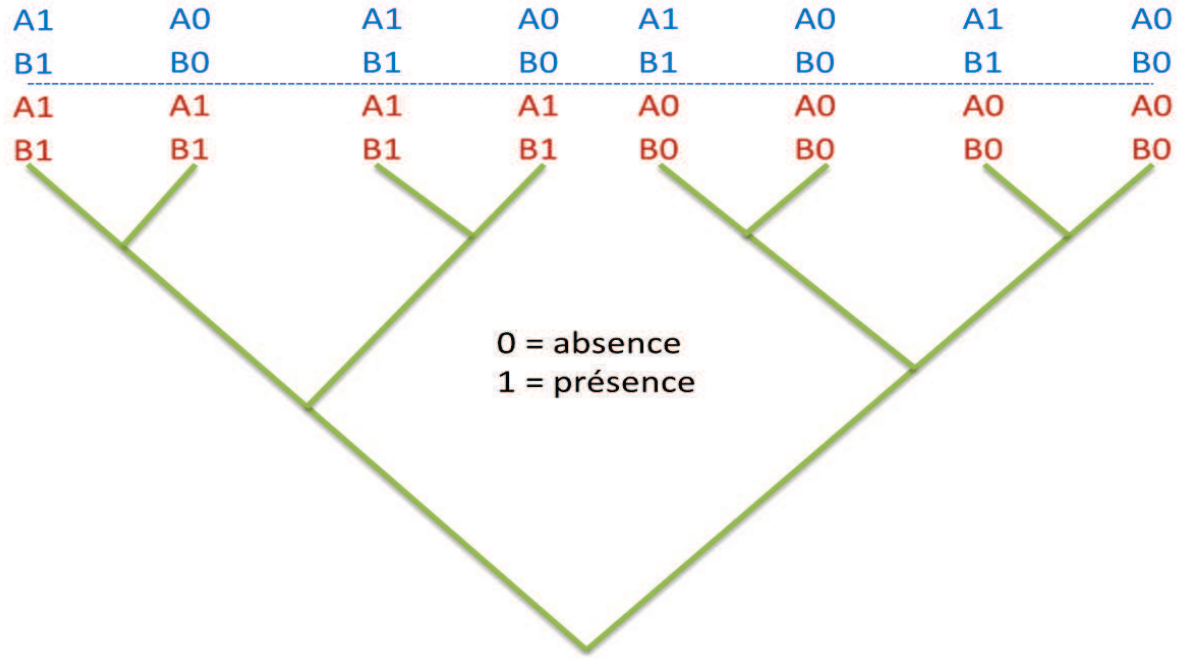


Illustration 43 : Caractères 0 et 1 de 8 espèces et relations phylogénétiques entre ces espèces

Un module de DAGOBAH permet de clustériser les données choisies par dates et par types. Il utilise ensuite la méthode de Pagel (Pagel, 1994) pour appliquer l'approche corrélative expliquée ci-dessus. Je souhaite utiliser ce module sur les données de pertes de gènes unitaires pour mettre en évidence des co-pertes et ainsi déterminer des liens fonctionnels forts entre des gènes perdus. Un cas de ce type a été décrit chez l'homme. C'est celui de la perte du gène UOX qui code pour une enzyme nécessaire dans la voie du catabolisme de la purine. La perte de ce gène a entraîné dans cette voie, en amont immédiat de l'activité de UOX, un relâchement entraînant la perte des gènes HIU et OCHU codant respectivement l'hydrolase et la decarboxylase (Keebaugh & Thomas, 2010). Cette approche corrélative permettra de mettre en relation des événements génétiques de natures différentes collectés avec DAGOBAH dans

l'ontologie. Des recherches au laboratoire sur l'approche corrélative de co-occurrence d'événements sont menées pour permettre à terme de corréler plusieurs caractères à la fois et pas seulement deux à deux comme c'est actuellement le cas.

3 Production de nouvelles données

L'étude qui a été menée sur la lignée humaine peut être appliquée à n'importe quelle autre lignée Eucaryote. Son approche a été développée dans une optique intégrative. Avec ses résultats, une partie de l'histoire de perte de gènes unitaires est désormais connue dans les lignées analysées, qui partagent des ancêtres communs avec l'homme. Ainsi pour étudier la lignée menant à *M. musculus*, il ne manque que l'étude de la lignée menant à M. musculus depuis le dernier ancêtre commun avec l'homme. Les GOs nécessaire pour mener à bien une telle étude avec GLADX, existent déjà. Il suffit de les sélectionner et les filtrer. En effet les GOs créés concernent l'ensemble des espèces utilisées. Si une nouvelle espèce nécessite une étude, elle peut être ajoutée à l'étude existante. Il suffit donc de faire une passe d'OrthoMCL supplémentaire pour rajouter les protéines de cette espèce aux résultats déjà disponibles. Les protéines s'agrègent à des GOs existants, ou permettent de créer de nouveaux GOs avec des séquences qui n'avaient jusqu'alors pas été clustérisées. Les pertes putatives ainsi détectées peuvent ensuite être analysées par GLADX pour approfondir les résultats.

Pour mener des études de pertes de gènes unitaires chez les procaryotes, il sera nécessaire d'apporter des modifications à GLADX. Les modifications doivent prendre en compte le phénomène de transfert horizontal de gènes qui est omniprésent chez les procaryotes. Pour tirer le meilleur parti de GLADX, il faudrait prendre également en compte les THGs au niveau de l'analyse des GOs qui servent de données en entrée de GLADX. Des algorithmes de parcimonie plus adaptés que celui de Dollo peuvent prendre en compte ce type d'événements (Mirkin *et al.*, 2003).

Annexes

Annexe 1

Publication	Espèces utilisées	Méthode utilisé	But de l'étude	Résultats
(R. L. Tatusov, 1997)	5 *Bacteria* (*E. Coli, H. influenza, M. genitalium, M. pneumoniae, Synechocystis*) + 1 *Archaea* (*M. jannaschii*) + 1 *Fungi*. (*S. cerevisiae*)	Création de l'approche des COG par best BLAST hit réciproque. Un COG doit contenir au moins des protéines de 3 génomes.	Création de groupes d'orthologues pour répondre à l'étude de génomes complets	Création de 720 COGs. Les espèces absentes de ces COGs représentent des pertes putatives
(Braun, 2000)	2 *Fungi* : *N. crassa* & *S. cerevisiae*	BLAST de *N. crassa* sur *S. cerevisiae*	Etude comparative des deux génomes de *Fungi*	Découvre quelques dizaines de pertes chez *S. cerevisiae*
(Aravind *et al.*, 2000)	*S.pombe, S. cerevisiae* & tous les autres *Eukaryota* disponibles.	BLAST des protéines de *S. pombe* sur *S. cerevisiae* et autres génomes *Eukaryota* disponibles	Analyse des pertes lignée spécifique	Découvre plus de 300 gènes perdus chez *S. cerevisiae* depuis l'ancêtre commun avec *S. pombe*
(Snel, Bork, & Huynen, 2002)	6 *Archaea*, 11 *Proteobacteria* & 16 « *outgroup* »	Construction de groupes d'orthologues putatifs par comparaison all-againt-all (T. F. Smith & Waterman, 1981). Méthode conceptuellement similaire aux COGs. Utilise reconstruction parcimonieuse en considérant les THG	Etude des flux de gènes chez les Proteobactéries et les *Archaea*	2 500 gènes inférés à l'ancêtre des *Proteobacteria* et 2 050 chez les *Archaea*. 950 semblent perdus chez *E. coli*
(Zdobnov *et al.*, 2002)	7 *Eukaryota* : *A. thaliana, S. cerevisiae, A. gambiae, D. melanogaster, C. elegans, M. musculus, H. sapiens*	Utilisation de la méthode des COG (R. L. Tatusov, 1997)	Etude des génomes de *A gambiae* et *D. melanogaster*	Plus de 187 pertes communes à *A. gambiae, D. melanogaster, M. msuculus* et *H. sapiens* sont trouvées
(Dehal *et al.*, 2002)	5 *Metazoa* : *C. elegans, D. melanogaster, C. intestinalis, F. rubripes, H. sapiens*	BLAST des protéines de *C. intestinalis* sur les autres génomes	Mettre en lumière l'origine des *Chordata*	Trouve quelques centaines de gènes qui match bien avec *D. melanogaster* et *C. elegans* mais pas avec *F. rubripes* et *H. sapiens*. Ce sont des pertes potentielles apparues dans la lignée des *Vertebrata*
(Krylov *et al.*, 2003)	7 *Eukaryota* : *A. thaliana, E. cuniculi, S. cerevisiae, S.pombe, C. Elegans, D. melanogaster, H. sapiens*	Fabrication de cluster par une méthode inspirée de celle des COGs (Roman L	Analyse des pertes comme source de diversité	3 140 KOGs inférés à l'ancêtre des *Eukaryota*. De nombreuses pertes détectées dans 10 lignées. Dont 381

		Tatusov *et al.*, 2003)		dans la lignée menant à *H. sapiens* et près de 2 000 chez *E. cuniculi*
(IHGSC, 2004)	*P. troglodytes, H. sapiens, M. musculus, T. norvegicus*	Comparaison des séquences deux à deux	Amélioration de l'annotation du génome humain	32 pseudogènes d'*H. sapiens* présents chez *M. musculus* dont 6 spécifiques à *H. sapiens* car présents chez *P. troglodytes*
(Hughes & Friedman, 2004a)	7 *Eukaryota* : *A. thaliana, S. cerevisiae, S.pombe, C. Elegans, D. melanogaster, T. rubripes, H. sapiens*	Clusterisation par l'utilisation de BLASTCLUST (Altschul *et al.*, 1997)	Analyses des pertes chez les animaux	2 106 familles de gènes inférés au dernier ancêtre commun de *C. elegans* et *H. sapiens*. De nombreuses pertes inférées dans 6 lignées dont 2 ancestrales. Au total 154 familles de gènes perdus dans la lignée humaine
(Hughes & Friedman, 2004b)	8 *Eukaryota* : *A. thaliana, S. cerevisiae, S.pombe, C. Elegans, A. gambiae D. melanogaster, T. rubripes, H. sapiens*,	Clusterisation par l'utilisation de BLASTCLUST (Altschul *et al.*, 1997)	Analyses des pertes chez les animaux	2 949 familles de gènes inféré au dernier ancêtre commun de *C. elegans* et *H. sapiens*. De nombreuses pertes inférées dans 8 lignées dont 3 ancestrales. Au total 112 familles de gènes perdus dans la lignée humaine
(Koonin *et al.*, 2004)	7 *Eukaryota* : *A. thaliana, E. cuniculi, S. cerevisiae, S.pombe, C. Elegans, D. melanogaster, H. sapiens*	Utilisation méthode de COG amélioré par l'ajout de co-orthologues par la procédure COGNITOR (R L Tatusov *et al.*, 2001) + vérification manuelle pour enlever les faux positif et négatif	Classification évolutive des protéines codées par les génomes des *Eukaryota*	3 413 familles de gènes inférés à l'ancêtre des *Eukaryota*. En tout ils reconstruisent 6 protéomes ancestraux. Ils analysent le gain de familles de gènes dans 12 lignées et les pertes dans 10 lignées. Plus de 300 pertes de familles de gènes détectées dans la lignée menant à *H. sapiens* (cette valeur varie en fonction du placement de *D. melanogaster* dans l'arbre utilisé pour l'inférence)
(Hughes & Friedman, 2005)	8 *Eukaryota* : S. cerevisiae, S.pombe, C. Elegans, A. gambiae D. melanogaster, C. intestinalis, T. rubripes, H. sapiens,	Clusterisation par l'utilisation de BLASTCLUST (Altschul *et al.*, 1997)	Analyse des pertes chez *C. intestinalis* et reconstruction protéome de l'ancêtre des *Chordata*	3 921 familles de gènes présents chez l'ancêtre des Chordés. 798 pertes de familles de gènes détectées dans la lignée menant à *C. intestinalis*, 217 dans le LCA des Vertébrés, puis 217 dans la lignée menant à *T. rubripes* et 230 dans celle menant à *H. sapiens*
(Hahn & Lee, 2005)	*P. troglodytes* & *H. sapiens*	Comparaison des séquences deux à deux	Recherches des pseudogènes unitaires chez *H. sapiens*	Identification de 9 pseudogènes unitaires spécifiques à *H. sapiens*
(Danchin *et al.*, 2006)	7 *Eukaryota* : *A. thaliana, S. cerevisiae, C. elegans, A. gambiae D. melanogaster, M. musculus, H. sapiens*	Utilisation de BD de COGs, vérification phylogénétique et approfondissement manuel	Recherche de gènes perdus chez les *Opisthokonta*	11 pertes de gènes dans la lignée menant aux *Mammalia*
(Blomme	7 *Metazoa* : *D.*	Création de	Etudes des	2 972 duplications à la base

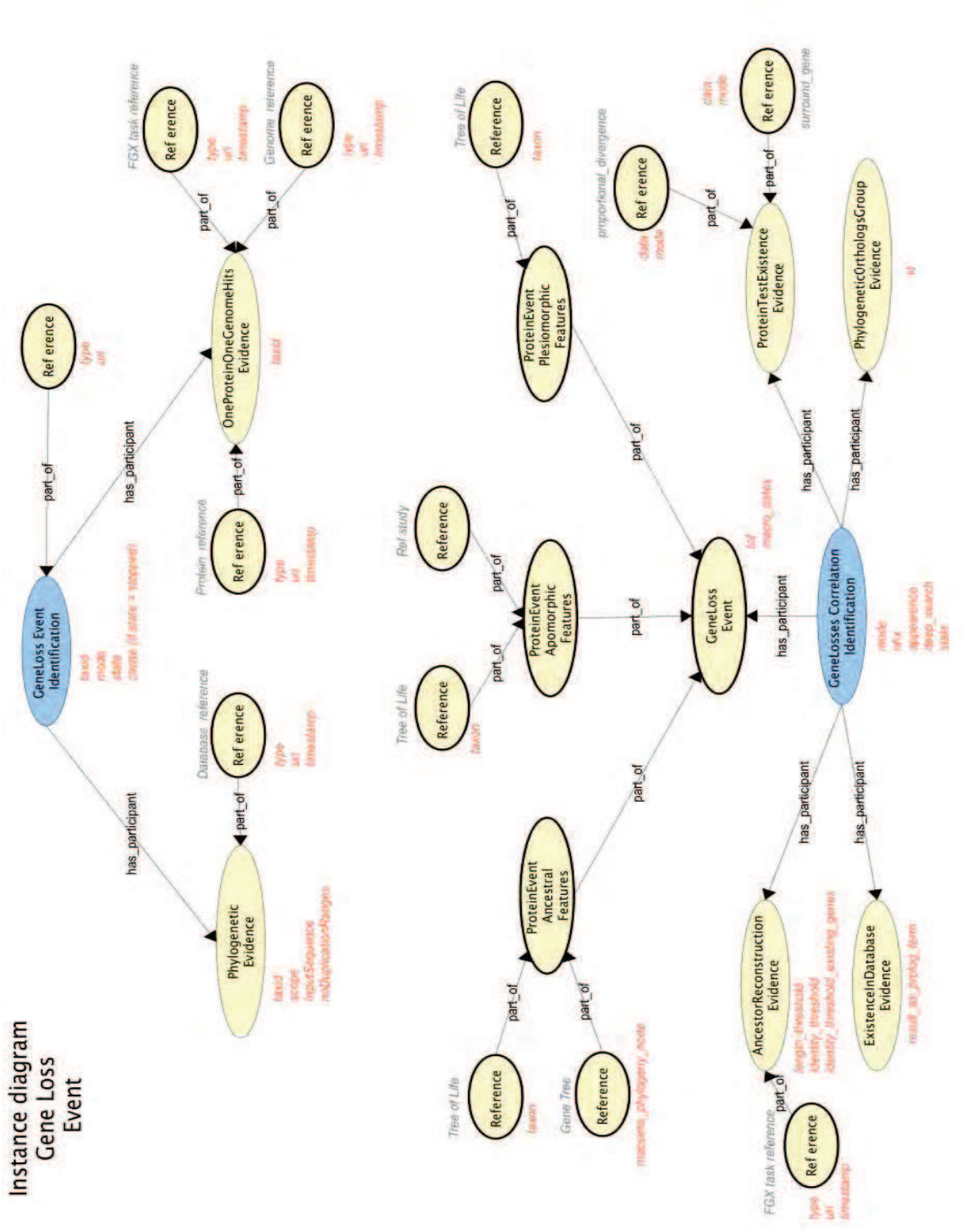
et al., 2006)	*melanogaster, C. intestinalis, T. nigroviridis, D. rerio, X. tropicalis, G. gallus, M. musculus, R. norvegicus, H. sapiens*	familles de gènes par BLASTP puis création de phylogénies et analyses	pertes et des gains chez les *Vertebrata*	des Vertébrés et 484 duplications apparues dans la lignée menant à *H. sapiens*. 363 pertes sont détectées le long de la lignée de *H. sapiens*
(X. Wang *et al.*, 2006)	*P. troglodytes* & *H. sapiens*	Comparaison des ORF de *P. troglodytes* avec ceux de *H. sapiens*	Recherche des pseudogènes unitaires chez *H. sapiens*	67 pseudogènes unitaires spécifiques à *H. sapiens* détectés
(Wyder *et al.*, 2007)	10 *Metazoa* : *A. gambiae, D. melanogaster, A. aegypti, T. castaneum, A. mellifera T. nigroviridis, G. gallus, M. domesticus, M. musculus, H. sapiens*	Utilisation du BLAST réciproque clustérisé par seuil de similarité. Procédure proche de celle des COGs	Analyse des pertes chez les *Arthropoda* et *Vertebrata*	7 144 familles de gène orthologue chez l'ancêtre des *Coelomata*. De nombreuses pertes détectées chez les insectes et les *Vertebrata*. Beaucoup de pertes détectées chez ancêtre des *Amniota*, des *Theria* et *Eutheria* et dans la lignée de *H. sapiens* après la séparation avec dernier ancêtre commun des *Eutheria*. Au total 249 pertes dans la lignée humaine de familles de gènes présents dans l'ancêtre des *Coelomata*, et 425 parmi ceux apparus après la séparation de la lignée des *Chordata* et des *Arthropoda*
(Putnam *et al.*, 2007)	9 *Eukaryota* : *A. thaliana, S. cerevisiae, D. discoideum, C. elegans, D. melanogaster, N. vectensis, T. rubripes, X. tropicalis, H. sapiens*	Utilisation de la méthode du meilleur hit de BLAST réciproque + clusterisation (de Hoon, Imoto, Nolan, & Miyano, 2004)	Etude de l'évolution du génome de l'ancêtre des Eumetazoaires et de *Coelomata*	A partir de 7 766 familles de gènes orthologues chez l'ancêtre des *Eumetazoa*, on est passé à 8 748 chez l'ancêtre des *Coelomata*. Apparemment de nombreuses pertes chez *D. melanogaster, C. elegans et C. intestinalis*
(Wapinski *et al.*, 2007a ; Wapinski, Pfeffer, Friedman, & Regev, 2007b)	17 *Fungi*	SYNERGY : méthode de clusterisation basée sur la similarité réciproque et la phylogénie des espèces utilisées	Identification des gains et des pertes chez les *Fungi*	Reconstruction des familles présentes chez chaque ancêtre et détection de nombreuses pertes et gains de familles de gènes dans toutes les branches de la phylogénie des espèces
(Zhu *et al.*, 2007)	3 *Eutheria* : *M. musculus, C. lupus familiaris, H. sapiens* + autres espèces pour approfondissement manuel	Comparaison structure mRNA de M. musculus sur H. sapiens et *C. lupus familiaris* grâce à au BLAT de l'outil TransMap	Détection des pseudogènes unitaires dans la lignée humaine	26 pseudogènes unitaires détectés chez *H. sapiens*
(Costello *et al.*, 2008)	9 espèces du sous-genre *Sophophora*	Méthode : Fuzzy Reciprocal BLAST	Détection des pertes et des pseudogènes	109 pertes détectées dont 18 pseudogènes. Seulement 1 perte récente par délétion

		(extension du BLAST réciproque pour étudier de multiples espèces simultanément) + clusterisation + analyse manuelle	unitaires dans la lignée menant à *D. melanogaster*	complète (2-5 Millions d'années)
(Fritz-Laylin *et al.*, 2010)	*N. gruberi* et 13 autres *Eukaryota* dont *H. sapiens* et *M. brevicollis* comme représentant des Opisthokontes	BLAST sur une BD de KOGs	Placement de *N. gruberi* chez les *Eukaryota* et analyse des familles de protéines	L'ancêtre des *Eukaryota* semble posséder 4 133 familles de protéines
(Z. D. Zhang *et al.*, 2010)	*H. sapiens & M. musculus* + autres espèces pour approfondissement manuel	BLAST réciproque & clusterisation avec InParanoid (O'Brien, Remm, & Sonnhammer, 2005) + analyses manuelles	Détection semi-automatisée de la présence des pseudogènes unitaires chez *H. sapiens*	76 pseudogènes unitaires trouvés chez *H. sapiens*
(Srivastava *et al.*, 2010)	17 *Eukaryota*	Meilleur hit mutuel	Recherche gain et perte chez *A. queenslandica* et placement parmi les *Eukaryota*	Parmi les 4 670 familles de gènes construites chez les *Metazoa*, 1 286 semblent leur être spécifiques
(Kuraku & Kuratani, 2011)	A) *D.rerio, T. nigroviridis, T. rubripes, G. aculeatus, O. latipes, X. tropicalis, G. gallus, M. musculus, C. familiaris, B. taurus, H. sapiens* + d'autres espèces utilisées uniquement pour la phylogénie provenant de Genbank et Ensembl B) *O. anatinus, M. domestica* + autres espèces	Recherche de perte par phylogénie et vérification par BLASTP	Détection perte chez les *Eutheria*	141 pertes détectées chez l'ancêtre des *Eutheria*

Annexe 1 : Principales études à grande échelle portant sur la perte de gènes lignées spécifiques

Les études qui se basent sur la création de groupes orthologues, utilisent la parcimonie de Dollo pour inférer les pertes.

