Détecter et exploiter les flux de gènes dans une biodiversité majoritairement inconnue

Théo Tricou

Université Claude Bernard Lyon 1

N°d'ordre NNT : 2021LYSE1318

THÈSE DE DOCTORAT DE L'UNIVERSITÉ DE LYON

opérée au sein de
l'Université Claude Bernard Lyon 1

École doctorale n°341
Évolution, Écosystèmes, Microbiologie, Modélisation

Discipline: Évolution

Soutenue publiquement le 14/12/2021 par

Théo Tricou

Détecter et exploiter les flux de gènes dans une biodiversité majoritairement inconnue

---

Devant le jury composé de :

Dr. Stéphanie BEDHOMME, CR, CNRS, CEFE	Rapportrice
Dr. Maud TENAILLON, DR, CNRS, GQE-Le Moulon	Rapportrice
Dr. Karine Van DONINCK, Pr., Université Libre de Bruxelles	Rapportrice
Dr. Céline BROCHIER-ARMANET, Pr., UCBL Lyon 1	Présidente du jury
Dr. Damien de VIENNE, CR, CNRS, LBBE	Directeur de thèse
Dr. Éric TANNIER, DR, INRIA, LBBE	Co-Directeur de thèse

Université Claude Bernard – LYON 1

Président de l'Université	M. Frédéric FLEURY
Président du Conseil Académique	M. Hamda BEN HADID
Vice-Président du Conseil d'Administration	M. Didier REVEL
Vice-Président du Conseil des Etudes et de la Vie Universitaire	M. Philippe CHEVALLIER
Vice-Président de la Commission de Recherche	M. Petru MIRONESCU
Directeur Général des Services	M. Pierre ROLLAND

COMPOSANTES SANTE

Département de Formation et Centre de Recherche en Biologie Humaine	Directrice : Mme Anne-Marie SCHOTT
Faculté d'Odontologie	Doyenne : Mme Dominique SEUX
Faculté de Médecine et Maïeutique Lyon Sud - Charles Mérieux	Doyenne : Mme Carole BURILLON
Faculté de Médecine Lyon-Est	Doyen : M. Gilles RODE
Institut des Sciences et Techniques de la Réadaptation (ISTR)	Directeur : M. Xavier PERROT
Institut des Sciences Pharmaceutiques et Biologiques (ISBP)	Directrice : Mme Christine VINCIGUERRA

COMPOSANTES & DEPARTEMENTS DE SCIENCES & TECHNOLOGIE

Département Génie Electrique et des Procédés (GEP)	Directrice : Mme Rosaria FERRIGNO
Département Informatique	Directeur : M. Behzad SHARIAT
Département Mécanique	Directeur M. Marc BUFFAT
Ecole Supérieure de Chimie, Physique, Electronique (CPE Lyon)	Directeur : Gérard PIGNAULT
Institut de Science Financière et d'Assurances (ISFA)	Directeur : M. Nicolas LEBOISNE
Institut National du Professorat et de l'Education	Administrateur Provisoire : M. Pierre CHAREYRON
Institut Universitaire de Technologie de Lyon 1	Directeur : M. Christophe VITON
Observatoire de Lyon	Directrice : Mme Isabelle DANIEL
Polytechnique Lyon	Directeur : Emmanuel PERRIN
UFR Biosciences	Administratrice provisoire : Mme Kathrin GIESELER
UFR des Sciences et Techniques des Activités Physiques et Sportives (STAPS)	Directeur : M. Yannick VANPOULLE
UFR Faculté des Sciences	Directeur : M. Bruno ANDRIOLETTI

Résumé

En première approximation, toutes les espèces sont éteintes. Et celles qui ne le sont pas nous sont pour la plupart inconnues.

Durant Les 4 milliards d'années d'évolution ayant engendré cette immense biodiversité, des organismes se sont transmis du matériel génétique, soit verticalement, de manière généalogique, soit horizontalement, par des transferts entre espèces distinctes. Cette deuxième composante est maintenant reconnue comme une force évolutive majeure ayant remodelé les génomes tout au long de l'évolution.

L'observation conjointe d'une biodiversité largement inconnue et de l'existence de flux génomiques horizontaux implique que certains gènes, présents dans les espèces observées aujourd'hui, sont vraisemblablement apparus et ont évolué pendant un certain temps dans des organismes aujourd'hui éteints ou encore inconnus. Il est donc légitime de se demander si l'étude des flux de gènes, menée habituellement sans prendre en compte ces lignées fantômes, ne peut pas conduire à des conclusions erronées.

Au cours de ma thèse, j'ai développé et utilisé des approches *in silico* pour explorer cette question.

J'ai tout d'abord collaboré au développement d'un outil bioinformatique, *Zombi*, permettant de simuler les composantes verticales et horizontales de l'évolution des génomes le long des branches d'un arbre d'espèces tout en considérant des lignées fantômes. J'ai ensuite utilisé cet outil pour explorer l'influence des lignées fantômes sur certains résultats en évolution moléculaire, dans trois directions. Premièrement j'ai examiné les erreurs commises lors de l'interprétation de la statistique-D pour détecter les introgressions si on néglige les espèces fantômes. J'ai montré que les interprétations erronées, qui étaient considérées comme des exceptions, étaient en réalité probablement la règle. Deuxièmement, j'ai ré-analysé trois études qui ont utilisé les longueurs de branches pour l'analyse des flux génomiques horizontaux et j'ai montré que leurs conclusions s'inversaient lorsque l'on considérait que des lignées fantômes pouvaient être impliquées. Enfin, j'ai fourni une preuve du concept que les flux de gènes provenant de lignées fantômes pouvaient être utilisés pour révéler l'existence et la nature de certaines lignées fantômes. La détection des flux de gènes pourrait ainsi de remplacer les fossiles lorsqu'ils sont indisponibles, comme chez les micro-organismes.

L'apport de cette thèse est à la fois de montrer l'importance de prendre en compte la biodiversité fantôme dans l'étude des flux de gènes, de fournir des outils

pour cette prise en compte, et donc d'offrir un nouveau cadre de travail pour la recherche future sur les flux de gènes dans tous les domaines du vivant.

Abstract

Our current view of Biodiversity is only a glimpse of all that has existed and exists on earth. Only a few million species have been described, which is a small percentage of those suspected of living today, and an even smaller fraction if we consider extinct diversity. This disparity may be stronger in microbes than in eukaryotes because they hardly fossilize and because their estimated diversity is several orders of magnitude greater than their known biodiversity.

The evolutionary history of Life on Earth that gave rise to this huge biodiversity has two components : one vertical, through the transmission of genetic material in a genealogical manner ; the other horizontal, through genetic transmission across species boundaries. This second component is now widely accepted as a major evolutionary force that constantly reshaped genomes throughout evolution.

The joint observation that most species are unknown or extinct, and that horizontal genomic flows occurred regularly in the history of Life, implies that some genes that are present in today's species may have originated and evolved for some time in organisms that are now extinct, are yet unknown or are simply not considered. As a consequence, one may wonder whether the study of gene flows without taking into account these *ghost lineages* may not lead to spurious conclusions.

During my PhD, I developed and used *in silico* approaches to explore this question.

Before all, to better comprehend the impact of ghost lineages on evolutionary processes there was