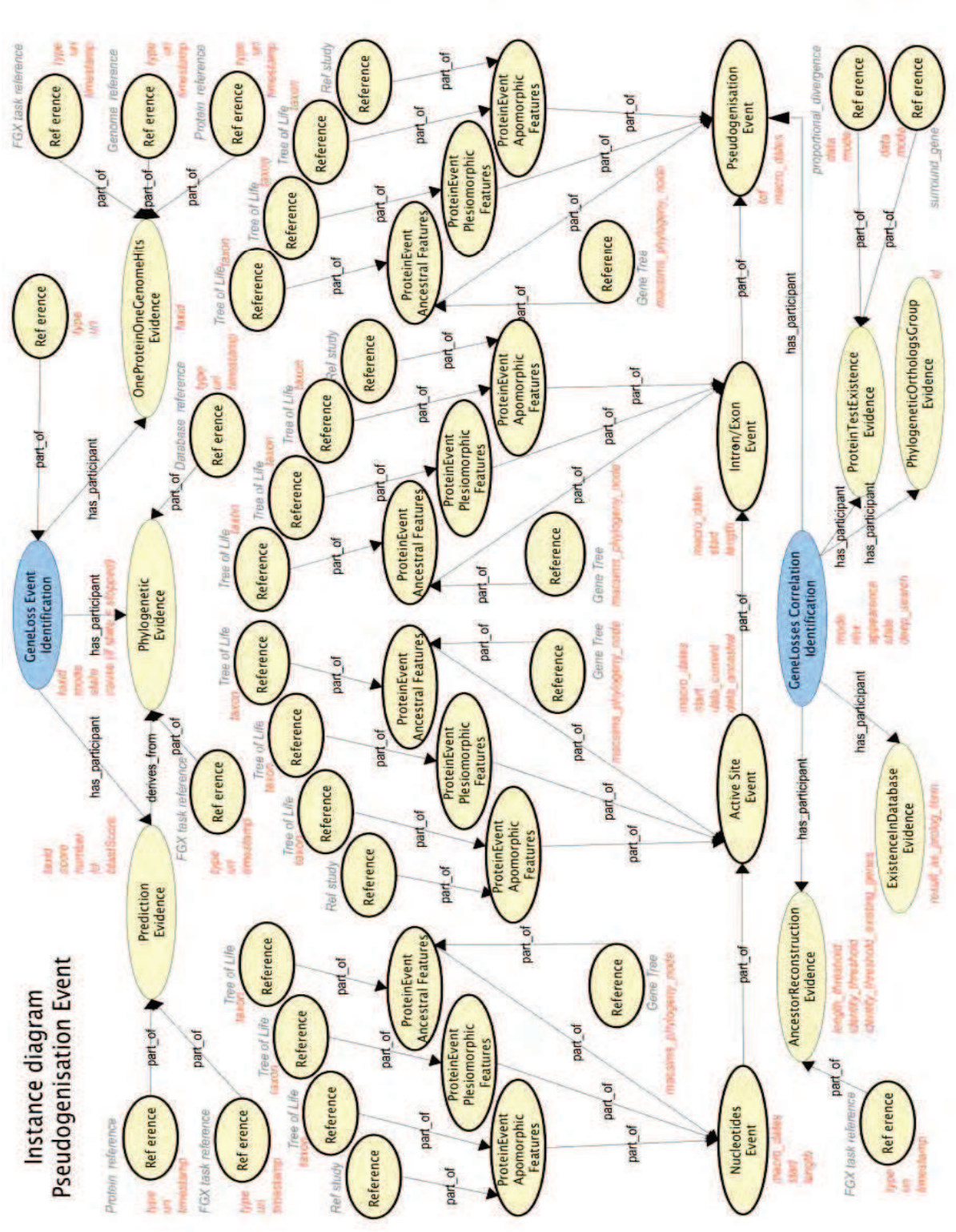
Annexe 2



Annexe 2 : Diagramme d'instance de l'ontologie pour une étude de perte

Ce diagramme représente l'ensemble des classes pouvant avoir une instance lors d'une étude de perte avec GLADX. Pour chaque classe, les données enregistrées associées sont indiquées en rouge. En bleu sont indiqués les processus informatiques. Des instances de natures différentes utilisant une même classe sont indiquées en gris à côte de la classe.

Annexe 3



Annexe 3 : Diagramme d'instance de l'ontologie pour une étude de pseudogénisation

Ce diagramme représente l'ensemble des classes pouvant avoir une instance lors d'une étude de pseudogénisation avec GLADX. Pour chaque classe, les données enregistrées qui lui sont associées sont indiquées en rouge. En bleu sont indiqués les processus informatiques. Des instances de natures différentes utilisant une même classe sont indiquées en gris à côte de la classe.

Annexe 4

Text S1: GLADX parameters

Numerous parameters are available to adjust the behavior of GLADX. Some are essential, such as species and used database, and mode of study (verification of putative lost genes or not). These parameters must be defined before analysis is launched. They are contained in an XML file accessible at:

/home/tower/TOWER_1.03/prod/DGH_2/dagobah.xml

The parameters of agents are defined between the following markups:

```
<engine-def>
   <type>Agent_Name</type>
   ...
</engine-def>
```

A) Parameters defined in the fasta_protein_phylo agent:

"**species_scope_for_phylogeny_study(*'9598,9606,9544,10116,10090,9601,9615,8090,9031,13616, 7719,8364,99883,9593,9103,9913,9796,9823,9258,59729,69293,7955'*)**" and

"**species_scope_list_for_phylogeny_study([*'9598','9606','9544','10116','10090','9601','9615','8090','9031','13616','7719','8364','99883','9593','9103','9913','9796','9823','9258','59729','69293','7955'*])**" are two identical species scopes (identified by taxid) with different formats employed to choose species used during the study. Phylogenies will be built with these species. The default value is that described above (22 species).

"**database(*'Path_database_used'*)**" defines the path of the protein database used. The default path is '../AlgoTools/Blast/db/ensembl'.

B) Parameters defined in the geneloss_event_search agent:

"**nucleotide_in_more_by_side(*10000*)**" is the number of nucleotides taken on each side of a TBLASTN hit, to output a prediction (value must be identical to the genelosses_synthetic_analysis value). The default value is 10000.

"**orthologs_group_mode(*mode('TaxidAncestor')*)**" is the ortholog sequence analysis mode launched. There are two ***mode*** options: ***lineage*** or ***species***. In the publication we speak only about the lineage mode that is appropriate to focus on the lineage-specific losses.

In ***lineage*** mode, GLADX searches the sub-tree having the ***TaxidAncestor*** ancestor as root and containing the reference given as input. All the sequences present in this subtree form an orthologous group. From this orthologous group it deduces the lineage-specific losses comparing the species present in the group to the species-set selected for the study. **Note: An agent allowing analyzing systematically all nodes of the lineage leading to the input reference from the selected ancestor can be activated. => See G) section**

In ***species*** mode, it searches in the phylogeny the species that have orthologs to the reference protein given as input until the TaxidAncestor ancestor and deduces losses comparing species that have an ortholog to the species-set selected for the study. This mode is less appropriate to analyze lineage-specific losses.

The default value is *lineage('117571')* that corresponds to a search of species that have no representative of a gene established at least since the last common ancestor of *Euteleostomi*.

"**do_not_study_when_species_exist([*'9606', '9544'*])**" defines species that will stop the study if an ortholog exists in the first phylogeny. Should be empty if you want to analyse all the species where the gene is missing. If you need to concentrate on losses in a specific species, note its taxid here. If a database-described ortholog already exists for your species in the first phylogeny, there is no need to continue the study (to save your time). By default the value is empty.

"**minimum_size_of_orthologs_group_for_begin_the_study(*3*)**" is the minimum size of an ortholog group required in the first phylogeny to continue the study. The default value is 3.

"**search_missing_cause_in_genome(*choice*)**" defined if you want to use GLADX in complete mode to search for the genome of a species where orthologs are missing in the first phylogeny. ***Choice*** can be **yes** or **no**. If **no** is chosen, no verification of loss is made, and the results output come exclusively from analyses of the first phylogeny built from the chosen database (making the process much faster). The default value is **yes**.

"**translate_in_gene_to_detect_ortholog_if_necessary(*choice*)**" is defined when you have a tree of proteins that you want to translate into genes. This parameter allows comparing two ortholog protein groups using their respective genes. *Choice* can be **yes** or **no**. No is faster but a little less accurate. Indeed, different proteins of a same gene can be present in two trees. If no translation in gene is performed, the both protein will be not found as similar.

"**force_to_analyse_this_species([*'9593', '9606'*])**" This parameter allows to annotate the list of selected species, even if an ortholog is found by phylogeny in the first step. By default the value is empty.

C) Parameters defined in the best_hit_fgx agent:

"**max_nb_managed_hits(*'5'*)**" is the number of hits retained from TBLASTN to continue the analysis. The higher this number is, more the GLADX analysis can be long. Indeed, GLADX tests one by one the orthology of hits. As long as no hit is found orthologous, GLADX continue to test the following hit. If there is no ortholog this step is only limited by the number of hit to be tested. Naturally it is possible that the number of hit found by BLAST can be inferior to the fixed value. The default value is 5.

D) Parameters defined in the genelosses_checkpoint_all_events_by_study agent:

"**length_threshold(50)**" is the minimum overlapping threshold between an orthologous sequence retrieved by GLADX and a known protein in order to continue the study at nucleotide level. The default value is 50.

"**identity_threshold(50)**" is the minimum identity threshold needed between an orthologous sequence retrieved by GLADX and a known protein to continue the study at nucleotide level. The default value is 50.

"**identity_threshold_for_real_gene(70)**" is the minimum identity threshold needed between known protein and used reference protein to be used in study at nucleotide level. The default value is 70.

E) Parameters defined in the genelosses_synthetic_analysis agent:

"**nucleotide_in_more_by_side(*10000*)**" is the number of nucleotides taken on each side of an orthologous gene to build an alignment with orthologs retrieved during the study. It is the step just before the reconstruction (The value must be identical to the geneloss_event_search value). The default value is 10000.

F) Parameters defined in the verify_prediction_existence agent:

When GLADX retrieves an ortholog, it systematically checks the database used to see whether there is an annotation on its position. Sometimes previously-described genes are present on the same area.

"**overlap_threshold(*50*)**" is the minimum overlap threshold in percentage for a previously-described gene in the database to consider that they are on the same position. The default value is 50.

"**identity_threshold_to_conclude_gene_already_exist(*70*)**" is the minimum identity threshold in percentage for a previously-described gene in the database overlapping the GLADX-retrieved ortholog sequence to be considered as the same prediction. The default value is 70.

G) Activation of the *gladx_driver* agent to automate the search of lineage-specific losses on all nodes:

Activation of this agent allows analyzing systematically the lineage-specific losses from all nodes available along the lineage leading to the input reference from the selected ancestor.

"**Targets([*'9606'*])**" is a parameter defining the species concerned by lineage-specific loss, searched by GLADX. It allows focusing the search on the interest species. When no species are specified, GLADX searches all lineage-specific losses along the studied lineage. By default the value is empty.

The activation of agents is defined with these following markups:

```
<master>
   <type>Agent_Name</type>
   ...
</master>
```

By default *gladx_driver* agent is deactivated by comment markups. To activate it, the comment markups of the *gladx_driver* agent must be removing, and the line of the **orthologs_group_mode** parameter of the *geneloss_event_search* agent must be commented.

Note: When new studies are performed with the *gladx_driver* agent, its **orthologs_group_mode(*lineage('TaxidAncestor')*)** parameter is used to define from which ancestor the study begin. While if the *gladx_driver* agent is launched after a first round of analysis with default mode, its **orthologs_group_mode(*lineage('TaxidAncestor')*)** parameter does not used. In this case, all the nodes of the lineage are analyzed from the ancestor that was defined at first round in the **orthologs_group_mode(*lineage('TaxidAncestor')*)** of geneloss_event_search agent.

Annexe 4 : Supplément texte S1 article 2 – Description des paramètres de GLADX

Annexe 5

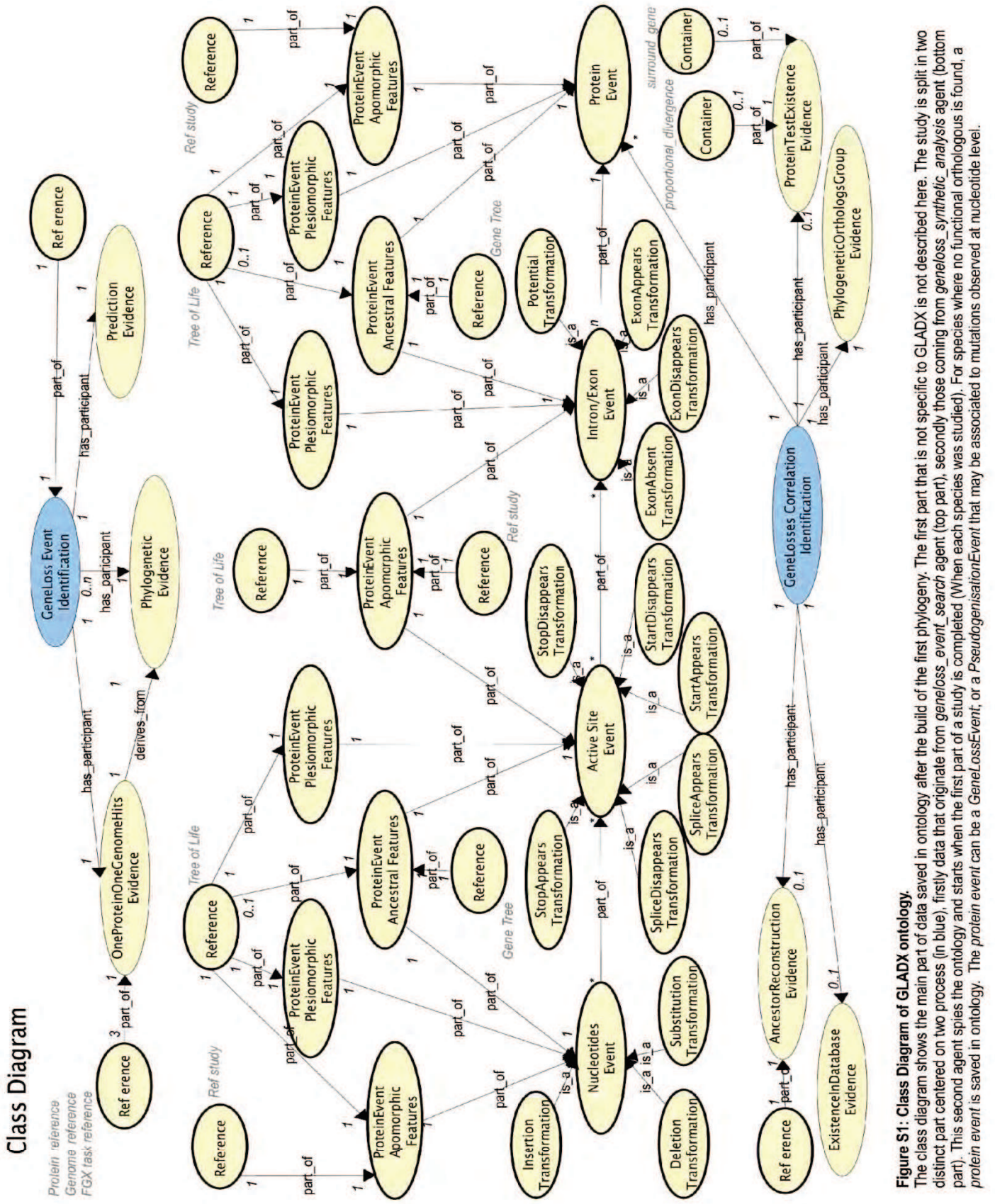


Figure S1: Class Diagram of GLADX ontology.
The class diagram shows the main part of data saved in ontology after the build of the first phylogeny. The first part that is not specific to GLADX is not described here. The study is split in two distinct part centered on two process (in blue), firstly data that originate from *geneloss_event_search* agent (top part), secondly those coming from *geneloss_synthetic_analysis* agent (bottom part). This second agent spies the ontology and starts when the first part of a study is completed (When each species was studied). For species where no functional orthologous is found, a *protein event* is saved in ontology. The *protein event* can be a *GeneLossEvent*, or a *PseudogenisationEvent* that may be associated to mutations observed at nucleotide level.

Annexe 5 : Supplément figure S1 article 2 – Diagramme de classe de l'ontologie de GLADX

Annexe 6

Table S1: Summary of benchmarking results

Symbol	Gene name	Publication results	Species used for GLADX Study	Lineage studied	appea-rance	GLADX results	Artifacts
2310042E22Rik	RIKEN cDNA 2310042E22 gene	(Zhu et al. 2007)(Zhang et al. 2010) Detected as pseudogene in *Homo*.	ENSMUSP0000010 0495	9347	9347 (D)	The gene is saved in *Homo* and *Macaca*, and lost in Sus. This is a pseudogene in *Pan*, where we observe a start codon mutation (ATG to CCG), an acceptor splice site mutation of exon 2 (TA to CA), a 5-bp insertion bringing a nonsense codon by frameshift and directly after, a 5-bp deletion restoring the initial reading frame.	
Gulo	gulonolactone(-L) oxidase	(Nishikimi et al. 1994) A short piece of sequence was retrieved in *Homo*, within indels and nonsense codons. (Zhu et al. 2007) Pseudogenization began before the *Callithrix jacchus* diverged from the human lineage by indels and nonsense codons. (Zhang et al. 2010) Pseudogenization began in the LCA of *Catarrhini*.	ENSMUSP0000006 0912	117571	32523	A gene loss is observed in the LCA of *Catarrhini*. The gene is present at least since the LCA of *Tetrapoda* and seems to be a pseudogene in *Xenopus*. The gene is lost in *Taeniopygia*. Lot of mutation observed at genome level in *Monodelphis* and *Ornithorhynchus* allow concluding on a pseudogenization process in these species.	
Acyl3	Acyltransferase 3	(Zhu et al. 2007) Analysis at genome level shows a nonsense mutation (TGG to TGA) that inactivated Acyl3 after the divergence of *Gorilla* from the *Homo* lineage and before the *Homo-Pan* split. (Zhang et al. 2010) Pseudogenization occurs after the divergence of *Gorilla* from the *Homo* lineage and before the *Homo-Pan* split.	ENSMUSP0000011 0749	7711***	7711	The gene is lost in *Danio*, *Branchiostoma* and in *Neognathae* lineage after the split with the LCA of *Amniota*, and is a pseudogene in *Xenopus*. It is found as a pseudogene in *Homo*, *Pan* and *Pongo* and seems to be intact in *Gorilla*. A splice site acceptor mutation (CT to CC) observed in the LCA of *Hominidae* may be the first event leading to the pseudogenization in *Homo*, *Pan* and *Pongo*. The gene is pseudogenized in *Pongo* independently by 3 nonsense codons that appeared by substitution (CAG to TAG; TAT to TAA; CGA to TGA) and an insertion of 4-bp at the end of the sequence, leading to a nonsense codon TGA by frameshift. A nonsense codon occurred in the LCA of *Homo-Pan* by substitution (TGG to TGA). This last mutation has already been described in the literature.	

		(Wu et al. 1989) They reported that the function of the enzyme has been lost in *Homo* and certain other primates. They observed two nonsense codons found in *Homo* that occurred by substitution (CGA to TGA; AGA to TGA). Pseudogenization seems to appear in *Hominoid* phyla after the divergence from *Old World monkeys*, but it's not clear whether it occurred after a common event or due to independent events. (Oda et al. 2002) They studying the coding region and described a nonsense mutation shared by *Homo*, *Pan*, *Gorilla* and *Pongo* in exon 2 codon 33 that can explain the start of the pseudogenization process in the *Hominidae* phylum. They found two mutations in the LCA of *Homininae*, such as the acceptor signal mutation in exon 3 and a nonsense mutation at codon 187. They found in *Gorilla* a mutation in the initiation codon and a nonsense mutation. The *Pongo* has two independent nonsense mutations and the *Pan* one. In total they found 6 nonsense codons in *Hominidae* phylum. An independent pseudogenization also occurred in the *Gibbon* lineage. Indeed, they found a nonsense mutation,a single-base deletion and a single-base insertion shared by *Hylobates lar*, *Hylobates agilis*, *Hylobates muelleri*, *Hylobates concolor* and *Hylobates syndactylus*. (Zhu et al. 2007) They found a pseudogenization of the Uox gene by nonsense mutation dated to before the *Homo-Pan* split. (Zhang et al. 2010) They found a pseudogenization of Uox gene since the LCA of *Hominidae*.						
Uox	Urate oxidase		ENSMUSP00000029837	117571	117571	All 6 nonsense codons already described in the *Hominidae* phylum and the acceptor site mutation of exon 3 were found. We found also a donor splice site mutation of exon 4 and an acceptor splice site mutation in exon 6 in *Pongo*. We also found a 7-bp insertion in the LCA of *Homininae* and a deletion of this insertion in *Pan*. The explanation for this indel is probably tied to allele sorting. To conclude, we found a pseudogenization of the Uox gene in *Hominidae* phyla, in agreement with other publications.	
Ctf2	Cardiotrophin 2	(Derouet et al. 2004) They reported the gene as intact in *Pan*. *Homo* has an 8-bp deletion in the putative third exon leading to shift and loss of the reading frame. (Zhu et al. 2007) They found an 8-bp deletion engendering a pseudogene in *Homo*. (Zhang et al. 2010) They detected the pseudogenization that occurred only in *Homo*.	ENSPTRP00000054644	117571	32523	The 8-bp deletion is observed and two nonsense codons inducted by this deletion are detected. We retrieve the gene in *Gorilla* and *Macaca*. GLADX was also able to show a gene loss in the *Neognathae* lineage after the split with the LCA of *Amniota*.	

Nradd	neurotrophin receptor associated death domain	(IHGSC 2004) They reported the pseudogene in *Homo* and in *Pan*. The *Homo* pseudogene contains one disruption. (Zhu et al. 2007) They observed a nonsense mutation and an indel that allows dating the start of pseudogenization to the LCA of *Hominidae*. (Zhang et al. 2010) They dated the start of pseudogenization to the LCA of *Hominidae*.	ENSMUSP00000035069	7711***	117571	The gene is lost in *Ornithorhynchus* and *Taeniopygia*. The gene started the pseudogenization in the LCA of *Hominidae* by a previously-described nonsense codon. Afterwards, there are independert mutations in *Pongo*, *Gorilla*, and Pan. In *Pongo* we observe two substitutions leading to nonsense codons (CGA to TGA; CAA to TAA). In *Gorilla,* there is a 1-bp deletion leading to the appearance of four nonsense codons by frameshift. In addition, we found a fifth nonsense codon exists, that appeared by addition of the frameshift and a substitution (TGG to TGA). In *Pan* there is a donor splice site mutation (GT to GC) in exon 2. No other mutations were found in *Homo* other than that found in the LCA of *Hominidae.*	* ENSGGOP00000028206 (intron length 4, 2, 12)
Nepn	nephrocan	(IHGSC 2004) They reported the pseudogene in *Homo* and *Pan*. The *Homo* pseudogene contains four disruptions. (Zhu et al. 2007) They found a nonsense mutation and an indel that allow dating the pseudogenization to the LCA of *Catarrhini*. (Zhang et al. 2010) They dated the start of pseudogenization to the LCA of *Catarrhini*.	ENSMUSP00000070130	7711***	7735	The gene is lost in *Clupeocephala* lineage after the split with the LCA of *Euteleostomi* and is also lost in *Bos*. GLADX save the gene in *Branchiostoma*. In the human lineage, the pseudogenization process began at least to the LCA of *Catarrhini*. We cannot observe the pseudogenization in *Macaca* as the gene is too extensively pseudogenized and is considered lost by GLADX. The gene composed of three exons was intact in the LCA of *Eutheria*, and we can see mutations from the LCA of *Hominidae*. The mutations observed in the LCA of *Hominidae* against the LCA of *Catarrhini* are two deletions (5 bp and 2 bp) in exon 2, a 1-bp insertion at the end of the sequence, 30 nonsense codons due to frameshifts, six of which are also linked to a substitution event. The indels engender a new reading frame for all the descendant species. In *Pongo* we observe the loss of the first and last exon. In the remaining exon there is a 1-bp deletion, which adds to the other frameshifts appearing in the LCA of *Hominidae*. There are six nonsense codons present in *Pongo*, all linked to the frameshifts, and among them two are also linked to substitution events (TAT to TAA; CGA to TGA). In the LCA of *Homininae* there are six nonsense mutations due to substitution events (TTA to TAA; TGG to TAG (*2) ; GAA to TAA; CAG to TAG; GGA to TGA). In *Gorilla* there are several indels in exon 2: two deletions (9 bp; 3 bp) and a 9-bp insertion that engender no frameshifts, and a 7-bp insertion that adds to the other frameshifts appearing in the LCA of *Hominidae*. The frameshifts and insertions give birth to two nonsense codons (--- to TAA (*2)). In *Pan* we found one nonsense codon by substitution (CGA to TGA).	* ENSGGOP00000023229 (intron length 1)
Mup4	Major unitary protein 4	(Chamero et al. 2007) They reported that this is a unitary pseudogene only in *Homo*. (Zhang et al. 2010) They found a mutation at the splicing donor (GT to AT) of exon2 of the *Homo* pseudogene. It is a pseudogene in *Homo*, and they observed an acceleration of the non-synonymous substitution rate in primates.	ENSMUSP00000095648	7711***	9347	The pseudogenization process began at least to the LCA of *Catarrhini*, by appearance of a nonsense codon (TAG) in the reading frame. No other mutations appear in *Macaca*. No harmful mutation appears in the LCA of *Hominidae*. In *Pongo,* a nonsense codon appears by substitution (CGA to TGA) in the penultimate exon. No harmful mutation appears in the LCA of *Homininae*. In *Gorilla,* there is a 1-bp insertion that induces seven nonsense codons. No new mutations appear in the LCA of *Homo-Pan*, or in *Pan*. On the other hand, a 41-bp insertion appears in *Homo*. The long insertion brings 3 new nonsense codons. We also found	* ENSGGOP00000020262 (intron length 1)

							the previously-described donor splice site mutation of exon 2.
T2r2	Bitter taste receptor T2R2	(Go et al. 2005) They claimed the gene is polymorphic in terms of two-base deletion at codon position 160 in the *Homo* population.	ENSPTRP0000005 4801	117571	32523	The gene exists at least since the LCA of *Tetrapoda*, and was pseudogenized in the *Phasianidae* lineage after the split with the LCA of *Neognathae*, in *Sus* and *Homo* and was lost in *Gallus* and *Macaca*. Automated analysis at *Homo* genome level shows a pseudogenization due to a 2-bp deletion already described in other publications. This deletion brings six nonsense codons. GLADX also retrieved the gene in *Mus*, *Bos* and *Equus*, which have no harmful mutations.	
Tas2R134	taste receptor, type 2, member 134	(Go et al. 2005) They reported it as a pseudogene in *Homo* due to two nonsense mutations (CAG to TAG and GAA to TAA). (Zhang et al. 2010) They found it pseudogenized in *Homo* only.	ENSPTRP0000005 4241	9347	9347 (D)	The gene appeared in the LCA of *Eutheria* and only became a pseudogene in *Homo* due to the two nonsense mutations already described in other publications.	
1110012 D08Rik	RIKEN cDNA 1110012D08 gene	(Zhang et al. 2010) They detected the pseudogene in *Homo*.	ENSMUSP0000005 0451	9254	32524 (D)	Pseudogenization began in the LCA of *Catarrhini* by a 1-bp insertion event that induced 8 nonsense codons, one of which is also linked to a substitution event (CAC to TAA). In *Macaca* there is one mutation in an initiation codon (ATG to CTG) and a 10-bp deletion engendering seven nonsense codons, 4 of which are also linked to substitution events (TTA to TGA; CAG to TAG (*2); TAC to TAA). In the LCA of *Hominidae* there is no mutation. In *Pongo* there is a donor splice site mutation (GT to AT) and an initiation codon mutation (ATG to GTG). One nonsense codon appeared by substitution in the LCA of *Homininae* * (CAA to TAA). In *Gorilla* there is a 1-bp deletion that induces one nonsense codon. In the LCA of *Homo-Pan* and in *Homo* there is no new harmful mutation. In *Pan* there is a 1-bp deletion that induces one nonsense codon.	ENSGGOP000 00021588 (introns length 4 and 11)
Gpr33	G protein-coupled receptor 33	(IHGSC 2004) They reported a pseudogene in *Homo* by a disruptive element, and also a pseudogene in *Pan*. (Zhu et al. 2007) They reported a pseudogene due to a nonsense codon that occurred only in *Homo*. (Zhang et al. 2010) They reported a duplicated pseudogene in *Homo*.	ENSPTRP0000005 4852	32523	32523 (D)	The gene is pseudogenized in *Homo* by a nonsense codon due to a substitution (CGA to TGA). The gene is also pseudogenized in *Pongo* by a nonsense codon that appeared by substitution (CAT to TAA). The gene seems to be lost in *Taeniopygia*, *Gallus*, *Monodelphis* and *Equus*. There is also a pseudogenization in Sus. Finally, the gene was saved in *Ornithorhynchus*.	

							The gene is saved in *Oryzias*, pseudogenized in *Branchiostoma* and * *Catarrhini* phylum, and lost in *Xenopus*, *Neognathae* phylum, ENSGGOP000 *Ornithorhynchus*, *Equus*, *Gasterosteus* and *Danio*. Pseudogenization of 00022534 the gene began in the LCA of *Catarrhini* by two deletions (1 bp * 2) and (intons length: two insertions (8 bp; 319 bp). There is also an acceptor splice site 1, 1, 14, 2) mutation in the third exon (AG to CG) and nine nonsense codons. ENSMMUP00 Among them, one is present in the inserted sequence, and two are 000029429 linked to substitution events (CAG to TAG; CGA to TGA). In *Macaca*, (exon length: the first exon disappeared; we also found one acceptor splice site 5) mutation (CG to CA) in exon three, and no indels that induce nonsense The study was codons. The LCA of *Hominidae* cannot be studied due to the relaunched reconstruction profile built by the "Ortheus" tool. In *Pongo* there is one with weak exon disappearance, two deletions (1 bp; 4 bp), and 21 nonsense threshold codons, 5 of which are linked to a substitution events (CGA to TGA; allowing to TGG to TGA; TGC to TGA; TGG to TAG; TAC to TAG). In the LCA analyzed of *Homininae* there is one donor splice site mutation in the first exon *Macaca* at (GT to AT), a 5-bp deletion also in the first exon, and 14 nonsense nucleotide codons. In *Gorilla* there is a 22-bp insertion that induces seven level. nonsense codons and one acceptor splice site mutation in the third exon ** Error in (CG to TG). The LCA of *Homo–Pan* contains no mutation. In *Pan* there reconstruction is a 1-bp deletion. In *Homo* there is a 1-bp insertion that induces three that induces nonsense codons and one acceptor splice site mutation (CG to TG), as we cannot described in *Gorilla*. These common splice mutations between *Homo* and observe the *Gorilla* may stem from allele sorting. first deletion.
Slc7a15	solute carrier family 7 (cationic amino acid transporter, y+ system), member 15	(IHGSC 2004) They reported pseudogene in *Homo* and in *Pan*. The *Homo* pseudogene contains two undescribed disruptions. (Zhu et al. 2007) They observed that the pseudogenization began before the *Callithrix jacchus* and *Catarrhini* lineage split by a nonsense codon. (Zhang et al. 2010) They found disruptive mutations that gave rise to *Homo* unitary pseudogenes dated to before the *Callithrix jacchus* and *Catarrhini* lineage split.	ENSMUSP0000009 3548	7735***	7735		** Error in alignment at the end of exons
Sult3a1	sulfotransferase family 3A	(Freimuth et al. 2004) They observed four nonsense codons and a frameshift in the *Homo* pseudogene.	ENSMUSP0000009 0259	117571	117571		: Exon 2 it miss one base, at exon 3 and The pseudogenization process began at least in the LCA of *Catarrhini*. 4 it miss 2 We cannot observe the pseudogenization in *Macaca*, since although the bases. sequence is present in the genome, it is too pseudogenized to pass the identity threshold and thus be scanned at genome level. In the LCA of Frame *Hominidae*, we found a 23-bp insertion and 10 nonsense codons, two of scanned not which are due to a substitution events (TTA to TGA; ATA to TGA). In always well. It *Pongo* we found two nonsense codons by substitution (TGG to TGA; engenders AGA to TGA). There is no mutation in the LCA of *Homininae*. *Gorilla* observation of has one nonsense codon that appeared by substitution (TGG to TGA) some mutation and is similar to that observed in *Pongo* but which seems to have not implicated appeared independently. There are no new mutations in the LCA of in *Homo–Pan* or in *Homo*. In *Pan* we found a nonsense codon that pseudogenizati appeared by substitution (CCA to TAG). on processes.

(D) = This symbol in “appearance” columns means that the orthologous group studied is occurred by a duplication which has been observed in phylogeny.
* = Some protein coding genes are described as intact in the Ensembl database but after manual inspection we observed that they seems to be over predictions. In these cases, the protein causing the problem was cited, and the study was relaunched with a parameter-set making it possible to avoid considering the protein as present. That allows re-annotating the gene by GLADX.
** = indicates tool-induced artifacts. An explanation is given for the kinds of problem engendered.
*** = In these cases we have implemented specifically in GLADX the capacity to use the *Branchiostoma floriade* species; this engenders the presence of a ancestral node with the taxid 7735.

Note: Some indels are not described here as they are multiples of three and do not seem to be essential to the pseudogenization analysis.

Annexe 6 : Supplément tableau S1 article 2 – Résumé des résultats du benchmark

Annexe 7

Text S2: Analyses of artifacts:

Through our analyses using GLADX, we highlighted three types of artifacts: errors in genomic sequences, prediction errors, and errors due to the limits of the tools used.

Detection and analysis of mis-predictions and over-predictions present in the used database:

Here we will explain prediction errors, which appear to be the event the most widely encountered, and which we have not yet tackled. Among the 14 benchmarked cases of pseudogenes, the first round of GLADX analysis found 6 non-parsimonious cases in terms of number events, kinds of events, and location on the evolutionary timeline. In these cases, Homo, Pan, Pongo and sometimes *Macaca* had detected pseudogenes, but a gene was detected as present in *Gorilla* (*Acyl3*, *Nradd*, *Nepn*, *Mup4*, *1110012D08Rik*, *Slc7a15*). Furthermore, 3 of these cases do not fit with published results, where Macaca gene (Slc7a15) and Gorilla genes (Nradd, Nepn, Slc7a15) should be pseudogene. After manual verification in the Ensembl database, we observed that small introns with less than 13 bases are present in 5 genes (*Nradd*, *Nepn*, *Mup4*, *1110012D08Rik*, *Slc7a15*). Very few introns are known to be less than 20 bases in length. The smallest intron found in protist genes was 13-20 bases long [1]. The minimum CDS intron size observed in 2903 genes from 10 eukaryotes was 13 bases [2]. The minimum intron size in ESTs in a collection of fungi was 27 bases [3]. In the case of the Acyl3 case, there is just suspected short 8-base and 4-base-long exons in Gorilla. In the case of the Slc7a15 gene, there are suspected short exons in Macaca. There is some concern that very short exons might not be real exons. Short exons or introns may be the result of annotation correcting for apparent frameshifts.

This evidence led us to suspect that these genes were over-predicted, with varying degrees of certainty according to clues.

To verify these suspicions, we relaunched the analysis for each of the 6 cases, with a parameter that, depending on the case, allowed us to consider the *Gorilla* and/or *Macaca* gene as absent from the database. Using this parameter, found that a potential mutation may have occurred in these species, as well as in ancestors. The results obtained on the 5 genes with intron size problems show more parsimonious scenarios with a common mutation existing in ancestors. These results may point to the mis-prediction of introns in these *Gorilla* genes. In the case of *Slc7a15* with suspect exon size in the *Macaca* gene, we again found a more parsimonious scenario, and confirmed the pseudogenization.

Based on these new results, we agree with the 3 cases of *Gorilla* and *Macaca* pseudogenization already described. Furthermore, in two cases (*Mup4, 1110012D08Rik)*, the pseudogenization process was found to be older.

The analysis of the *Acyl3* gene represents a particular case. Indeed, no pseudogenization has been reported in the *Gorilla* gene, and the gene described on Ensembl (ENSGGOP00000028123) showed no sequence problems except exons that were relatively short, at 8 bases and 5 bases long. Relaunching *Acyl3* with *Gorilla* considered as absent from the database came up with the same result as in the previous study. *Homo*, *Pan* and *Pongo* are found as pseudogenes due to mutation. The nonsense codon that occurred by substitution was always present in the ancestor before the split between *Homo* and *Pan* and after *Gorilla* diverged from the human lineage. The differences in this relaunched study are that the LCAs of *Hominidae* and *Homininae* are scanned at genome level, as the functionality of the genes in these ancestors is considered as unknown. In the LCA of *Hominidae*, we observed an acceptor splice site mutation, but the gene is found as functional in *Gorilla* as there is no existing harmful mutation. No further mutation is shown in the gene of the LCA of *Homininae*. The *Gorilla* gene is consequently found as potentially functional (noted as saved by the tool) as there is no harmful mutation in the *Gorilla* sequence or its ancestral sequences (*Hominidae* and *Homininae*). The splice site mutation in the LCA of *Hominidae* seems to be the first event leading to pseudogenization. Our results are in agreement with Ensembl and with previously published results [4] about functionality of the Gorilla gene. This splice site mutation in the LCA of *Hominidae* may have triggered the pseudogenization process in *Homo, Pan* and *Pongo*, but *Gorilla* seems to have found a way to keep its gene intact. Our results show several missing exons in *Gorilla*, some of which seem to be in sequencing gaps. The results exposed need to be treated with caution, as there may be mutation in these missing exons.

Problem encountered with reconstruction of ancestral sequences:

The GLADX analysis of *Slc7a15* highlights some nonsense codons, with no substitution or frameshift events observed. We searched for an explanation by manually running an analysis of the ancestral sequence reconstruction product and its alignment. We found one deletion of 1 base at the start of the primates genes that was absent in genes of other species. The gene in the LCA of *Eutheria* should have the gene without this deletion, but there is an indel error in the reconstruction, and the deletion has been ascended in the LCA of *Eutheria*. Consequently, when we compared the gene of the LCA of *Catarrhini* against that of the LCA of *Eutheria*, the deletion event was not seen. GLADX cannot observe these indels but is able to observe the event engendered. We also observed that a substitution leading to the first nonsense codon (linked to the frameshift) also went undetected. Indeed, focusing on the indel, a nonsense codon normally only present in primates was ascended in the *Eutheria* ancestral sequence reconstructed. The codon concerned in *Canis* is the CAG in position 17. It becomes a TGA in primates. *Slc7a15* is an example that underlines the limits of

tools integrated in GLADX. Here it engenders no error, but some information is missed.

1. Russell CB, Fraga D, Hinrichsen RD (1994). Extremely short 20–33 nucleotide introns are the standard length in Paramecium tetraurelia. *Nucleic Acids Research*, *22*(7), 1221-1225.
2. Deutsch M, Long M (1999). Intron-exon structures of eukaryotic model organisms. Nucleic Acids Research, 27(15), 3219-28.
3. Kupfer DM, Buchanan KL, Lai H, Zhu H, Dyer DW et al. (2004). Introns and splicing elements of five diverse fungi. *Eukaryotic Cell*, *3*(5), 1088-1100.
4. Zhu J, Sanborn JZ, Diekhans M, Lowe CB, Pringle TH et al (2007). Comparative genomics search for losses of long-established genes on the human lineage. PLoS Computational Biology, 3(12), e247.

Annexe 7 : Supplément texte S2 article 2 – Analyse des artefacts

Annexe 8

Gene Loss Analyzer DAGOBAH eXtension (GLADX) User's Manual

EBM (Evolutionary Biology and Modeling) Laboratory - UMR 7353

Aix-Marseille University - France

Download the (VirtualBox) GLADX image... (~17Go)

Link to benchmark analysis of 14 reported cases (Currently only available for identified users)

The current GLADX version enables using 22 species from Ensembl v57: Bos taurus, Canis familiaris, Ciona intestinalis;, Danio rerio, Equus caballus, Gasterosteus aculeatus, Gallus Gallus, Gorilla gorilla, Homo Sapien, Monodelphis Domestica, Meleagris Gallopavo, Macaca Mulatta, Mus Musculus, Ornithorhynchus anatinus, Oryzias latipes, Pan Troglodytes, Pongo Pygmaeus Abelii, Rattus norvegicus, Sus scrofa, Taeniopygia guttata, Tetraodon nigroviridis, Xenopus Tropicalis.

GLADX corresponds to a set of agents. The sources are freely available and could be retrieved in the */home/tower/TOWER_1.03/prod/DGH_2/src/* **directory.**

The distributed GLADX version could easily be modified (as described in the procedure below) in order to increase the number of studied species. Thus, GLADX can also function with any species present in Ensembl from the version 48 to 58 (included) giving the option to work with 51 different species.

Table of contents

Introduction

GLADX is a module included in a software application: DAGOBAH (Gouret et al., 2011). According to its name (Gene Loss Analyzer DAGOBAH eXtension), it is dedicated to gene losses and pseudogenizations automatic detection and analysis.

DAGOBAH relies on other relevant software tools (FIGENIX (Gouret et al., 2005), PhyloPattern (Gouret et al., 2009), IODA (In press, http://ioda.univ-provence.fr)).

All these components form the lab's bioinformatic software platform, called: T.O.W.E.R (Tools Operating With Evolutive Resources). GLADX work in the TOWER framework.

For us and for external users, TOWER is now very complex to install, because one has to deploy many software components, many bioinformatics binaries, many databases and many genomic data.

So we chose the virtualization strategy, that means the installation of all TOWER's components, on a virtual machine image. Several image instances can be started, as virtual computers, on computers which disposes of a virtualisation software like VirtualBox, VMWare, ... and on Clouds.

Technical requirements

We decided to build an Ubuntu 11.04, 64-bit image on VirtualBox 4.1.2 (Oracle TM). Therefore, this image will work efficiently on 64-bit architecture host computers.

To run one image of TOWER, we recommend using a four cores workstation, with 4Go of RAM (minimal configuration). Our image is configured, as a default, to run with eight cores and with 8 Go (current workstation producing our tests).

Please, note that hardware virtualization technology has to be activated on the host computers. (VT-X for Intel, AMD-V for AMD) in order to obtain most advantageous performances.

Warning: the hyperthreading technology with OpenMPI, that is a software layer used to exploit parallel computing with bioinformatics softwares like Tree-Puzzle, ClustalW, is not recommended because of reported bugs.

Note: the eight cores are not strictly required for the image, users could modify the scripts as below:

- in /home/tower/TOWER_1.03/prod/FGX_API/scripts/puzzle_cmd_perl, change “-np 8” by “-np X”, where X is the number of cores you want to use
- same procedure for: /home/tower/TOWER_1.03/prod/FGX_API/scripts/clustalw_cmd
- same procedure for: /home/tower/TOWER_1.03/prod/FGX_API/pipelines/Templates/__CassiopePhylo+M__, replace “-a 8” in block:

```
<nodeRef>blast</nodeRef>
<parameterAssignation>
<parameterName>options</parameterName>
<parameterValue>-a 8</parameterValue>
```

About the network configuration of the image, the NAT mode was set as a default. This mode doesn't allow 'ssh' access but it is very much faster than Bridge mode.

Images can be run with or without X Window GUI (quite slow in the emulation). In NAT mode, RDP clients can be used to access to a non graphical image. In Bridge mode, one can use ssh.

Important : the *tower* user has *t0wer* as password in ssh or graphical mode.

GLADX image is downloadable on the following link: GLADX image. First uncompress it, then add it to VirtualBox with the GUI or with "vboxmanage -registervm …." command.

GLADX launch

GLADX is started on boot of the image (with VirtualBox). In order to start a gene study with GLADX, one just has to deposit one or several FASTA files (amino acid sequences) in the following directory: */home/tower/GLADX_DATA*

The FASTA files require to be named as follows: *EnsemblProteinSequenceName.Taxid.fasta*

This file must contain a sequence in FASTA format with an header in the following format:

>lcl|EnsemblProteinSequenceName|Taxid|Species~Name|OptionalyADescription

corresponding in this actual example:

>lcl|ENSP00000375415|9606|HOMO~SAPIENS|

A golden dataset of 14 FASTA files corresponding to the cases reported at http://ioda.univ-provence.fr/ is available in the directory /home/tower/Examples/Fastas.

Additional options:

Users can deactivate automatic start of GLADX on boot of the image by commenting with a '#' the line 'su tower -c /home/tower/TOWER_1.03/prod/DGH_2/start' of the file '/etc/rc.local'.

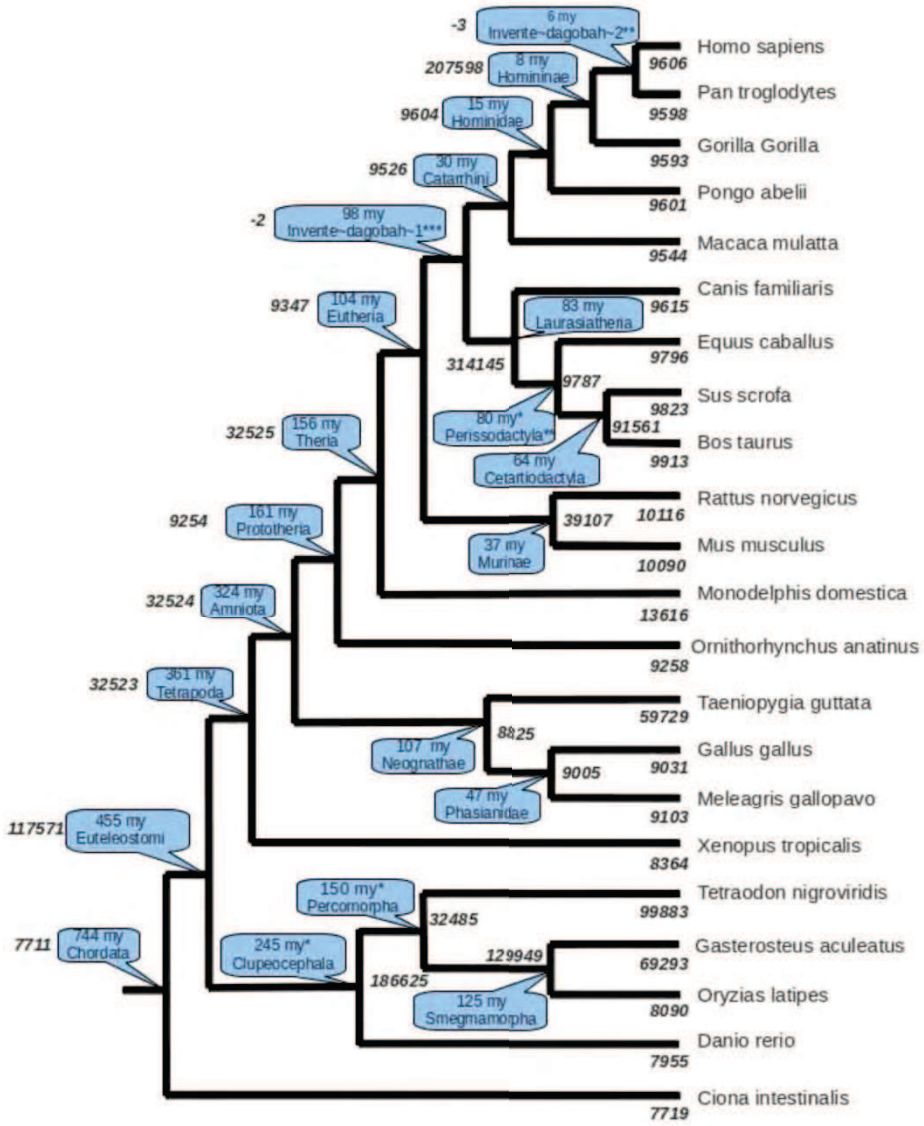
In this configuration you need to launch DAGOBAH using the command 'start' in a Terminal from the current directory */home/tower/TOWER_1.03/prod/DGH_2*.

To stop DAGOBAH you just need to press 'CTRL+C' or alternatively to kill the process.

Choice of phylum and species studied

The default parameters of GLADX allow to analyze lineage-specific gene losses in Euteleostomi (or from the closest ancestor in leaves direction whether the gene is appeared later) by studying the orthologous group containing the protein reference given as input. By default 22 chordates species are used with the topology described below:

Tree of life of 22 species implemented in GLADX



Date of speciations: They come from *TimeTree* (http://www.timetree.org/) and are displayed by million of years (my). When the date was not available it is indicates by "*" and we have chose a coherent date.

Species and ancestor names: The names are taken from *NCBI taxonomy*. When a common ancestor remain unclear and it exists a rake in NCBI, we have noted it by "**". In these case we have decided one topology and chose a name for the ancestors. When an ancestor name does not exist due to a topology incoherent between our choice and the NCBI topology, there is noted "***".

Taxid identifiers: For each leaves and ancestral nodes it exists an unique taxonomic identifier. These identifiers are numbers that are noted close to leaves or ancestral nodes. To follow our topology two ancestor names were invented and consequently no taxid exist, so we have tied to the ancestral species a negative unique identifier.

On these 22 species, by default 21 species of Euteleostomi are studied because the 'orthologs_group_mode' parameter defined in the */home/tower/TOWER_1.03/prod/DGH_2/dagobah.xml* file is parameterized to analyze losses

in Euteleostomi (taxid = 117571) in ***lineage*** mode. However, the analysis of largest phylum such as Chordates (including *Ciona*) is conceivable by using the taxid 7711. In the contrary, smallest phylum could be studied by using the taxid of any ancestor described in the figure 1. The number of species studied in a phylum may be modified by choosing among the 22 species those kept in the scope parameters (*species_scope_for_phylogeny_study* & *species_scope_list_for_phylogeny_study*).

Produced data and results

Results are automatically produced as .report files and databases contents. Report files can be easily read by our "user friendly" viewer FGXView (/home/TOWER/FGXView). The most important result is the final species tree of species-set in which all the results are pinpointed.

Databases contents are of two kinds:

- FIGENIX results produced on a SGBDR PostgreSQL in the database: *figenix_db*

- DAGOBAH results produced as an ontological database (see supplement 1), that relies also on a SGBDR PostgreSQL database: *dagobah_db*.

Note that these databases can be deployed on our IODA web site through collaborations.

1) Manipulate databases (*figenix_db* and *dagobah_db*) in SQL:

To manipulate these databases, please use the following commands in a Terminal:

- To backup the database in SQL format: pg_dump *DatabaseName* -f *SavingFileName*
- To delete a database: dropdb *DatabaseName*
- To create a database: createdb *DatabaseName*
- To install a new database: psql *DatabaseName* -f *DatabaseToInstall*

Note: *DatabaseToInstall* may be a database saved earlier or a clean figenix_db or dagobah_db database available in the directory: /home/tower/Examples/Databases

Warning: There is an incompatibility with the SGBDR PostgreSQL when the version >8.2 is used (we used 8.4). When a new database is created, before database installation you must be connected to the database (as postgres user, “sudo su postgres”, then “psql *DatabaseName*”) and past the text present in the */home/tower/jena_with_postgres_higher_than_8.2* file.

2) Manipulate ontological database (only *dagobah_db*) in OWL:

The ontological results can be exploited by *Protege* software and exported in “.OWL” files.

A script named *clear_with_file* is available at the directory */home/tower/TOWER_1.03/prod/DGH_2.*

It allows to delete a database and to install a new database from an .OWL file. In this case, you need to be in the *DGH_2* directory and use the command in a Terminal like this:

clear_with_file dagobah_model *CompletePathOfTheDatabaseName.owl*

Example to install a *dagobah_db* ontological database empty:

clear_with_file dagobah_model */home/tower/Examples/Databases/dagobah_db_empty.owl*

To backup the *dagobah_db* database in owl without use of *Protege* you need to be in the directory *DGH_2* and launch the following command:

owldump *NameOfBackup.owl*

How to add new species retrieved from Ensembl ?

The current GLADX version enables using 22 species, but more species can be used by some manipulation.

1) Install proteome and/or genome

- Genomes are required when you use GLADX in “complete mode” (parameter 'search_missing_cause_in_genome' in the *dagobah.xml* file). The genomes of species already present are in the directory */home/tower/TOWER_1.03/prod/FGX_API/GenomicDB/ensembl_dna/*. To add new species you need to add the formatted (command *formatdb* in Blast package) genome in this directory. You need also to add the path of the formatted file containing the DNA in the file */home/tower/TOWER_1.03/prod/DGH_2/src/project_specific.pl* like this: “species_dna_database('*Taxid*', '*PathOfTheSpeciesDNAFile*').”

- The proteomes of species already present are in the file */home/tower/TOWER_1.03/prod/FGX_API/GenomicDB/ensembl*. To add new species you need to add the proteome in this file and re-formatted it (command *formatdb* in Blast package).

Note: When you add a new proteomes or/and a new genomes, you need to format the FASTA headers as follows :

>lcl|ENSP00000375415 |9606|HOMO~SAPIENS|

corresponding to

>lcl|*SequenceName* |*Taxid*|*Species~Name*|*OptionalyADescription*

2) Modify the tree topology

The binary species tree defined in GLADX needs to contain the species chosen for analyses.

2.1) Database modifications

The tree topology of species is provided into FIGENIX database (called figenix_db) in the *dagobahtreeoflife* table. The topology is described branch by branch where each taxid is linked to its parent taxid and a description of its rank (*class* if it is an ancestral node, *species* if it is a leaf). An ancestral node must be linked to two taxid corresponding to their child nodes.

Warning: if you add new species that are outgroup of species already present: the farthest ancestor must always be linked to the ghost root taxid *1*.

2.2) Advise the length of branches

The length of branches of the species tree topology is defined in the file */home/tower/TOWER_1.03/prod/DGH_2/src/project_specific.pl* as follows:

"tof_branch_length_to_node('*taxid*','*branch_length*')."

You need to add all the new branch lengths.

Note: When you change the version of Ensembl data you must change the value of the 'maxEnsemblBuildNumber' parameter with the corresponding Ensembl version. This parameter is in the file */home/tower/TOWER_1.03/prod/DGH_2/ENSJHelper.properties*

Particular Case: If you add new species that are outgroup of chordate you need to change the taxid of the far ancestor of the new tree topology in the */home/tower/TOWER_1.03/prod/DGH_2/src/project_specific.pl* file replaced the taxid defined in "*dagobah_treeoflife_database_root*('7711')."

Can I use an other kind of protein database ?

Yes but only in "***simple*** mode". To use the ***simple*** mode you need to modify the *dagobah.xml* file available at this path */home/tower/TOWER_1.03/prod/DGH_2/.*

1. change the mode as described below:

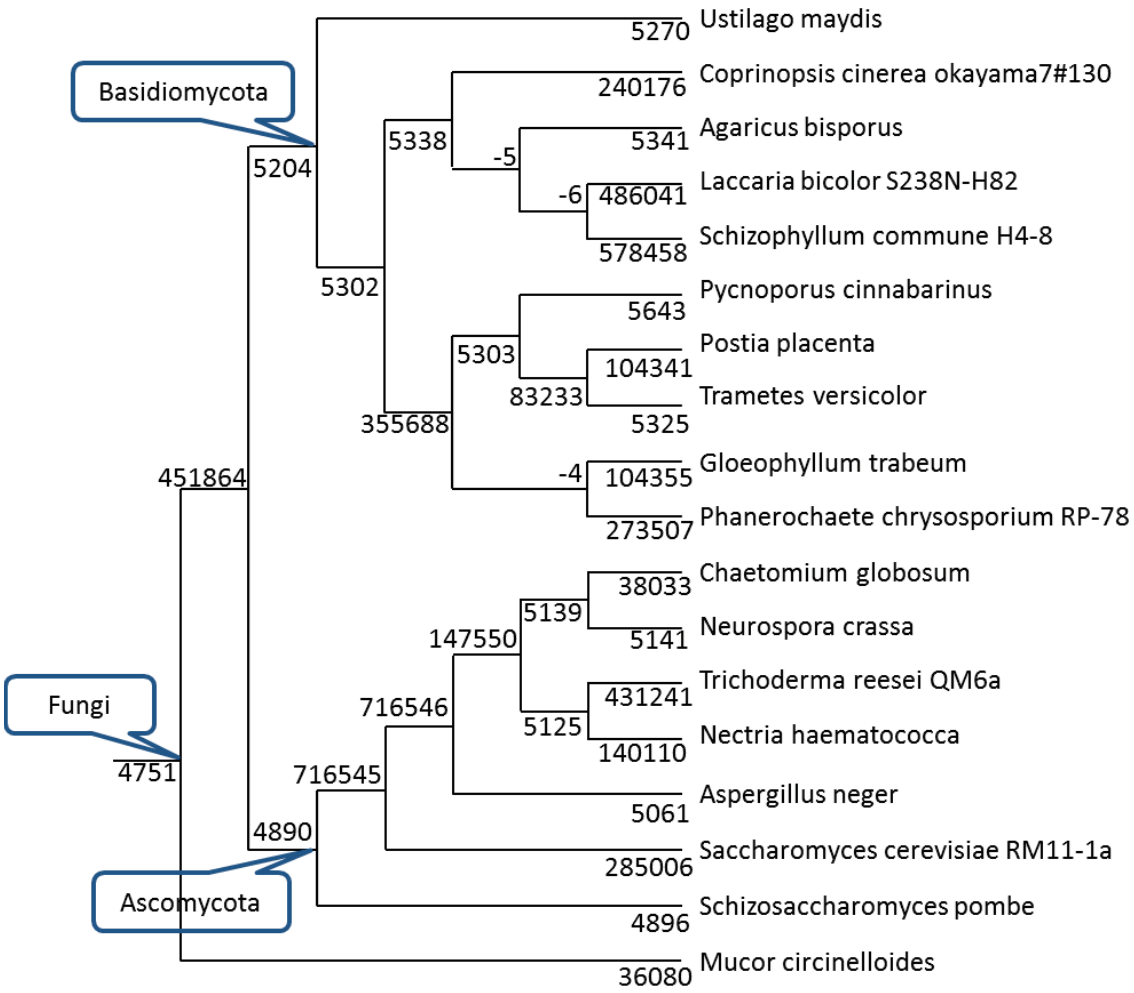
   search_missing_cause_in_genome('no')

2. change the path to your new database

   database('*Path_database_used*')

Annexe 8 : Supplément texte S3 article 2 – Manuel d'utilisateur de GLADX

Annexe 9



Annexe 9 : Arbre des 18 espèces de champignons (Fungi) utilisées

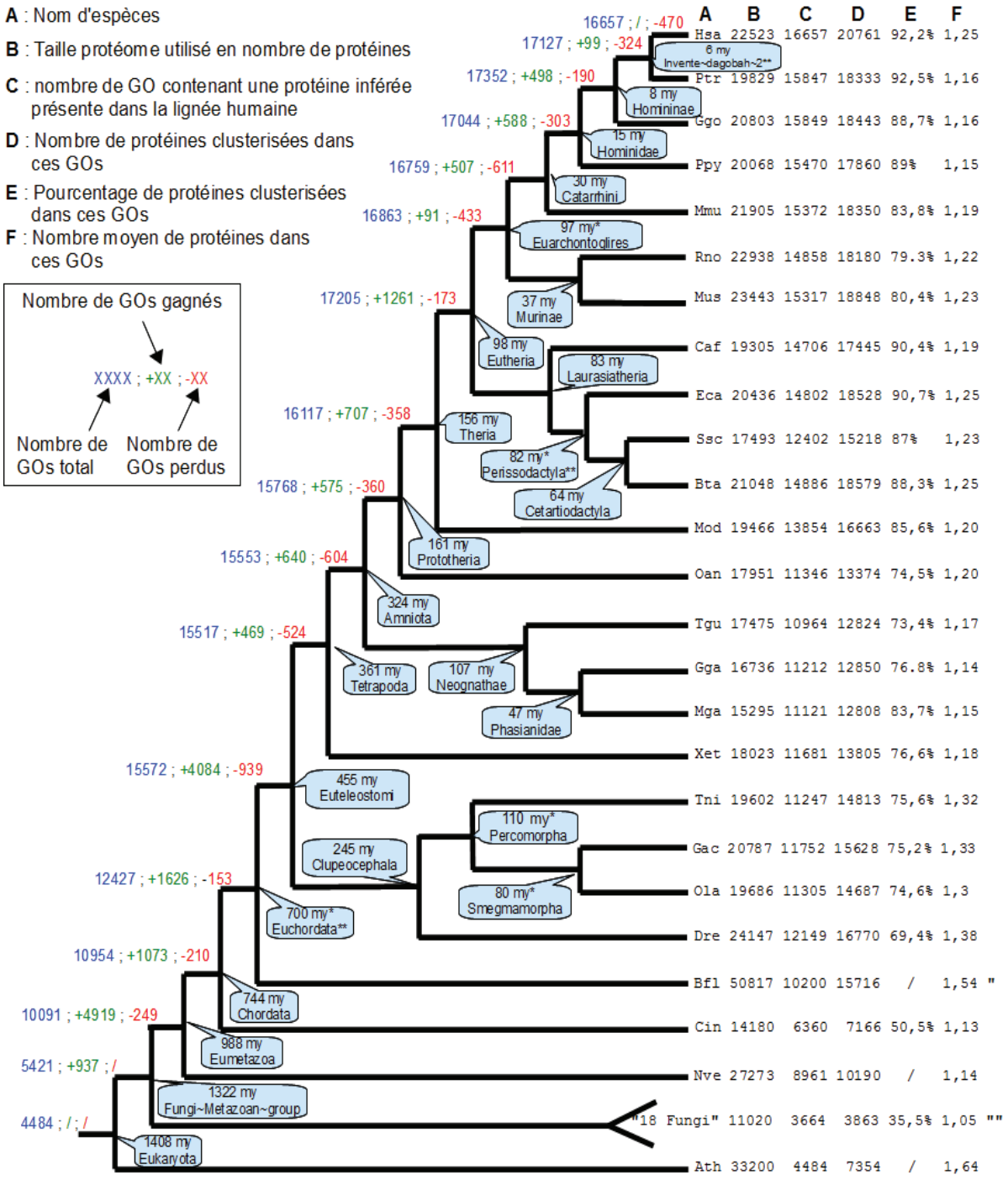
Annexe 10

Espèces	Préfixes utilisés	Base de données utilisées	Assemblage	ADN	version	Nombre de protéines	Nombres de protéines isoformes filtrées	Préfixes identifiants Ensembl
Homo sapiens	Hsa	Ensembl	GRCH37	Chromosomal & non chromosomal Non masqué	57	76592	22523	ENS
Pan troglodytes	Ptr	Ensembl	CHIMP2.1	//	57	34142	19829	ENSTPTR
Gorilla gorilla	Ggo	Ensembl	gorGor3	//	57	27325	20803	ENSGGO
Pongo pygmaeus	Ppy	Ensembl	PPYG2	//	57	23533	20068	ENSPPY
Macaca mulatta	Mmu	Ensembl	MMUL_1	//	57	36384	21905	ENSMMU
Canis familiaris	Caf	Ensembl	BROADD2	//	57	25559	19305	ENSCAF
Equus caballus	Eca	Ensembl	EquCab2	Chromosomal & Non chromosomal ; Non masqué	57	22641	20436	ENSECA
Sus scrofa	Ssc	Ensembl	Sscrofa9	Chromosomal ; Non masqué	57	19083	17493	ENSSSC
Bos taurus	Bta	Ensembl	Btau_4.0	Chromosomal & Non chromosomal ; Non masqué	57	26977	21048	ENSBTA
Mus musculus	Mus	Ensembl	NCBIM37	//	57	50068	23443	ENSMUS
Rattus norvegicus	Rno	Ensembl	RGSC3.4	Chromosomal ; Non masqué	57	32971	22938	ENSRNO
Monodelphis domestica	Mod	Ensembl	BROADO5	Chromosomal ; Non masqué	57	32541	19466	ENSMOD
Ornithorhynchus anatinus	Oan	Ensembl	OANA5	Chromosomal & Non chromosomal ; Non masqué	57	26836	17951	ENSOAN
Taeniopygia guttata	Tgu	Ensembl	TaeGut3.2.4	//	57	18191	17475	ENSTGU
Gallus gallus	Gga	Ensembl	WASHUC2	//	57	22194	16736	ENSGAL
Meleagris gallopavo	Mga	Ensembl	UMD2	Chromosomal ; Non masqué	57	17210	15295	ENSMGA
Xenopus tropicalis	Xet	Ensembl	JGI4.1	Non chromosomal ; Non masqué	57	27710	18023	ENSXET

Tetraodon nigroviridis	Tni	Ensembl	TETRAODON8	Chromosomal & Non chromosomal ; Non masqué	57	23118	19602	ENSTNI
Gasterosteus aculeatus	Gac	Ensembl	BROADS1	Non chromosomal ; Non masqué	57	27576	20787	ENSGAC
Oryzias latipes	Ola	Ensembl	MEDAKA1	Chromosomal & non chromosomal ; Non masqué	57	24661	19686	ENSORL
Danio_rerio	Dre	Ensembl	Zv8	//	57	28630	24147	ENSDAR
Branchiostoma floridae	Bfl	JGI	Brafl1	non masqué		50817	50817	/
Ciona intestinalis	Cin	Ensembl	JGI2	//	57	19858	14180	ENSCIN
Nematostella vectensis	Nve	JGI	/	/	1	27273	27273	/
Fungi *	Fungi	JGI	/	/	/	11020	11020	/
Arabidopsis Thaliana	Ath	JGI	/	/	9	33200	33200	/
							Total : 555449	

Annexe 10 : Informations sur le génome et le protéome des 26 espèces utilisées

Annexe 11



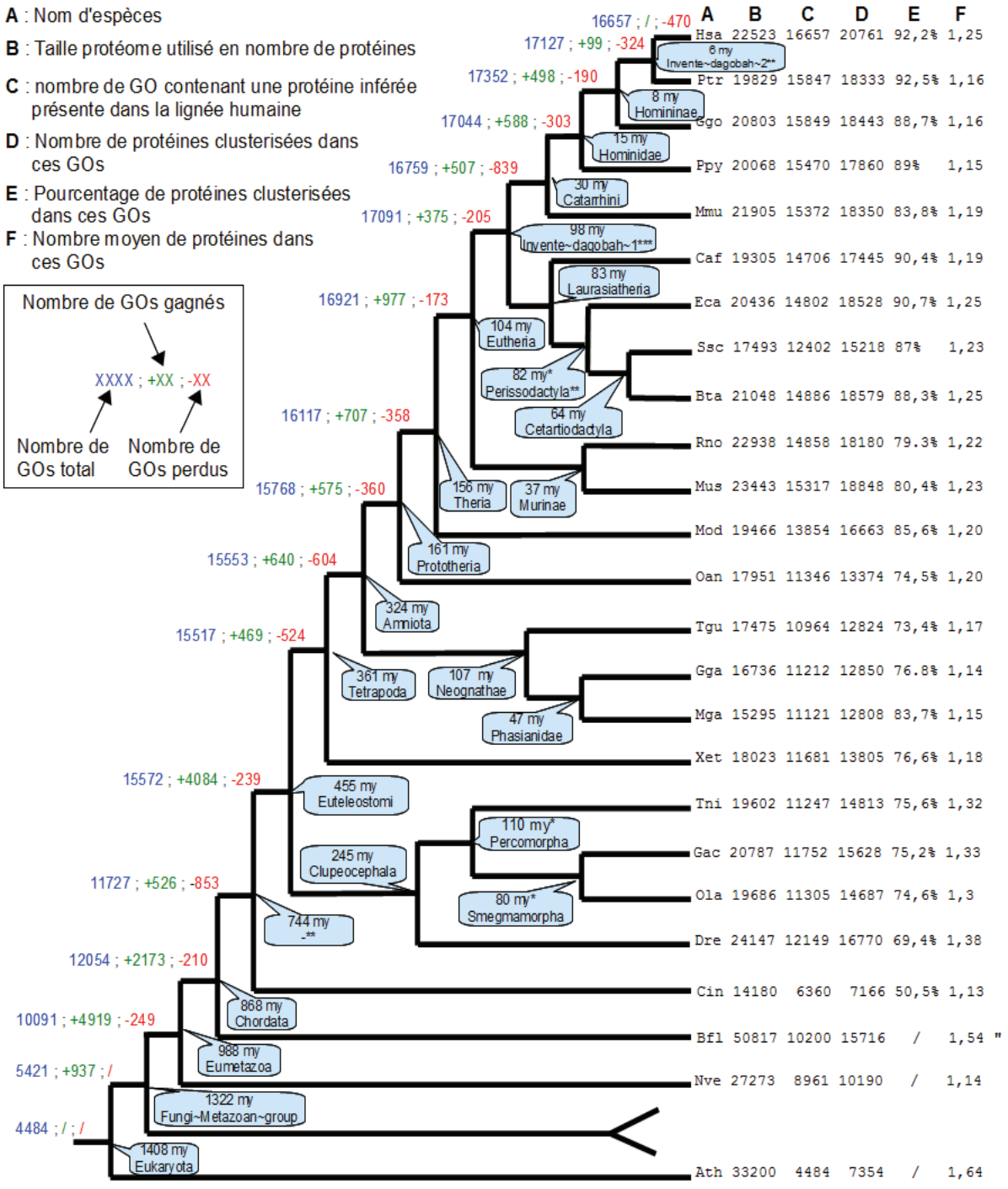
Annexe 11 : Inférence des 22558 GOs sur l'arbre des espèces avec la topologie ciblant les Laurasathériens et les Murinés inversée

La topologie présente est différente de celle utilisée actuellement comme référence. Ici le nœud ancestral de la lignée des Laurasathériens et des Murinés avec la lignée humaine a été inversé. Pour la description de la topologie se référer à l'Illustration 27.

"Les isoformes n'ont pas été filtrés

""Le pan protéome de Fungi a été créé par OrthoMCL à partir de 18 protéomes d'espèces de Fungi. Ce pan protéome a été défini en prenant une protéine de référence par GO construit

Annexe 12



Annexe 12 : Inférence des 22558 GOs sur l'arbre des espèces avec la topologie ciblant C. intestinalis et B. floridae inversée

La topologie présentée est différente de celle utilisée actuellement comme référence. Le nœud ancestral de la lignée menant à C. intestinalis *et celle menant à* B. floridae *avec la lignée humaine a été inversé. Pour la description de la topologie se référer à l'Illustration 27.*

"Les isoformes n'ont pas été filtrés

""Le pan protéome de Fungi a été créé par OrthoMCL à partir de 18 protéomes d'espèces de Fungi. Ce pan protéome a été défini en prenant une protéine de référence par GO construit

Annexe 13

Reference protein used by GLADX study	Saved gene phylum	Observed mutations	Annotation in Ensembl 57 (BD utilisé par GLADX)	Annotation in recent DB Ensembl 61	GLADX vs recent DB Ensembl 61	BLAST of GLADX signal on transcripts DB from NCBI	BLAST of GLADX signal on EST DB from NCBI	Synthesis of results
ENSPTRP00000057518	9606 sc	no	PT: ENSG00000215113	PT: ENSG00000215113	-	NM_001145140.1 (100%) Product: hypothetical protein LOC100130361"	BX283792.1 (65%) Product: cDNA clone	**Pr~**
ENSRNOP00000052333	-3 sc	no	PT: ENSG00000204663	PT: ENSG00000204663	-	XM_939505.4 (100%) hypothetical protein Product: hypothetical protein LOC285556	DA762306 20% Product: cDNA clone	**Pr~**
ENSRNOP00000047640	9604 sc	3 bp deletion	P: ENSG00000229354	P: ENSG00000229354	-	no similarity found	CD557272.1. (100%) Product: cDNA clone	**Tr**
ENSPTRP00000059703	9606 sc	no	UP: ENSG00000240939	PC: ENSG00000253309 UP: ENSG00000215461	+	NM_001101320.1 90% Product: serpin E3 precursor	BM726878.1 (34%) Product: cDNA clone	**Pr**
ENSPTRP00000058392	9606 sc	no	PT: ENSG00000182183	PC: ENSG00000182183 PT: ENSG00000254368	+	NM_001042693.1 (100%) Product: membrane protein FAM159A	AW170404.1 73% Product: tumor Homo sapiens cDNA clone	**Pr**
ENSPTRP00000057227	9606 sc	two mutated splice sites	PP: ENSG00000239290	PP: ENSG00000239290	-	NR_027442.1 (14%) Product: betaGal beta-1,3-N-acetylglucosaminyltransferase 5 pseudogene"	BP872636.1 62% Product cDNA clone	**Tr**
ENSPTRP00000056568	9606 sc	no	PT: ENSG00000214481	PT: ENSG00000214481	-	no similarity found	AI760105 27% Product: Kid11 cdna clone	**-**
ENSPTRP00000056308	9606 sc	no	P: ENSG00000235883	PC: ENSG00000255302 P: ENSG00000235883	+	NM_014335.2 (100%) Product: EID1	DB454504.2 et DB294570.1 (100%) Product: testis cDNA clone H013065M07 et cDNA clone Y79AA100201	**Pr**
ENSPTRP00000056108	9606 sc	no	PT: ENSG00000204623	PT: ENSG00000204623	-	AF032110.1 Product: putative open reading frames can be found in the sequence, but there is no protein data yet for this gene	BI827073.1 (35%) Product: IH_MGC_119 cDNA clone	**Tr**
ENSPTRP00000055033	9606 sc	no	P: ENSG00000179449	P: ENSG00000179449 PC: ENSG00000254585	+	NM_019066.4 (100%) Product: MAGE-like protein 2	AL528236 (85%) Product: NEUROBLASTOMA COT 25-NORMALIZED cDNA clone	**Pr**
ENSPTRP00000054836	9606 sc	no	PT: ENSG00000203772	PC: ENSG00000203772	+	NM_001012508.3 (100%) Product: shadow of prion protein precursor	BG702476.1 (64%) Product: cdna clone	**Pr**
ENSPTRP00000052322	9606 sc	one mutated splice site	PT: ENSG00000242054	PT: ENSG00000204663	-	NR_001279.2 (pseudogène) Product: cystatin pseudogene et cystatin pseudogene (cdna clone)	BG772612.1 (50%) Product: NIH_MGC_97 cDNA clone	**Tr**
ENSPTRP00000051722	9606 sc	no	PT: ENSG00000203759	PT: ENSG00000203759 PC: ENSG00000254726	+	NM_001093725 (100%) Product: RNA-binding protein MEX3A	BQ883752.1 93% Product:cDNA clone	**Pr**
ENSPTRP00000051696	9606 sc	one mutated splice site	P & UP: ENSG00000203758	P: ENSG00000203758	-	no similarity found	DB342969.1.1 (59%) Product:cDNA clone	**Tr**

ENSPTRP00000050833	9606 sc	no	P & UP: ENSG00000223816	UP: ENSG00000223816	-	no similarity found	DN995274 (92%) Product: TC121786 Human breast cancer cDNA clone	**Tr**
ENSPTRP00000049745	9606 sc	no	PT: ENSG00000243587	PT: ENSG00000243587	-	XM_001726130.2 (100%) Product: predicted hypothetical protein FLJ37396	BG572220.1 (100%) Product: cDNA clone IMAGE:4719772 5-	**Pr~**
ENSPTRP00000049566	9606 sc	no	PT: ENSG00000203733	PT: ENSG00000203733	-	no similarity found	no similarity found	-
ENSPTRP00000047979	9606 sc	no	PT: ENSG00000185290	PT: ENSG00000185290	-	NM_001145712.1 (100%) Product: hypothetical protein LOC389493	BI831486.1 et BI825227.1 (100%) Product: 603074595F1 cDNA clone et 603071914F1 cDNA clone	**Pr~**
ENSPTRP00000047306	9606 sc	no	PT: ENSG00000204025	PT: ENSG00000204025	-	NM_001195578.1 Product: cDNA clone IMAGE:4822062 et hypothetical LOC100329135	BF376155 (100%) Product: (cDNA clone)	**Pr~**
ENSPTRP00000041494	9606 sc	no	UP: ENSG00000196403	UP: ENSG00000196403	-	no similarity found	no similarity found	-
ENSPTRP00000037787	9606 sc	no	PT: ENSG00000147160	NS ET PC retained intron : ENSG00000147160	+	NM_001002254.1 (100%) Product: acyl-CoA wax alcohol acyltransferase 2	BG743236.1 (76%) Product: cDNA clone	**Pr**
ENSPTRP00000033937	9606 sc	no	P: ENSG00000176510	P: ENSG00000176510	-	NM_001005471.1 (7%*) Product: olfactory receptor 2T6	BX327855 (3%*) Product: cdna clone	-
ENSPTRP00000031144	9606 sc	no	PT: ENSG00000157606	PC: ENSG00000249481 PT: ENSG00000157606	+	NM_145026.3 (100%) Product: spermatogenesis-associated serine-rich protein	BX096175.1 (97%) Product: cDNA clone	**Pr**
ENSPTRP00000030458	9606 sc	no	P: ENSG00000168126	P: ENSG00000168126	-	NM_014053.3 (4%*) Product: feline leukemia virus subgroup C	CF597147.1 (5%*) Product: cDNA clone	-
ENSPTRP00000025208	9606 sc	no	PT: ENSG00000134075	PC: ENSG00000254999 PT: ENSG00000134075	+	NM_018462.4 (100%) Product: probable protein BRICK1	BJ996510.1 et DB024801.1 (100%) Product:cdna clone	**Pr**
ENSPTRP00000025206	9606 sc	no	PT: ENSG00000163705	PT: ENSG00000253453 PC: ENSG00000163705	+	NM_001164839.1 et NM_173472.1 (100%) Product: hypothetical protein LOC115795	BX360156.2 (100%) Product: cDNA clone	**Pr**
ENSPTRP00000020492	9606 sc	no	PT: ENSG00000162997	PT: ENSG00000162997	-	NR_027258.1 Product: prolyl-tRNA synthetase associated domain containing 1, pseudogene	AA987380 (100%) Product: cDNA clone	**Tr**
ENSPTRP00000011472	9606 sc	no	Nothing	PC: ENSG00000254656	+	NM_001134888.2 (100%) Product: retrotransposon-like protein 1	BE294363.1 (41%) Product: cDNA clone	**Pr**
ENSPTRP00000010010	9606 sc	no	PT: ENSG00000242674	PT: ENSG00000242674 PC: ENSG00000186047	+	NM_198989.2 (70%, manque 3'utr) Product: leukemia-associated protein	DN992279 94% Product: cDNA clone	**Pr**
ENSPTRP00000005373	9606 sc	One disappeared exon	PT: ENSG00000188916	PC: ENSG00000188916	+	NM_001039762.2 (100%) Product: "hypothetical protein LOC642938"	DW408011.1 35% (manque 5' et 3 utr) Product: cDNA	**Pr**
ENSPTRP00000004939	9606 sc	no	PT: ENSG00000155254	PT: ENSG00000155254	-	NM_031484.3 (100%) Product: putative MARVEL domain-containing protein 1	BG325189.1 (100%) Product: cDNA clone	**Pr~**
ENSPTRP00000002929	9606 sc	no	PT: ENSG00000121446	PT: ENSG00000121446	-	NM_001137669 (92%) Product: regulator of G-protein signaling protein-like	DB461953.2 (39%) Product: cDNA clone	**Pr~**

ENSPTRP00000002862	9606 sc	no	PT: ENSG00000188585	PT: ENSG00000188585	-	XR_114945.1 100% Product: PREDICTED: Homo sapiens non-protein coding RNA	DB340699.1 (11%*) Product: cDNA clone	-
ENSPTRP00000002547	9606 sc	no	P: ENSG00000180409	P: ENSG00000180409	-	NM_014818.1 (3%) + NR_027287.1 (3%*) Product: tripartite motif-containing protein 66 et CKL1 antisense RNA 1 (non-protein coding)	BY997135.1 (5*%) Product: cDNA clone	-
ENSPTRP00000001217	9606 sc	no	PT: ENSG00000186118	PT: ENSG00000186118	-	NM_001145474.1 (100%) Product: hypothetical protein LOC374973	BI462021.1 (100%) Product: cDNA clone	**Pr~**
ENSPTRP00000001083	9606 sc	no	PT: ENSG00000184157	PT: ENSG00000184157 PC:ENSG00000253313	+	NM_001164829.1 (100%) PRODUCT: putative uncharacterized protein C1orf210	CV575780 (100%) Product: cDNA clone	**Pr**
ENSPTRP00000001082	9606 sc	no	PT: ENSG00000179178	PC: ENSG00000179178	+	NM_144626.2 (100%) Product: transmembrane protein 125	BI600176.1 (100%) Product: cDNA clone	**Pr**
ENSPTRP00000000773	9606 sc	no	PT: ENSG00000187975	PT: ENSG00000187975 PC: ENSG00000253304	+	NM_001171868 (98%) Product: transmembrane protein 200B	BM546171 (65%) Product: cDNA clone	**Pr**
ENSMUSP00000109800	9606 un	/	PC: ENSG00000186470	PC: ENSG00000186470	+?	NM_197974.2 (100%) Product: butyrophilin subfamily 3 member A3 isoform b	BQ054495.1 (96%) Product: cDNA clone	**Pr**
ENSMUSP00000108580	9604 sc	no	Nothing	Nothing	+	NM_001195279.1 (100%) Product: hypothetical protein LOC100129480	DT932515.1 (72%) Product : cDNA clone	**Pr**
ENSMUSP00000095349	207598 sc	no	PT : ENSG00000205116	PT: ENSG00000205116	-	NM_001146685.1 (100 % ncbi) Product: transmembrane protein 88B	DN831326 (62%) Product: cDNA clone	**Pr~**
ENSMUSP00000094042	9526 sc	6 bp insertion	Nothing	PT: ENSG00000254470	-	NM_138368.3 (94%) Product: hypothetical protein LOC91056	BQ057983.1 62 % Product: cDNA clone	**Pr~**
ENSMUSP00000093970	9604 sc	no	PC: ENSG00000162194	PC: ENSG00000162194	+?	NM_024099.3 (100%) Product: hypothetical protein LOC79081	BX362851.2 (100%) Product: cDNA clone	**Pr**
ENSMUSP00000088808	9526 sc	no	Nothing	PC: ENSG00000253251	+	NR_003545.1 (100%) Product: -	BG537558.1 (99%) Product: cDNA clone	**Pr**
ENSMUSP00000087605	207598 un	/	UP: ENSG00000234757	UP: ENSG00000234757	-	XM_001719381 (51%) Product: carbonic anhydrase 15-like predicted	CB957335.1 (92%*) Product: cDNA clone	**Pr~**
ENSMUSP00000079081	9443 sc	no	PC: ENSG00000186577	PC: ENSG00000186577	+?	NM_001008703.1 (100%) Product: hypothetical protein LOC221491 precursor	DA400180.1 (100%) Product: cDNA clone	**Pr**
ENSMUSP00000061190	-3 sc	no	Nothing	PC: ENSG00000255181	+	NM_001162914.1 (100%) Product: coiled-coil domain containing 121-like	EG327530.1 (12%) Product: cDNA clone	**Pr**
ENSMUSP00000059069	-3 sc	no	PT:ENSG00000236980	PT: ENSG00000236980	-	NM_001080528.2 (100%) Product: hypothetical protein LOC646498"	AI190914.1 (100%) Product: cDNA clone	**Pr~**
ENSMUSP00000049725	9526 sc	no	P: ENSG00000176305	PC: ENSG00000253958 P: ENSG00000176305	+	NM_194284.2 (94%) Product: claudin-23	BG477807.1 (83%) Product: cDNA clone	**Pr**
ENSMUSP00000023728	9604 sc	no	Nothing	Nothing	+	NR_029449.1 (100%) Product: hypothetical LOC255411 non-coding RNA.	BI463658.1 (100%) Product: cDNA clone	**Pr**
ENSMODP00000000661	9526 sc	one mutated splice site	PC: ENSG00000156869	PC: ENSG00000156869	+?	NM_001013660.2 (100%) Product: ferric-chelate reductase 1	DY654935.1 (55%) Product: cDNA clone	**Pr**

Annexe 13 : Test d'un échantillon aléatoire de 51 gènes sauvés par rapport aux annotations disponibles dans les

bases de données

sc: séquence scannée au niveau nucléotidique; un: séquence non scannée au niveau nucléotidique; /: mutations non observées au niveau nucléotidique.

P: Pseudogene; PT: Processed Transcript; UP: Unprocessed pseudogene; PP: Processed pseudogene; PC: Protein coding; NS: Non sense mediated decay.

+: annotations de la BD Ensembl V61 en accord avec les résultats de GLADX; +?: Résultats de GLADX en accord avec ceux de Ensembl, le gène étant déjà annoté dans la BD Ensembl V57 utilisée par GLADX; -: Les informations des BD ne sont pas en accord avec les résultats de GLADX; Pr: La protéine existe; Pr~: La protéine existe potentiellement.

Annexe 14

Ensembl Protein as Ref	ratio dN/dS	Ensembl Protein as Ref	ratio dN/dS	Ensembl Protein as Ref	ratio dN/dS
ENSPTRP00000057518	0.64	ENSPTRP00000047979	0.25	ENSPTRP00000001217	1,61
ENSRNOP00000052333	0.36	ENSPTRP00000047306	0.0001	ENSPTRP00000001083	0
ENSRNOP00000047640	0.18	ENSPTRP00000041494	0.71	ENSPTRP00000001082	0
ENSPTRP00000059703	0.31	ENSPTRP00000037787	0	ENSPTRP00000000773	0.0001
ENSPTRP00000058392	0.42	ENSPTRP00000033937	17	ENSMUSP00000109800	0.30
ENSPTRP00000057227	0,19	ENSPTRP00000031144	0.47	ENSMUSP00000108580	0.28
ENSPTRP00000056568	0.0001	ENSPTRP00000030458	0.32	ENSMUSP00000095349	0.11
ENSPTRP00000056308	0.25	ENSPTRP00000025208	0	ENSMUSP00000094042	0.14
ENSPTRP00000056108	0	ENSPTRP00000025206	0	ENSMUSP00000093970	0.14
ENSPTRP00000055033	0	ENSPTRP00000020492	0	ENSMUSP00000088808	0.39
ENSPTRP00000054836	0.11	ENSPTRP00000013433	0	**ENSMUSP00000087605**	**0.85**
ENSPTRP00000052322	**0.97**	ENSPTRP00000010010	0	ENSMUSP00000079081	0,22
ENSPTRP00000051722	0.06	ENSPTRP00000005373	0.17	ENSMUSP00000061190	0.19
ENSPTRP00000051696	0.22	ENSPTRP00000004939	0,31	ENSMUSP00000059069	0.26
ENSPTRP00000050833	0.47	ENSPTRP00000002929	0.34	ENSMUSP00000049725	15
ENSPTRP00000049745	0	ENSPTRP00000002862	1,72	ENSMUSP00000023728	0.26
ENSPTRP00000049566	0.0001	ENSPTRP00000002547	1,59	ENSMODP00000000661	0.37

Annexe 14 : Ratio dN/dS d'un échantillon de gènes sauvés

Les calculs ont été effectués avec codeml du package PAML (Yang, 2007), plusieurs modèles ont été utilisés pour faire plus ou moins varier le ratio dans les différentes branches de la topologie utilisée. Un test de khi² a été effectué pour choisir le ratio du meilleur modèle. Un ratio de 0 définit une conservation très importante (une à deux mutations). dN/dS > 1 signifie une sélection positive, dN/dS < 1 signifie une sélection purificatrice, dN/dS ~ 1 équivaut à une sélection relâchée (évolution neutre).

Annexe 15

Gène de référence utilisé dans l'étude GLADX	Symbole / nom	Identification MGI du gène utilisé dans les études de Zhu et Zhang	Pertes	(Zhu *et al.*, 2007)	(Z. D. Zhang *et al.*, 2010)
ENSPTRP00000054241	Tas2r134	MGI:2681300 / UNIPROT:Q50KC1	P 9606		X
ENSMUSP00000052557	Nr1H5	MGI:3026618	P 9526		X
ENSMUSP00000071120	Tssk5	MGI:1920792	P 9526		X
ENSMUSP00000068253	Gucy2g	MGI:106025	P -3		X
ENSMUSP00000033166	4933427G17Rik	MGI:1921716	P 9526		X
ENSPTRP00000054852	Gpr33	MGI:1277106 / UNIPROT:Q49SQ3	P 9606	X	
ENSMUSP00000107398	Gm766	MGI:2685612	L 207598	X	X
ENSMUSP00000081903	Trpc2	MGI:109527	P 9526		X
ENSMUSP00000060912	Gulo	MGI:1353434	L 9526	X	X
ENSMUSP00000021406	2700097O09Rik	MGI:1919908	P 207598	X	X
ENSMUSP00000091442	4933422H20Rik	MGI:3588186	P 9526		X
ENSMUSP00000063340	0610012H03Rik	MGI:1921338	P 9526	X	X
ENSMUSP00000104022	Cyp2t4	MGI:2686296	P 9604		X
ENSMUSP00000025064	2610034M16Rik	MGI:1916489	P 9526		X
ENSMUSP00000070130	Nepn	MGI:1913900	L 9606	X	X
ENSMUSP00000023271	170016M24Rik	MGI:1916689	L 9606		X
ENSMUSP00000095648	Mup4	MGI:97236	P -3		X
ENSMUSP00000035069	Nradd	MGI:1914419	P -3	X	X
ENSPTRP00000054644	Ctf2	MGI:2684607 / UNIPROT:Q6R2R2	P 9606	X	X
ENSMUSP00000110749	Acyl 3	MGI:2442915	P -3	X	X
ENSMUSP00000033230	Tha1	MGI:1919026	P -3		X
ENSMUSP00000082150	4921517D21Rik	MGI:1914972	P -3	X	X
ENSMUSP00000093548	Slc7a15	MGI:3045351	P -3	X	X
ENSMUSP00000047621	AC102611.2	MGI:1916703	P 9526		X
ENSMUSP00000050451	1110012D08Rik	MGI:1921077	P -3		X
ENSMUSP00000029837	Uox	MGI:98907	P 9604	X	X
ENSMUSP00000059240	E230025N22Rik	MGI:3687212	P 9604		X
ENSMUSP00000075770	Sult1d1	MGI:1926341	P 9526	X	X
ENSMUSP00000022232	4933425L06Rik	MGI:1914013	P -2		X
ENSMUSP00000074381	Tlr12	MGI:3045221	P 9526		X
ENSMUSP00000036817	Taar3	MGI:3527427	P 9604		X
ENSMUSP00000028430	Cyct	MGI:88579	L -3		X
ENSMUSP00000034903	Gsta4	MGI:1309515	P 9526	X	
ENSMUSP00000090710	Slc42a2 / 4933429E10Rik	MGI:3588190	P 9526	X	
ENSMUSP00000065658	Art2b	MGI:107545	P 9526		X
ENSMUSP00000068158	BC048502	MGI:2652828	P 9526	X	
ENSMUSP00000028933	8030411F24Rik	MGI:1925859	P 9526		X

ENSMUSP00000021323	1700023F06Rik	MGI:1916691	P 9526		X
ENSMUSP00000043461	1700013G24Rik	MGI:1916630	P 9526	X	
ENSMUSP00000031690	Hyal6	MGI:1921659	P 207598		X
ENSPPYP00000012197	BC018465	MGI:2385160	L 207598	X	X
ENSMUSP00000068753	Prame	MGI:1923079	P -3		X
ENSMMUP00000011254	Sirpb3	MGI:3045317	L 207598		X
ENSMMUP00000012161	Tcam1	MGI:1923120	P 207598		X
ENSMMUP00000040945	Tex16	MGI:1890545	P 9604		X
ENSMUSP00000031935	1700074P13Rik	MGI:1920731	P 207598		X
ENSMMUP00000034281	4930511M11Rik	MGI:1922305	P 9604		X

Annexe 15 : Pertes détectées par GLADX communes aux deux principales études

L'identifiant MGI est la référence du gène utilisé pour les études de Zhu et Zhang. Dans la colonne pertes, P signifie que GLADX a trouvé un pseudogène, L une perte. Le nombre qui suit est l'identifiant du phylum dans lequel la perte est détectée. Les croix des dernières colonnes montrent les études communes avec GLADX.

Annexe 16

Reférence d'étude	Nom du gène	ESTs trouvés	Publications
ENSMODP00000009684	LOC100026279	/	
ENSGALP00000003217	LOC426514	/	
ENSCAFP00000012582	/	/	
ENSCAFP00000027551	LOC611724	/	
ENSMODP00000000863	LOC100030370	/	
ENSMODP00000005387	/	/	
ENSMODP00000022403	/	/	
ENSMUSP00000049931	Gm6377	/	
ENSMUSP00000052054	A530099J19Rik	/	
ENSOANP00000003470	/	/	
ENSOANP00000020384	/	/	
ENSRNOP00000046274	LOC688559	/	
ENSPTRP00000054644	Ctf2	/	(Derouet *et al.*, 2004 ; Z. D. Zhang *et al.*, 2010 ; Zhu *et al.*, 2007)
ENSMUSP00000070130	Nepn	/	(IHGSC, 2004 ; Mochida *et al.*, 2006 ; Z. D. Zhang *et al.*, 2010 ; Zhu *et al.*, 2007)
ENSMUSP00000093548	Slc7a15	/	(IHGSC, 2004 ; Z. D. Zhang *et al.*, 2010 ; Zhu *et al.*, 2007)
ENSMUSP00000095648	Mup4	/	(Chamero *et al.*, 2007 ; Z. D. Zhang *et al.*, 2010)
ENSMUSP00000033230	Tha1	/	(Z. D. Zhang *et al.*, 2010)
ENSMUSP00000099640	2810408A11Rik	/	(IHGSC, 2004)

Annexe 16 : Liste des 18 études qui n'ont pas d'ESTs dans les BDs parmi l'échantillon de 41 pseudogènes

Annexe 17

Reférence d'étude	Nom du gène	ESTs trouvés	Mutations détectées par GLADX	Mutations détectées dans EST	Résultat	Remarques / Publications
ENSMUSP00000082427	Adam4	complet	codon stop	indel	+	
ENSRNOP00000033535	LOC498236	partiel	Indel (codon stop), codon stop, site d'épissage	Indel (codon stop)	+	
ENSPTRP00000052598	LOC735480	complet	Indel (codon stop)	Indel (codon stop)	+	
ENSMUSP00000030011	1300002K09Rik	Partiel (la deuxiéme moitié)	non scanné	Indel (codon stop)	+	
ENSPTRP00000059300	env	complet	Indel (codon stop)	Indel (codon stop)	+	cas de rétrovirus
ENSMODP00000020269	MGC133804	partiel	Indel (codon stop),codon stop	Indel (codon stop),codon stop	+	
ENSMUSP00000073114	BC048679	complet	non scanné	Indel (codon stop)	+	
ENSPTRP00000059502	RGSL1	Partiel (la deuxiéme moitié)	Indel (codon stop)	Indel (codon stop)	+	
ENSMUSP00000074553	Dusp 13	complet	indel(codon stop)	pas de mutation	-	Mauvaise reconstruction des séquences ancestrales dans GLADX
ENSMUSP00000075923	Gm1574	complet	non scanné	pas de mutation	~+	Décrit sur NCBI et Ensembl comme ne produisant pas de protéines
ENSPTRP00000015214	Poldip 2	complet	indel(codon stop)	pas de mutation	-	/!\ erreur de séquençage. Délétion engendrant pseudogène uniquement sur Ensembl
ENSRNOP00000037327	Tril	complet	indel(codon stop)	pas de mutation	-	/!\ erreur de séquençage. Manque un nucléotide dans la séquence correspondante sur Ensembl
ENSPTRP00000047603	Moxd2	partiel	codon stop	codon stop	+	(Hahn, Jeong, & Lee, 2007)
ENSMUSP00000071120	Tssk5	partiel (premier exon)	non scanné	pas de mutation	~+	(Caenepeel, Charydczak, Sudarsanam, Hunter, & Manning, 2004 ; Z. D. Zhang *et al.*, 2010)
ENSPTRP00000054852	Gpr33	complet	codon stop	pas de mutation	polymorphisme	(IHGSC, 2004 ; Z. D. Zhang *et al.*, 2010 ; Zhu *et al.*, 2007)
ENSPTRP00000047678	Casp 12	complet	codon stop	pas de mutation	polymorphisme	(Fischer, Koenig, Eckhart, & Tschachler, 2002 ; IHGSC, 2004 ; Saleh *et al.*, 2004 ; X. Wang *et al.*, 2006 ; Xue *et al.*, 2006)
ENSMUSP00000059240	E230025N22Rik	partiel	non scanné	Indel (codon stop)	+	(Z. D. Zhang *et al.*, 2010)
ENSPTRP00000054241	Tas2r134	partiel (la fin)	codon stop	codon stop, indel	+	(Z. D. Zhang *et al.*, 2010)
ENSMUSP00000047621	AC102611.2	partiel (la fin)	indel(codon stop)	Pas de mutation	+	(Z. D. Zhang *et al.*, 2010)
ENSMUSP00000050451	1110012D08Rik	partiel (la deuxiéme moitié)	indel(codon stop)	indel(codon stop)	+	(Z. D. Zhang *et al.*, 2010)
ENSMUSP00000029837	Uox	partiel	codon stop	codon stop	+	(Oda, Satta, Takenaka, & Takahata, 2002 ; Wu *et al.*, 1989 ; Z. D. Zhang *et al.*, 2010 ; Zhu *et al.*, 2007)
ENSMUSP00000110749	Acyl 3	partiel	codon stop	indel(codon stop)	+	(Z. D. Zhang *et al.*, 2010 ; Zhu *et al.*, 2007)
ENSMUSP00000082150	4921517D21Rik	partiel (la fin)	codon stop	codon stop	+	(Z. D. Zhang *et al.*, 2010 ; Zhu *et al.*, 2007)

Annexe 17 : Liste des 23 études qui ont des ESTs dans les BDs, parmi l'échantillon de 41 pseudogènes

+ Le pseudogène est confirmé ; ~+ Le pseudogène ne peut pas être confirmé mais n'est pas infirmé ; - Le pseudogène est infirmé.

Annexe 18

Ensembl protein used as reference	gene appears (Taxid)	gene loss (Taxid)	ESTs trouvés	GLADX blast result exists	GLADX failed phylogenies exist	Result
ENSOANP00000001155	117571	32525	/	no	/	ok
ENSGALP00000014438	117571	9254	/	no	/	ok
ENSGALP00000006574	117571	9254	/	yes	no	ok
ENSMODP00000028478	117571	9526	/	yes	no	ok
ENSOANP00000019294	117571	9526	/	yes	no	ok
ENSCAFP00000000996	117571	9526	/	yes	no	ok
ENSMODP00000024801	117571	9347	/	yes	no	ok
ENSMODP00000037964	117571	9526	/	no	/	ok
ENSMODP00000035935	117571	9347	/	no	/	ok
ENSMODP00000019929	117571	9347	/	no	/	ok
ENSMODP00000027682	117571	9526	/	yes	yes	ok; the first hit is paralog, the second is bad maybe pseudogene
ENSMODP00000036751	117571	9526	/	yes	no	ok
ENSMODP00000005681	117571	9347	/	yes	yes	ok
ENSMUSP00000042575	117571	9526	/	yes	no	ok
ENSOANP00000011824	117571	32525	/	yes	no	ok
ENSMUSP00000068238	32525	9526	/	yes	yes	ok
ENSOANP00000012473	117571	32525	/	no	/	ok
ENSOANP00000013098	117571	32525	/	no	/	ok
ENSOANP00000022073	117571	32525	/	no	/	ok
ENSOANP00000017555	117571	32525	/	yes	no	ok
ENSOANP00000018156	117571	32525	/	no	/	ok
ENSOANP00000022832	117571	32525	/	yes	no	ok
ENSOANP00000014324	117571	9347	/	no	/	ok
ENSOANP00000015058	117571	32525	/	yes	no	ok
ENSMUSP00000041750	117571	9526	/	yes	no	ok
ENSOANP00000006898	117571	32525	/	yes	yes	ok
ENSOANP00000008940	117571	32525	/	yes	yes	ok
ENSMODP00000001284	117571	9347	DW43 0621	no	/	ok; the human putative orthologous EST found is very damaged; no hit found in human but a putative orthologous pseudogene exists in others Primates
ENSMUSP00000051917	9254	9526	/	yes	yes	ok ; one hit putative ortholog is pseudogene
ENSOANP00000013978	117571	32525	/	yes	no	ok
ENSMUSP00000055600	117571	9526	/	no	/	ok
ENSMODP00000011477	117571	9347	/	no	/	ok
ENSMODP00000007297	117571	9347	/	no	/	ok
ENSMODP00000000122	117571	9347	/	no	/	ok
ENSMODP00000022957	117571	9347	/	no	/	ok

ENSOANP00000013869	117571	32525	/	yes	yes	ok

Annexe 18 : Vérification d'un échantillon de pertes anciennes où une recherche de séquences orthologues a été faite par GLADX

La colonne « gene loss » indique le taxid de l'ancêtre où le gène semble être perdu. La colonne « GLADX blast result exists» contient yes lorsqu'au moins un hit de blast existe. La colonne « GLADX failed phylogenies exist » contient no lorsque aucun problème n'a été révélé, yes lorsque des phylogénies ont échoué. La colonne « Result » contient le résultat final après l'expertise manuelle.

La recherche d'ESTs a été faite auprès de la DB du NCBI.

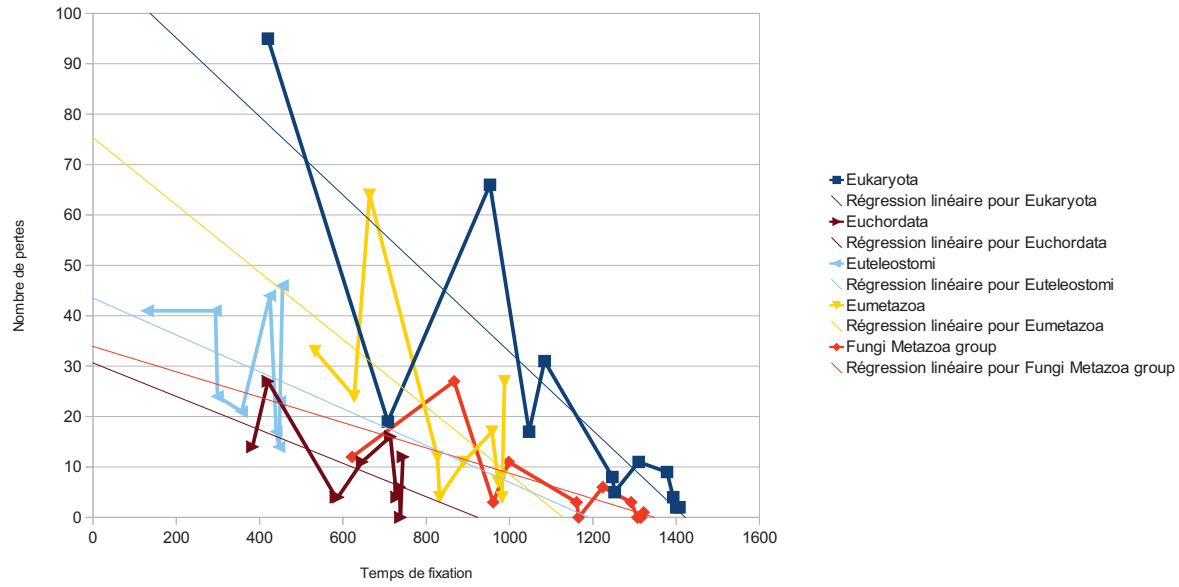
Annexe 19

Ensembl protein used as reference	Gene name / function	gene appears (Taxid)	gene loss (Taxid)	GLADX blast result exists	GLADX failed phylogenies exist	Result	Comment
ENSMUSP00000065456	RIKEN cDNA 5430435G22 gene	117571	9606	yes	no	ok	The first hit found ortholog in other Primates. This hit seems to not exists in Human
ENSGALP00000015482	hypothetical LOC420562	117571	207598	yes	no	ok	The loss event seems to be older
ENSMUSP00000050833	Olfr1018	117571	9606	yes	yes	ok	Missed pseudogene due to a failed phylogeny
ENSMUSP00000053925	Olfr1342	32525	207598	yes	no	ok	
ENSMUSP00000053887	Olfr600	32525	207598	yes	no	ok	
ENSGALP00000015482	LOC420562	117571	207598	yes	yes	ok	
ENSMUSP00000095803	Olfr599	32525	-3	yes	no	ok	
ENSPTRP00000017441	/	32523	9606	yes	yes	ok	Missed pseudogene due to a failed phylogeny
ENSCAFP00000007585	LOC485601 similar to Caspase-14 precursor	9254	-3	yes	no	ok	The loss event seems to be older
ENSPTRP00000050259	/	117571	9606	yes	yes	ok	Missed pseudogene due to a failed phylogeny
ENSMUSP00000062700	Olfr736	9254	207598	yes	no	ok	
ENSMUSP00000076633	Olfr644	32525	207598	yes	no	ok	
ENSMUSP00000079489	Olfr605	32524	207598	yes	no	ok	
ENSMUSP00000049569	Olfr630	32523	9604	yes	no	ok	The loss event seems to be older
ENSMUSP00000095791	Olfr632	32525	9604	yes	no	ok	
ENSMUSP00000049887	Olfr1026	32523	9606	yes	no	ok	
ENSPTRP00000055195	Olfr1231	9254	9606	yes	no	ok	
ENSMODP00000022265	/	117571	9606	yes	no	ok	The loss event seems to be older
ENSMUSP00000095816	Olfr572	32525	9604	yes	no	ok	
ENSPTRP00000035732	/	9254	9606	yes	yes	ok	Missed pseudogene due to a failed phylogeny

Annexe 19 : Vérification d'un échantillon de pertes récentes où une recherche de séquences orthologues a été faite par GLADX

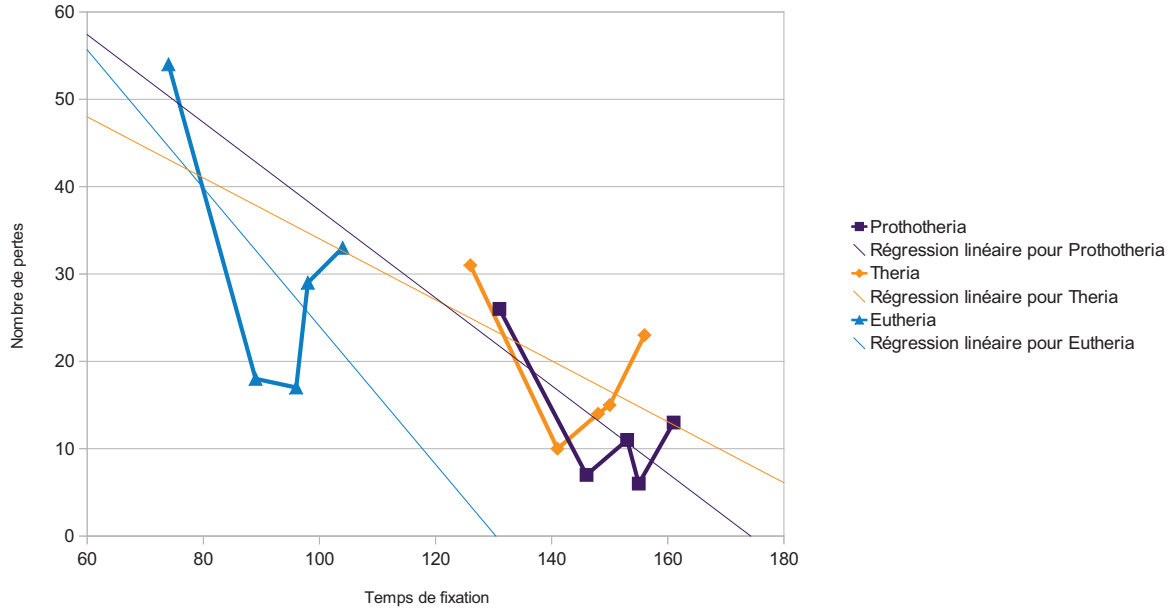
La colonne « gene loss » indique le taxid de l'ancêtre où le gène semble être perdu. La colonne « GLADX blast result exists » contient yes lorsqu'au moins un hit de blast existe. La colonne « GLADX failed phylogenies exist » contient no lorsque aucun problème n'a été révélé, yes lorsque des phylogénies ont échoué. La colonne « Result » contient le résultat final après l'expertise manuelle.

Annexe 20



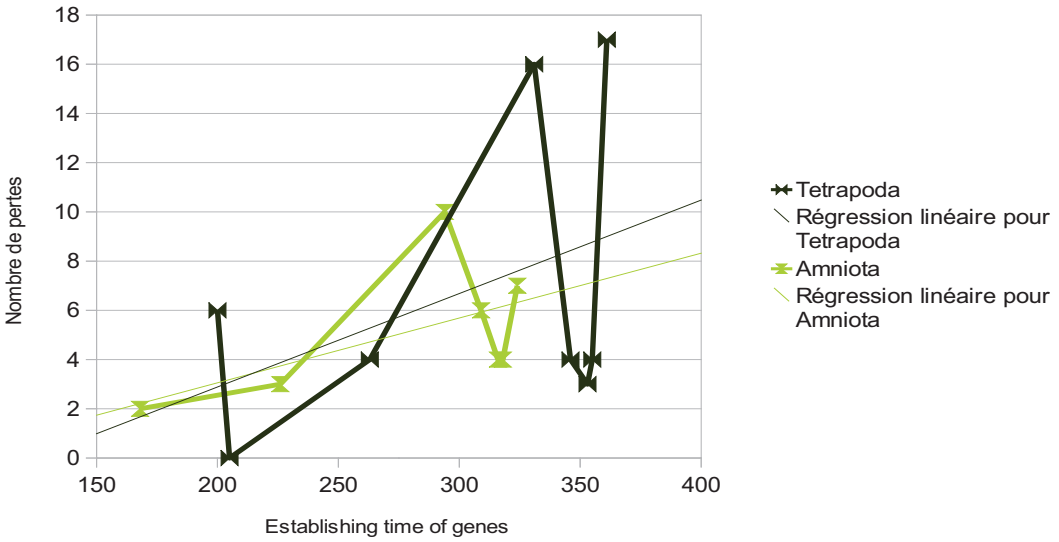
Annexe 20 : Nombre de gènes apparus chez un ancêtre spécifique, notés comme « perdus » en fonction de leur temps de fixation

Annexe 21



Annexe 21 : Nombre de gènes apparus chez un ancêtre spécifique, notés comme « perdus » en fonction de leur temps de fixation

Annexe 22



Annexe 22 : Nombre de gènes apparus chez un ancêtre spécifique, notés comme « perdus » en fonction de leur temps de fixation

Annexe 23

GO:0008150 : biological_process	GO:0005575 : cellular_component	GO:0003674 molecular_function
GO:0022610 : biological adhesion	GO:0005623 : cell	GO:0016209 : antioxidant activity
GO:0065007 : biological regulation	GO:0030054 : cell junction	GO:0005488 : binding
GO:0009758 : carbohydrate utilization	GO:0044464 : cell part	GO:0003824 : catalytic activity
GO:0015976 : carbon utilization	GO:0005576 : extracellular region	GO:0016247 : channel regulator activity
GO:0001906 : cell killing	GO:0044421 : extracellular region part	GO:0042056 : chemoattractant activity
GO:0008283 : cell proliferation	GO:0032991 : macromolecular complex	GO:0045499 : chemorepellent activity
GO:0071840 : cellular component organization or biogenesis	GO:0031974 : menbrane-enclosed lumen	GO:0009055 : electron carrier activity
GO:0009987 : cellular process	GO:0043226 : organelle	GO:0030234 : enzyme regulator activity
GO:0016265 : death	GO:0044422 : organelle part	GO:0016530 : metallochaperone activity
GO:0032502 : developmental process	GO:0055044 : symplast	GO:0060089 : molecular transducer activity
GO:0051234 : establishement of	GO:0045202 : synapse	GO:0016015 : morphogen

localization		activity
GO:0040007 : growth	GO:0044456 : synapse part	GO:0001071 : nucleic acid binding transcription factor activity
GO:0002376 : immune system process	GO:0019012 : virion	GO:0045735 : nutrient reservoir activity
GO:0051179 : localization	GO:0044423 : virion part	GO:0000988 : protein binding transcription factor activity
GO:0040011 : locomotion		GO:0031386 : protein tag
GO:0008152 : metabolic process		GO:0004872 : receptor activity
GO:0051704 : multi-organism process		GO:0030545 : receptor regulator activity
GO:0032501 : multicellular organismal process		GO:0005198 : structural molecule activity
GO:0048519 : negative regulation of biological process		GO:0045182 : translation regulator activity
GO:0019740 : nitrogen utilization		GO:0005215 : transporter activity
GO:0006794 : phosphorus utilization		
GO:0043473 : pigmentation		
GO:0048518 : positive regulation of biological process		
GO:0050789 : regulation of biological process		
GO:0000003 : reproduction		
GO:0022414 : reproductive process		
GO:0050896 : response to stimulus		
GO:0048511 : rhythmic process		
GO:0023052 : signaling		
GO:0006791 : sulfur utilization		

Annexe 23 : Les trois grands types d'annotations contenues dans la BD QuickGO et leur sous-catégorie respective de premier niveau

Tous les termes sont associés à un identifiant Gene Ontology (GO) unique.

Références bibliographiques

Albà, M. M., & Castresana, J. (2005). Inverse relationship between evolutionary rate and age of mammalian genes. *Molecular biology and evolution*, *22*(3), 598–606.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, *215*(3), 403–10.

Altschul, S. F., Madden, T. L., Schäffer, a a, Zhang, J., Zhang, Z., Miller, W., & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, *25*(17), 3389–402.

Andersson, S. G., & Kurland, C. G. (1998). Reductive evolution of resident genomes. *Trends in microbiology*, *6*(7), 263–8.

Aravind, L., Watanabe, H., Lipman, D. J., & Koonin, E. V. (2000). Lineage-specific loss and divergence of functionally linked genes in eukaryotes. *Proceedings of the National Academy of Sciences of the United States of America*, *97*(21), 11319–24.

Arber, W., & Linn, S. (1969). DNA modification and restriction. *Annual Review of Biochemistry*, *38*(1), 467–500. Annual Reviews 4139 El Camino Way, PO Box 10139, Palo Alto, CA 94303-0139, USA.

Aristote. (n.d.). *Métaphysique, Z, 7* (pp. 1032 a 24–25).

Avery, O. T., MacLeod, C. M., & McCarty, M. (1944). Studies on the chemical nature of the substance inducing transformation of pneumococcal types: Induction of transformation by a desoxyribonucleic pneumococcus type acid III. *Journal of Experimental Medicine*, *79*(2), 137–158.

Barker, D., Meade, A., & Pagel, M. (2007). Constrained models of evolution lead to improved prediction of functional linkage from correlated gain and loss of genes. *Bioinformatics*, *23*(1), 14–20. OXFORD UNIV PRESS.

Barker, D., & Pagel, M. (2005). Predicting functional gene links from phylogenetic-statistical analyses of whole genomes. (D. Murray, Ed.)*PLoS Computational Biology*, *1*(1), e3. Public Library of Science.

Baumann, P., Baumann, L., Lai, C. Y., Rouhbakhsh, D., Moran, N. A., & Clark, M. A. (1995). Genetics, physiology, and evolutionary relationships of the genus Buchnera: intracellular symbionts of aphids. *Annual review of microbiology*, *49*, 55–94.

Beneden, E. V. (1883). Recherches sur la maturation de I'oeuf et la fecondation. Ascaris megalocephala. *Archives Biologie*, *4*, 265–640.

Bischof, J. M., Chiang, A. P., Scheetz, T. E., Stone, E. M., Casavant, T. L., Sheffield, V. C., & Braun, T. A. (2006). Genome-wide identification of pseudogenes capable of disease-causing gene conversion. *Human mutation*, *27*(6), 545–52.

Blomme, T., Vandepoele, K., De Bodt, S., Simillion, C., Maere, S., & Van de Peer, Y. (2006). The gain and loss of genes during 600 million years of vertebrate evolution. *Genome Biology*, *7*(5), R43.

Boucher, Y., Douady, C. J., Papke, R. T., Walsh, D. a, Boudreau, M. E. R., Nesbø, C. L., Case, R. J., *et al.* (2003). Lateral gene transfer and the origins of prokaryotic groups. *Annual Review of Genetics*, *37*, 283–328.

Braun, E. L. (2000). Large-Scale Comparison of Fungal Sequence Information: Mechanisms of Innovation in Neurospora crassa and Gene Loss in Saccharomyces cerevisiae. *Genome Research*, *10*(4), 416–430.

Brawand, D., Wahli, W., & Kaessmann, H. (2008). Loss of egg yolk genes in mammals and the origin of lactation and placentation. *PLoS Biology*, *6*(3), 508–517.

Broman, K. W., Murray, J. C., Sheffield, V. C., White, R. L., & Weber, J. L. (1998). Comprehensive human genetic maps: individual and sex-specific variation in recombination. *American journal of human genetics*, *63*(3), 861–9.

Buffon, & Daubenton. (1766). *Histoire naturelle, générale et particulière: avec la description du Cabinet du Roi, vol 14*. Paris: Imprimerie Royale.

Bult, C. J., White, O., Olsen, G. J., Zhou, L., Fleischmann, R. D., Sutton, G. G., Blake, J. A., *et al.* (1996). Complete Genome Sequence of the Methanogenic Archaeon, Methanococcus jannaschii. *Science*, *273*(5278), 1058–1073.

Caenepeel, S., Charydczak, G., Sudarsanam, S., Hunter, T., & Manning, G. (2004). The mouse kinome: discovery and comparative genomics of all mouse protein kinases. *Proceedings of the National Academy of Sciences of the United States of America*, *101*(32), 11707–12.

Chamero, P., Marton, T. F., Logan, D. W., Flanagan, K., Cruz, J. R., Saghatelian, A., Cravatt, B. F., *et al.* (2007). Identification of protein pheromones that promote aggressive behaviour. *Nature*, *450*(7171), 899–902.

Chen, K., Durand, D., & Farach-Colton, M. (2000). NOTUNG: a program for dating gene duplications and optimizing gene family trees. *Journal of computational biology : a journal of computational molecular cell biology*, *7*(3-4), 429–47.

Chen, M. J., Shimada, T., Moulton, a D., Harrison, M., & Nienhuis, a W. (1982). Intronless human dihydrofolate reductase genes are derived from processed RNA molecules. *Proceedings of the National Academy of Sciences of the United States of America*, *79*(23), 7435–9.

Church, D. M., Goodstadt, L., Hillier, L. W., Zody, M. C., Goldstein, S., She, X., Bult, C. J., *et al.* (2009). Lineage-specific biology revealed by a finished genome assembly of the mouse. (R. J. Roberts, Ed.)*PLoS biology*, *7*(5), e1000112. Public Library of Science.

Cleary, M. L., Schon, E. A., & Lingrel, J. B. (1981). Two related pseudogenes are the result of a gene duplication in the goat β-globin locus. *Cell*, *26*(2), 181–190.

Conant, G. C., & Wolfe, K. H. (2008). Turning a hobby into a job: how duplicated genes find new functions. *Nature Reviews Genetics*, *9*(12), 938–50.

The 1000 Genomes Project Consortium. (2010). A map of human genome variation from population-scale sequencing. *Nature*, *467*(7319), 1061–73.

Costello, J. C., Han, M. V., & Hahn, M. W. (2008). Limitations of Pseudogenes in Identifying Gene Losses. *Gene*, 14–25.

Danchin, E. G. J., Gouret, P., & Pontarotti, P. (2006). Eleven ancestral gene families lost in mammals and vertebrates while otherwise universally conserved in animals. *BMC Evolutionary Biology*, *6*, 5.

Darby, A. C., Cho, N.-H., Fuxelius, H.-H., Westberg, J., & Andersson, S. G. E. (2007). Intracellular pathogens go extreme: genome evolution in the Rickettsiales. *Trends in genetics : TIG*, *23*(10), 511–20.

Darwin, C. (1871). *The descent of man, and selection in relation to sex*. London: John Murray.

Darwin, C. R. (1837). *Transmutation of species* (Notebook B.). CUL-DAR121.

Darwin, Charles. (1859). *On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life*. London: John Murray.

Darwin, Charles. (1868). *The variation of animals and plants under domestication, 2 vol.* London: John Murray.

Darwin, Charles, & Wallace, A. R. (1858). On the Tendency of Species to Form Varieties, and on the Perpetuation of Varieties by Natural Means of Selection. *Journal of Proceedings of the Linnean Society of London*, *III*(9), 1–62.

Dawkins, R. (2003). *Le gène égoïste* (O. Jacob.).

Dayhoff, Eck, Chang, & Sochard. (1965). *Atlas of Protein Sequence and Structure*.

Dehal, P., & Boore, J. L. (2005). Two rounds of whole genome duplication in the ancestral vertebrate. *PLoS biology*, *3*(10), e314.

Dehal, P., Satou, Y., Campbell, R. K. R. K., Chapman, J., Degnan, B., De Tomaso, A., Davidson, B., *et al.* (2002). The draft genome of Ciona intestinalis: insights into chordate and vertebrate origins. *Science*, *298*(5601), 2157–67. American Association for the Advancement of Science.

Derouet, D., Rousseau, F., Alfonsi, F., Froger, J., Hermann, J., Barbier, F., Perret, D., *et al.* (2004). Neuropoietin, a new IL-6-related cytokine signaling through the ciliary neurotrophic factor receptor. *Proceedings of the National Academy of Sciences of the United States of America*, *101*(14), 4827–32.

Dobzhansky, T. G. (1937). *Genetics and the Origin of Species*. Columbia Univ Pr.

Domazet-Lošo, T., Brajkovic, J., & Tautz, D. (2007). A phylostratigraphy approach to uncover the genomic history of major adaptations in metazoan lineages. *Trends in genetics : TIG*, *23*(11), 533–539.

Eisen, J., & Fraser, C. M. (2003). Phylogenomics: intersection of evolution and genomics. *Science*, *300*(5626), 1706–7.

F. H. Crick. (1958). On protein synthesis. *Symposia of the Society for Experimental Biology*, *12*, 138–163.

Farris, J. S. (1977). Phylogenetic Analysis Under Dollo's Law. *Syst Biol*, *26*(1), 77–88.

Feschotte, C., & Pritham, E. J. (2007). DNA transposons and the evolution of eukaryotic genomes. *Annual review of genetics*, *41*, 331–68.

Fiers, W., Contreras, R., Duerinck, F., Haegeman, G., Iserentant, D., Merregaert, J., Min Jou, W., *et al.* (1976). Complete nucleotide sequence of bacteriophage MS2 RNA: primary and secondary structure of the replicase gene. *Nature*, *260*(5551), 500–507.

Fischer, H., Koenig, U., Eckhart, L., & Tschachler, E. (2002). Human caspase 12 has acquired deleterious mutations. *Biochemical and biophysical research communications*, *293*(2), 722–6.

Fitch, W. (1970). Distinguishing homologous from analogous proteins. *Systematic Zoology*, *19*(2), 99–113.

Fitch, W. M., & Margoliash, E. (1967). Construction of Phylogenetic Trees. *Science*, *155*(3760), 279–284. American Association for the Advancement of Science.

Fleischmann, R. D., Adams, M. D., White, O., Clayton, R. A., Kirkness, E. F., Kerlavage, A. R., Bult, C. J., *et al.* (1995). Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. *Science*, *269*(5223), 496–512. American Association for the Advancement of Science.

Force, A., Lynch, M., Pickett, F. B., Amores, A., Yan, Y. L., & Postlethwait, J. (1999). Preservation of duplicate genes by complementary, degenerative mutations. *Genetics*, *151*(4), 1531–45.

Freimuth, R. R., Wiepert, M., Chute, C. G., Wieben, E. D., & Weinshilboum, R. M. (2004). Human cytosolic sulfotransferase database mining: identification of seven novel genes and pseudogenes. *The Pharmacogenomics Journal*, *4*(1), 54–65.

Frith, M. C., Wilming, L. G., Forrest, A., Kawaji, H., Tan, S. L., Wahlestedt, C., Bajic, V. B., *et al.* (2006). Pseudo-messenger RNA: phantoms of the transcriptome. (J. Blake, J. Hancock, B. Pavan, & L. Stubbs, Eds.)*PLoS genetics*, *2*(4), e23. Public Library of Science.

Fritz-Laylin, L. K., Prochnik, S. E., Ginger, M. L., Dacks, J. B., Carpenter, M. L., Field, M. C., Kuo, A., *et al.* (2010). The genome of Naegleria gruberi illuminates early eukaryotic versatility. *Cell*, *140*(5), 631–42. Elsevier Ltd.

Fsihi, H., De Rossi, E., Salazar, L., Cantoni, R., Labò, M., Riccardi, G., Takiff, H. E., *et al.* (1996). Gene arrangement and organization in a approximately 76 kb fragment encompassing the oriC region of the chromosome of Mycobacterium leprae. *Microbiology (Reading, England)*, *142 (Pt 1*, 3147–61.

Garcia-Vallvé, S., Romeu, A., & Palau, J. (2000). Horizontal gene transfer in bacterial and archaeal complete genomes. *Genome research*, *10*(11), 1719–25.

Germonpre, M., Sablin, M., Stevens, R., Hedges, R., Hofreiter, M., Stiller, M., & Despres, V. (2009). Fossil dogs and wolves from Palaeolithic sites in Belgium, the Ukraine and Russia: osteometry, ancient DNA and stable isotopes. *Journal of Archaeological Science*, *36*(2), 473–490.

Gladyshev, E. A., Meselson, M., & Arkhipova, I. R. (2008). Massive horizontal gene transfer in bdelloid rotifers. *Science (New York, N.Y.)*, *320*(5880), 1210–3.

Go, Y., Satta, Y., Takenaka, O., & Takahata, N. (2005). Lineage-specific loss of function of bitter taste receptor genes in humans and nonhuman primates. *Genetics*, *170*(1), 313–26.

Goffeau, A., Barrell, B. G., Bussey, H., Davis, R. W., Dujon, B., Feldmann, H., Galibert, F., *et al.* (1996). Life with 6000 genes. *Science (New York, N.Y.)*, *274*(5287), 546, 563–7.

Gould, S.J. (2002). *The structure of evolutionary theory*. Belknap Press.

Gould, Stephen Jay. (1993). *Le sourire du flamant rose - Réflexions sur l'histoire naturelle* (Seuil.).

Gould, Stephen Jay. (2012). *Les équilibres ponctués* (Gallimard.).

Gouret, P. (2009). Automatisation de processus d'annotation génomique contrôlée par système expert. Thèse de sciences, Université de Provence - centre St-Charles.

Gouret, P., Thompson, J. D., & Pontarotti, P. (2009). PhyloPattern: regular expressions to identify complex patterns in phylogenetic trees. *BMC Bioinformatics*, *10*, 298.

Gouret, P., Vitiello, V., Balandraud, N., Gilles, A., Pontarotti, P., & Danchin, E. G. J. (2005). FIGENIX: intelligent automation of genomic annotation: expertise integration in a new software platform. *BMC Bioinformatics*, *6*, 198.

Green, R. E., Krause, J., Briggs, A. W., Maricic, T., Stenzel, U., Kircher, M., Patterson, N., *et al.* (2010). A draft sequence of the Neandertal genome. *Science*, *328*(5979), 710–22.

Gu, X., Wang, Y., & Gu, J. (2002). Age distribution of human gene families shows significant roles of both large- and small-scale duplications in vertebrate evolution. *Nature genetics*, *31*(2), 205–9.

Haegeman, A., Jones, J. T., & Danchin, E. G. J. (2011). Horizontal Gene Transfer in Nematodes: A Catalyst for Plant Parasitism? *Molecular Pant-Microbe Interactions*, *24*(8), 879–887.

Hahn, Y., Jeong, S., & Lee, B. (2007). Inactivation of MOXD2 and S100A15A by exon deletion during human evolution. *Molecular Biology and Evolution*, *24*(10), 2203–12.

Hahn, Y., & Lee, B. (2005). Identification of nine human-specific frameshift mutations by comparative analysis of the human and the chimpanzee genome sequences. *Bioinformatics (Oxford, England)*, *21 Suppl 1*, i186–94.

Haldane, J. B. S. (1933). The part played by recurrent mutation in evolution. *Am Nat*, *67*, 5–9.

Harrison, P., Kumar, A., Lan, N., Echols, N., Snyder, M., & Gerstein, M. (2002). A small reservoir of disabled ORFs in the yeast genome and its implications for the dynamics of proteome evolution. *Journal of molecular biology*, *316*(3), 409–19.

Harrison, P. M., Zheng, D., Zhang, Z., Carriero, N., & Gerstein, M. (2005). Transcribed processed pseudogenes in the human genome: an intermediate form of expressed retrosequence lacking protein-coding ability. *Nucleic acids research*, *33*(8), 2374–83.

Harvey, P. H., & Pagel, M. D. (1991). The comparative method in evolutionary biology. (R. M. May & P. H. Harvey, Eds.)*Trends in Ecology & Evolution*. Oxford University Press.

Haussler, D., O'Brien, S. J., Ryder, O. A., Barker, F. K., Clamp, M., Crawford, A. J., Hanner, R., *et al.* (2009). Genome 10K: a proposal to obtain whole-genome sequence for 10 000 vertebrate species. *Journal of Heredity*, *100*(6), 659–674.

Heinen, T. J. A. J., Staubach, F., Häming, D., & Tautz, D. (2009). Emergence of a new gene from an intergenic region. *Current biology : CB*, *19*(18), 1527–31.

Hennig, W. (1966). *Phylogenetic Systematics*. (E. Sober, Ed.)*Univ Illinois Press Urbana IL Humphries C J* (Vol. 1, p. 263). University of Illinois Press.

Hollis, G. F., Hieter, P. A., McBride, O. W., Swan, D., & Leder, P. (1982). Processed genes: a dispersed human immunoglobulin gene bearing evidence of RNA-type processing. *Nature*, *296*(5855), 321–325.

Hong, X., Scofield, D. G., & Lynch, M. (2006). Intron size, abundance, and distribution within untranslated regions of genes. *Molecular Biology and Evolution*, *23*(12), 2392–404.

Hughes, A. L., & Friedman, R. (2004a). Differential loss of ancestral gene families as a source of genomic divergence in animals. *Proceedings of the Royal Society B: Biological Sciences*, *271 Suppl*, S107–9.

Hughes, A. L., & Friedman, R. (2004b). Shedding genomic ballast: extensive parallel loss of ancestral gene families in animals. *Journal of Molecular Evolution*, *59*(6), 827–33.

Hughes, A. L., & Friedman, R. (2005). Loss of ancestral genes in the genomic evolution of Ciona intestinalis. *Evolution & Development*, *7*(3), 196–200.

Huxley, T. H. (1863). *Evidence as to Man's Place in Nature*. Williams and Norgate.

IHGSC. (2004). Finishing the euchromatic sequence of the human genome. *Nature*, *431*, 931–945.

Innan, H., & Kondrashov, F. (2010). The evolution of gene duplications: classifying and distinguishing between models. *Nature Reviews Genetics*, *11*(2), 97–108. Nature Publishing Group.

Jackson, S., Macleod, M., & Krauss, R. (1959). Determination of type in capsulated transformants on pneumococcus by the genome of non-capsulated donor and recipient strains. *The Journal of experimental medicine*, *109*(5), 429–38.

Jacq, C., Miller, J. R., & Brownlee, G. G. (1977). A pseudogene structure in 5S DNA of Xenopus laevis. *Cell*, *12*(September), 109–120.

Jaillon, O., Aury, J.-M., Noel, B., Policriti, A., Clepet, C., Cassagrande, A., Choisne, N., *et al.* (2007). The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature*, *449*(7161), 463–467.

Johann Gregor Mendel. (1866). *Experiments on plant hybrids (English translation)*. *Verhandlungen des naturforschenden Vereines in Brünn* (Vol. IV).

Keebaugh, A. C., & Thomas, J. W. (2010). The evolutionary fate of the genes encoding the purine catabolic enzymes in hominoids, birds, and reptiles. *Molecular biology and evolution*, *27*(6), 1359–69.

Keeling, P. J., & Palmer, J. D. (2008). Horizontal gene transfer in eukaryotic evolution. *Nature reviews. Genetics*, *9*(8), 605–18. Nature Publishing Group.

Kellis, M., Birren, B. W., & Lander, E. S. (2004). Proof and evolutionary analysis of ancient genome duplication in the yeast Saccharomyces cerevisiae. *Nature*, *428*(6983), 617–24.

Keren, H., Lev-Maor, G., & Ast, G. (2010). Alternative splicing and evolution: diversification, exon definition and function. *Nature reviews. Genetics*, *11*(5), 345–55. Nature Publishing Group.

Khelifi, A., Adel, K., Duret, L., Laurent, D., Mouchiroud, D., & Dominique, M. (2005). HOPPSIGEN: a database of human and mouse processed pseudogenes. *Nucleic acids research*, *33*(Database issue), D59–66.

Kimura, M. (1968). Evolutionary rate at the molecular level. *Nature*.

Koonin, E. V. (2005). Orthologs, paralogs, and evolutionary genomics. *Annual review of genetics*, *39*, 309–38.

Koonin, E. V., Fedorova, N. D., Jackson, J. D., Jacobs, A. R., Krylov, D. M., Makarova, K. S., Mazumder, R., *et al.* (2004). A comprehensive evolutionary classification of proteins encoded in complete eukaryotic genomes. *Genome Biology*, *5*(2), R7.

Krylov, D. M., Wolf, Y. I., Rogozin, I. B., & Koonin, E. V. (2003). Gene loss, protein sequence divergence, gene dispensability, expression level, and interactivity are correlated in eukaryotic evolution. *Genome Research*, *13*(10), 2229–35.

Kuraku, S., & Kuratani, S. (2011). Genome-wide Detection of Gene Extinction in Early Mammalian Evolution. *Genome Biology and Evolution*, 1–45.

Lacy, E., & Maniatis, T. (1980). The nucleotide sequence of a rabbit beta-globin pseudogene. *Cell*, *21*(2), 545–53.

Lamarck, J. B. (1809). *Philosophie zoologique*. Paris: Dentu.

Lamarck, J. B. (1822). *Histoire naturelle des animaux sans vertèbres, 7 vol.* Paris: Déterville.

Larhammar, D., Lundin, L.-G., & Hallböök, F. (2002). The human Hox-bearing chromosome regions did arise by block or chromosome (or even genome) duplications. *Genome research*, *12*(12), 1910–20. Cold Spring Harbor Laboratory Press.

Lee, Y., Sultana, R., Pertea, G., Cho, J., Karamycheva, S., Tsai, J., Parvizi, B., *et al.* (2002). Cross-referencing eukaryotic genomes: TIGR Orthologous Gene Alignments (TOGA). *Genome research*, *12*(3), 493–502.

Lequesne, W. J. (1972). Further Studies Based on the Uniquely Derived Character Concept. *Syst Biol*, *21*(3), 281–288.

Levasseur, A., Pontarotti, P., Poch, O., & Thompson, J. D. (2008). Strategies for reliable exploitation of evolutionary concepts in high throughput biology. *Evolutionary bioinformatics online*, *4*(33), 121–37.

Li, L., Stoeckert, C. J., & Roos, D. S. (2003). OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Research*, *13*(9), 2178–89.

Li, W.-H., Gojobori, T., & Nei, M. (1981). Pseudogenes as a paradigm of neutral evolution. *Nature*, *292*(5820), 237–239.

Lindblad-Toh, K., Garber, M., Zuk, O., Lin, M. F., Parker, B. J., Washietl, S., Kheradpour, P., *et al.* (2011). A high-resolution map of human evolutionary constraint using 29 mammals. *Nature*, *478*(7370), 476–82. Nature Publishing Group.

Linnaeus, C. (1735). *Systema Naturae, sive Regna Tria Naturae systematice proposita per classes, ordines, genera, & species*. (Lugduni Batavorum & Theodorum Haak, Eds.). Leiden.

Little, P. F. (1982). Globin pseudogenes. *Cell*, *28*(4), 683–4.

Loguercio, L. L., & Wilkins, T. a. (1998). Structural analysis of a hmg-coA-reductase pseudogene: insights into evolutionary processes affecting the hmgr gene family in allotetraploid cotton (Gossypium hirsutum L.). *Current Genetics*, *34*(4), 241–9.

Lucas-Lledó, J. I., & Lynch, M. (2009). Evolution of mutation rates: phylogenomic analysis of the photolyase/cryptochrome family. *Molecular biology and evolution*, *26*(5), 1143–53.

Lynn, A., Ashley, T., & Hassold, T. (2004). Variation in human meiotic recombination. *Annual review of genomics and human genetics*, *5*, 317–49.

M. O. Dayhoff, & Ledley, R. S. (1962). Comprotein: A Computer Program to Aid Primary Protein Structure Determination. *In Proceedings of the Fall Joint Computer Conference*, 262–274.

Makalowski, W. (2001). Are we polyploids? A brief history of one hypothesis. *Genome research*, *11*(5), 667–70.

Matthaei, J. H., Jones, O. W., Martin, R. G., & Nirenberg, M. W. (1962). Characteristics and composition of RNA coding units. *Proceedings of the National Academy of Sciences*, *48*(4), 666–677.

Mayr, E. (1942). *Systematics and the origin of species, from the viewpoint of a zoologist*. Harvard Univ Pr.

Mayr, E. (1963). *Animal species and evolution. Animal species and their evolution.* Harvard University Press; London: Oxford University Press.

McCarrey, J. R., & Thomas, K. (1987). Human testis-specific PGK gene lacks introns and possesses characteristics of a processed gene. *Nature*, *326*(6112), 501–5.

McLysaght, A., Hokamp, K., & Wolfe, K. H. (2002). Extensive genomic duplication during early chordate evolution. *Nature genetics*, *31*(2), 200–4.

Meselson, M., & Yuan, R. (1968). DNA restriction enzyme from E. coli. *Nature*, *217*(5134), 1110–1114.

Miescher, F. (1871). Ueber die chemische Zusammensetzung der Eiterzellen. *Medizinisch-chemische Untersuchungen*, *4*, 441–460.

Minocherhomji, S., Seemann, S., Mang, Y., El-schich, Z., Bak, M., Hansen, C., Papadopoulos, N., *et al.* (2012). Sequence and expression analysis of gaps in human chromosome 20. *Nucleic Acids Research*, 1–13.

Mirkin, B. G., Fenner, T. I., Galperin, M. Y., & Koonin, E. V. (2003). Algorithms for computing parsimonious evolutionary scenarios for genome evolution , the last universal common ancestor and dominance of horizontal gene transfer in the evolution of prokaryotes. *BMC Evolutionary Biology*, *1*, 1–34.

Mitchell, A., & Graur, D. (2005). Inferring the pattern of spontaneous mutation from the pattern of substitution in unitary pseudogenes of Mycobacterium leprae and a comparison of mutation patterns among distantly related organisms. *Journal of Molecular Evolution*, *61*(6), 795–803.

Mochida, Y., Parisuthiman, D., Kaku, M., Hanai, J., Sukhatme, V. P., & Yamauchi, M. (2006). Nephrocan, a novel member of the small leucine-rich repeat protein family, is an inhibitor of transforming growth factor-beta signaling. *The Journal of biological chemistry*, *281*(47), 36044–51.

Molinier, J., Ries, G., Zipfel, C., & Hohn, B. (2006). Transgeneration memory of stress in plants. *Nature*, *442*(7106), 1046–9.

Moran, N. A., & Mira, A. (2001). The process of genome shrinkage in the obligate symbiont Buchnera aphidicola. *Genome biology*, *2*(12), RESEARCH0054.

Moran, N., & Wernegreen, J. (2000). Lifestyle evolution in symbiotic bacteria: insights from genomics. *Trends in ecology & evolution*, *15*(8), 321–326.

Morgan, T. H., Sturtevant, A. H., Muller, H. J., & Bridges, C. B. (1915). *The Mechanism of Mendelian Heredity*.

Muller, H. J. (1935). The origination of chromatin deficiencis as minute deletions subject to insertion elsewhere. *Genetics*, *17*, 237–252.

Needleman, S. B., & Wunsch, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, *48*(3), 443–453.

Nishikimi, M., Fukuyama, R., Minoshima, S., Shimizu, N., & Yagi, K. (1994). Cloning and chromosomal mapping of the human nonfunctional gene for L-gulono-gamma-lactone oxidase, the enzyme for L-ascorbic acid biosynthesis missing in man. *The Journal of Biological Chemistry*, *269*(18), 13685–8.

Ochiai, K., Yamanaka, T., Kimura, K., & Sawada, O. (1959). Inheritance of drug resistance (and its transfer) between Shigella strains and between Shigella and E. coli strains (in Japanese). *Hihon Iji Shimpor*, *1861*.

Oda, M., Satta, Y., Takenaka, O., & Takahata, N. (2002). Loss of urate oxidase activity in hominoids and its evolutionary implications. *Molecular Biology and Evolution*, *19*(5), 640–53.

Ohno, S. (1970). Evolution by gene duplication. *Springer*.

Ohno, S. (1972). So much "junk" DNA in our genome. *Brookhaven symposia in biology*, *23*, 366–370.

Olson, M. (1999). When less is more: gene loss as an engine of evolutionary change. *American Journal of Human Genetics*, 18–23.

Ose, T., & Bush, O. J. (1962). Erythroblastosis fetalis. Report of cases, with one caused by gene deletion. *Bibliotheca haematologica*, *13*, 290–1.

O'Brien, K. P., Remm, M., & Sonnhammer, E. L. L. (2005). Inparanoid: a comprehensive database of eukaryotic orthologs. *Nucleic acids research*, *33*(Database issue), D476–80.

Pace, J. K., Gilbert, C., Clark, M. S., & Feschotte, C. (2008). Repeated horizontal transfer of a DNA transposon in mammals and other tetrapods. *Proceedings of the National Academy of Sciences of the United States of America*, *105*(44), 17023–8.

Pagani, I., Liolios, K., Jansson, J., Chen, I.-M. a, Smirnova, T., Nosrat, B., Markowitz, V. M., *et al.* (2011). The Genomes OnLine Database (GOLD) v.4: status of genomic and metagenomic projects and their associated metadata. *Nucleic Acids Research*, *40*(December 2011), 571–579.

Paganini J., G. P. (2012). Reliable phylogenetic trees building: a new web interface for FIGENIX. *Evolutionary Bioinformatics*, *in press*.

Pagel, M. (1994). Detecting Correlated Evolution on Phylogenies: A General Method for the Comparative Analysis of Discrete Characters. *Proceedings of the Royal Society B: Biological Sciences*, *255*(1342), 37–45.

Paten, B., Herrero, J., Fitzgerald, S., Beal, K., Flicek, P., Holmes, I., & Birney, E. (2008). Genome-wide nucleotide-level mammalian ancestor reconstruction. *Genome Research*, *18*(11), 1829–43.

Pionnier-Capitan, M., Bemilli, C., Bodu, P., Célérier, G., Ferrié, J.-G., Fosse, P., Garcià, M., *et al.* (2011). New evidence for Upper Palaeolithic small domestic dogs in South-Western Europe. *Journal of Archaeological Science*, *38*(9), 2123–2140. Elsevier Ltd.

Proudfoot, N. J., & Maniatis, T. (1980). The structure of a human alpha-globin pseudogene and its relationship to alpha-globin gene duplication. *Cell*, *21*(2), 537–44.

Putnam, N. H., Butts, T., Ferrier, D. E. K., Furlong, R. F., Hellsten, U., Kawashima, T., Robinson-Rechavi, M., *et al.* (2008). The amphioxus genome and the evolution of the chordate karyotype. *Nature*, *453*(7198), 1064–71.

Putnam, N. H., Srivastava, M., Hellsten, U., Dirks, B., Chapman, J., Salamov, A., Terry, A., *et al.* (2007). Sea anemone genome reveals ancestral eumetazoan gene repertoire and genomic organization. *Science*, *317*(5834), 86–94.

Race, R. R., Sanger, R., & Selwyn, J. G. (1951). A possible deletion in a human Rh chromosome; a serological and genetical study. *British journal of experimental pathology*, *32*(2), 124–35.

Ramos-Onsins, S., & Aguadé, M. (1998). Molecular evolution of the cecropin multigene family in Drosophila: functional genes vs. pseudogenes. *Brain, Behavior and Evolution*, *52*(4-5), 177–185.

Remm, M., Storm, C. E. V., & Sonnhammer, E. L. L. (2001). Automatic Clustering of Orthologs and In-paralogs from Pairwise Species Comparisons. *Online*, 1041–1052.

Roelofs, J., & Haastert, P. J. M. V. (2001). Genes lost during evolution. *Nature*, *411*(6841), 1013–1014.

Rumpho, M. E., Worful, J. M., Lee, J., Kannan, K., Tyler, M. S., Bhattacharya, D., Moustafa, A., *et al.* (2008). Horizontal gene transfer of the algal nuclear gene psbO to the photosynthetic sea slug Elysia chlorotica. *Proceedings of the National Academy of Sciences of the United States of America*, *105*(46), 17867–71.

Saleh, M., Vaillancourt, J. P., Graham, R. K., Huyck, M., Srinivasula, S. M., Alnemri, E. S., Steinberg, M. H., *et al.* (2004). Differential modulation of endotoxin responsiveness by human caspase-12 polymorphisms. *Nature*, *429*(6987), 75–9.

Sankoff, D. (1975). Minimal mutation trees in sequences. *Society for Industrial and Applied Mathematics*, *28*, 35–42.

Sankoff, D., & Rousseau, P. (1975). Locating the vertices of a Steiner tree in an arbitrary metric space. *Mathematical Programming*, *9*, 240–246.

Scherer, S. (2008). *A Short Guide to the Human Genome* (p. 173). CSHL Press.

Schrider, D. R., Costello, J. C., & Hahn, M. W. (2009). All human-specific gene losses are present in the genome as pseudogenes. *Journal of computational biology : a journal of computational molecular cell biology*, *16*(10), 1419–27.

Sherman, F., & Slonimski, P. P. (1964). Respiration-deficient mutants of yeast. *Biochimica et Biophysica Acta (BBA) - General Subjects*, *90*(1), 1–15.

Simpson, G. G. (1944). *Tempo and mode in evolution*. Columbia Univ Pr.

Smith, H. O., & Kelly, T. J. (1970). A restriction enzyme from Hemophilus influenzae II. Base sequence of the recognition site. *Journal of molecular biology*, *51*, 393–409.

Smith, H. O., & Wilcox, K. W. (1970). A Restriction enzyme from Hemophilus influenzae I. Purification and general properties. *Journal of Molecular Biology*, *51*, 379–391.

Smith, T. F., & Waterman, M. S. (1981). Identification of common molecular subsequences. *Journal of Molecular Biology*, *147*(1), 195–197.

Snel, B., Bork, P., & Huynen, M. A. (2002). Genomes in flux: the evolution of archaeal and proteobacterial gene content. *Genome Research*, *12*(1), 17–25.

Spencer, M., Susko, E., & Roger, A. J. (2006). Modelling prokaryote gene content. *Evolutionary bioinformatics online*, *2*(2003), 157–78.

Srivastava, M., Simakov, O., Chapman, J., Fahey, B., Gauthier, M. E. a., Mitros, T., Richards, G. S., *et al.* (2010). The Amphimedon queenslandica genome and the evolution of animal complexity. *Nature*, *466*(7307), 720–726. Nature Publishing Group.

Stedman, H. H., Kozyak, B. W., Nelson, A., Thesier, D. M., Su, L. T., Low, D. W., Bridges, C. R., *et al.* (2004). Myosin gene mutation correlates with anatomical changes in the human lineage. *Nature*, *428*(6981), 415–8.

Stegemann, S., & Bock, R. (2009). Exchange of genetic material between cells in plant tissue grafts. *Science (New York, N.Y.)*, *324*(5927), 649–51.

Sällström, B., & Andersson, S. G. E. (2005). Genome reduction in the alpha-Proteobacteria. *Current opinion in microbiology*, *8*(5), 579–85.

Tatusov, R L, Natale, D. a, Garkavtsev, I. V., Tatusova, T. a, Shankavaram, U. T., Rao, B. S., Kiryutin, B., *et al.* (2001). The COG database: new developments in phylogenetic classification of proteins from complete genomes. *Nucleic Acids Research*, *29*(1), 22–8.

Tatusov, R. L. (1997). A Genomic Perspective on Protein Families. *Science*, *278*(5338), 631–637.

Tatusov, Roman L, Fedorova, N. D., Jackson, J. D., Jacobs, A. R., Kiryutin, B., Koonin, E. V., Krylov, D. M., *et al.* (2003). The COG database: an updated version inicludes eukaryotes. *BMC Bioinformatics*, *4*(41).

Tiselius, A. W. K. (1930). The moving-boundary method of studying the electrophoresis of proteins. *Nova Acta Regiae Societatis Scientiarum Upsaliensis*, *7*(4).

Torrents, D., Suyama, M., Zdobnov, E., & Bork, P. (2003). A genome-wide survey of human pseudogenes. *Genome Research*, *13*(12), 2559–67.

Turing, A. M. (1937). On Computable Numbers, with an Application to the Entscheidungsproblem. *Proceedings of the London Mathematical Society*, *42*, 230–265.

Van Dongen, S. (2000). *Graph Clustering by Flow Simulation*. *University of Utrecht*. University of Utrecht.

Van de Peer, Y., Maere, S., & Meyer, A. (2009). The evolutionary significance of ancient genome duplications. *Nature reviews Genetics*, *10*, 725–732.

Varki, A. (2001). Loss of N-Glycolylneuraminic Acid in Humans: Mechanisms, Consequences, and Implications for Hominid Evolution. *Yearbook of Physical Anthropology*, *69*, 54–69.

Vignoles, A. (1738). *Chronologie de l'histoire sainte*. *Berlin* (Ambroise H.). Berlin.

Volff, J.-N. (2006). Turning junk into gold: domestication of transposable elements and the creation of new genes in eukaryotes. *BioEssays : news and reviews in molecular, cellular and developmental biology*, *28*(9), 913–22.

Wakasugi, T., Tsudzuki, J., Ito, S., Nakashima, K., Tsudzuki, T., & Sugiura, M. (1994). Loss of all ndh genes as determined by sequencing the entire chloroplast genome of the black pine Pinus thunbergii. *Proceedings of the National Academy of Sciences of the United States of America*, *91*(21), 9794–8.

Wang, X., Grus, W. E., & Zhang, J. (2006). Gene losses during human origins. *PLoS biology*, *4*(3), e52.

Wang, X., Thomas, S. D., & Zhang, J. (2004). Relaxation of selective constraint and loss of function in the evolution of human bitter taste receptor genes. *Human molecular genetics*, *13*(21), 2671–8.

Wang, Y., & Gu, X. (2000). Evolutionary patterns of gene families generated in the early stage of vertebrates. *Journal of molecular evolution*, *51*(1), 88–96.

Wapinski, I., Pfeffer, A., Friedman, N., & Regev, A. (2007a). Automatic genome-wide reconstruction of phylogenetic gene trees. *Bioinformatics (Oxford, England)*, *23*(13), i549–58.

Wapinski, I., Pfeffer, A., Friedman, N., & Regev, A. (2007b). Natural history and evolutionary principles of gene duplication in fungi. *Nature*, *449*(7158), 54–61.

Watson, J. D., & Crick, F. H. C. (1953). Molecular structure of nucleic acids. *Nature*, *171*(4356), 737–8.

Weaver, W. (1970). Molecular Biology: Origin of the Term. *Science*, *170*, 581–582.

Weigel, D., & Mott, R. (2009). The 1001 genomes project for Arabidopsis thaliana. *Genome biology*, *10*(5), 107.

Weismann, A. F. L. (1892a). La prétendue transmission héréditaire des mutilations. *Essais sur l'hérédité et la sélection naturelle* (pp. 424–426). Paris: C. Reinwald.

Weismann, A. F. L. (1892b). *Das Keimplasma: Eine Theorie der Vererbung*.

Weismann, A. F. L. (1892c). La Vie et la Mort. *Essais sur l'hérédité et la sélection naturelle* (p. 97). Paris: C. Reinwald.

Wirth, B., Leh-Louis, V., Potier, S., Souciet, J.-L., & Despons, L. (2005). Paleogenomics or the search for remnant duplicated copies of the yeast DUP240 gene family in intergenic areas. *Molecular Biology and Evolution*, *22*(9), 1764–71.

Wolf, Y. I., Novichkov, P. S., Karev, G. P., Koonin, E. V., & Lipman, D. J. (2009). The universal distribution of evolutionary rates of genes and distinct characteristics of eukaryotic genes of different apparent ages. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(18), 7273–80.

Wu, X. W., Lee, C. C., Muzny, D. M., & Caskey, C. T. (1989). Urate oxidase: primary structure and evolutionary implications. *Proceedings of the National Academy of Sciences of the United States of America*, *86*(23), 9412–6.

Wyder, S., Kriventseva, E. V., Schröder, R., Kadowaki, T., & Zdobnov, E. M. (2007). Quantification of ortholog losses in insects and vertebrates. *Genome Biology*, *8*(11), R242.

Xue, Y., Daly, A., Yngvadottir, B., Liu, M., Coop, G., Kim, Y., Sabeti, P., *et al.* (2006). Spread of an inactive form of caspase-12 in humans is due to recent positive selection. *American journal of human genetics*, *78*(4), 659–70. Elsevier.

Yamada, K., Lim, J., Dale, J. M., Chen, H., Shinn, P., Palm, C. J., Southwick, A. M., *et al.* (2003). Empirical analysis of transcriptional activity in the Arabidopsis genome. *Science (New York, N.Y.)*, *302*(5646), 842–6.

Yang, Z. (2007). PAML 4: phylogenetic analysis by maximum likelihood. *Molecular Biology and Evolution*, *24*(8), 1586–91.

Yasui, A., Eker, A. P., Yasuhira, S., Yajima, H., Kobayashi, T., Takao, M., & Oikawa, A. (1994). A new class of DNA photolyases present in various organisms including aplacental mammals. *The EMBO journal*, *13*(24), 6143–51.

Zdobnov, E. M., von Mering, C., Letunic, I., Torrents, D., Suyama, M., Copley, R. R., Christophides, G. K., *et al.* (2002). Comparative genome and proteome analysis of Anopheles gambiae and Drosophila melanogaster. *Science*, *298*(5591), 149–59.

Zhang, Z., Carriero, N., & Gerstein, M. (2004). Comparative analysis of processed pseudogenes in the mouse and human genomes. *Trends in Genetics*, *20*(2), 62–67.

Zhang, Z. D., Frankish, A., Hunt, T., Harrow, J., & Gerstein, M. (2010). Identification and analysis of unitary pseudogenes: historic and contemporary gene losses in humans and other primates. *Genome Biology*, *11*(3), R26.

Zhang, Z., Harrison, P. M., Liu, Y., & Gerstein, M. (2003). Millions of years of evolution preserved: a comprehensive catalog of the processed pseudogenes in the human genome. *Genome research*, *13*(12), 2541–58.

Zheng, D., & Gerstein, M. B. (2007). The ambiguous boundary between genes and pseudogenes: the dead rise up, or do they? *Trends in genetics : TIG*, *23*(5), 219–24.

Zheng, D., Zhang, Z., Harrison, P. M., Karro, J., Carriero, N., & Gerstein, M. (2005). Integrated pseudogene annotation for human chromosome 22: evidence for transcription. *Journal of molecular biology*, *349*(1), 27–45.

Zhu, J., Sanborn, J. Z., Diekhans, M., Lowe, C. B., Pringle, T. H., & Haussler, D. (2007). Comparative genomics search for losses of long-established genes on the human lineage. *PLoS Computational Biology*, *3*(12), e247.

Zomorodipour, A., & Andersson, S. G. (1999). Obligate intracellular parasites: Rickettsia prowazekii and Chlamydia trachomatis. *FEBS letters*, *452*(1-2), 11–5.

de Hoon, M. J. L., Imoto, S., Nolan, J., & Miyano, S. (2004). Open source clustering software. *Bioinformatics (Oxford, England)*, *20*(9), 1453–4.

Résumé

La biologie a connu une extraordinaire révolution avec l'arrivée de nombreux génomes entièrement séquencés. L'analyse de la quantité d'informations disponibles nécessite la création et l'utilisation d'outils informatiques automatisés. L'interprétation des données biologiques prend tout son sens à la lumière de l'évolution. En ce sens, les études évolutives sont incontestablement nécessaires pour donner un sens aux données biologiques. Dans ce contexte, le laboratoire développe des outils pour étudier l'évolution des génomes (et protéomes) à travers les mutations subies. Cette thèse porte sur l'étude spécifique des événements de pertes de gènes unitaires. Ces événements peuvent révéler des pertes de fonctions très instructives pour comprendre l'évolution des espèces. En premier lieu, j'ai développé l'outil GLADX qui mime l'expertise humaine afin d'étudier automatiquement et avec précision les événements de pertes de gènes unitaires. Ces études se basent sur la création et l'interprétation de données phylogénétiques, de BLAST, de prédictions protéiques, etc., dans un contexte automatisé. Ensuite, j'ai développé une stratégie utilisant l'outil GLADX pour étudier à grande échelle les pertes de gènes unitaires au cours de l'évolution du protéome humain. La stratégie utilise d'abord comme filtre l'analyse de groupes d'orthologues fabriqués par un outil de clustérisation à partir du protéome complet de nombreuses espèces. Cette analyse a permis de détecter 6237 pertes de gènes unitaires putatives dans la lignée humaine. L'étude approfondie de ces pertes avec GLADX a mis en évidence de nombreux problèmes liés à la qualité des données disponibles dans les bases de données. Elle a essentiellement permis de détecter 1318 pertes de gènes unitaires depuis l'ancêtre des Eucaryotes correspondants à près de 5% du protéome humain. Cette étude montre l'importance du phénomène de pertes de gènes unitaires dans l'évolution des génomes. La majorité des pertes identifiées sont décrites pour la première fois. Une des particularités de cette thèse est d'aborder et analyser aussi bien les pertes de gènes unitaires sans signaux des séquences d'origine, que celles liées à des pseudogènes. De plus, les événements de pertes et de pseudogénisations sont mis en évidence dans un contexte évolutif. L'apport d'informations fonctionnelles et des études comparatives permettront d'appréhender les phénomènes sous-jacents ayant induit ces pertes.

Mots clés : Perte de gènes unitaires, pseudogène unitaire, pseudogénisation, automatisation, phylogénie, homme

Abstract

Biology has undergone an extraordinary revolution with the appearance of numerous whole genomes sequenced. Analysis of the amount of information available requires creation and use of automated tools. The interpretation of biological data becomes meaningful in light of evolution. In view of all this, evolutionary studies are undoubtedly necessary to highlight the biological data. In this context, the laboratory develops tools to study the genomes (and proteomes) evolution through all the undergone mutations. The project of this thesis focuses specifically on the events of unitary gene losses. These events may reveal loss of functions very instructive for understanding the evolution of species. First, I developed the GLADX tool that mimics human expertise to automatically and accurately investigate the events of unitary gene losses. These studies are based on the creation and interpretation of phylogenetic data, BLAST, predictions of protein, etc., in an automated environment. Secondly, I developed a strategy using GLADX tool to study, at large-scale, the loss of unitary genes during the evolution of the human proteome. The strategy uses, in the first step, the analysis of orthologous groups produced by a clustering tool from complete proteomes of numerous species. This analysis used as a filter, allowed detecting 6237 putative losses in the human lineage. The study of these unitary gene loss cases has been deepened with GLADX and allowed to highlight many problems with the quality of available data in databases. It enabled to mainly identify 1318 unitary gene losses from the ancestor of Eukaryotes corresponding to about 5% of the human proteome. This study shows the importance of the loss phenomenon of well-established genes in the genome evolution. The majority of losses are described for the first time. One feature of this thesis is that losses without origin sequences signal and those associated with pseudogenes are both discussed and analyzed. Moreover, loss and pseudogenization events are highlighted in an evolutionary context. The provision of functional information and comparative studies could allow understanding the underlying phenomena of such losses.

Key words: unitary gene loss, unitary pseudogene, pseudogenization, automation, phylogeny, human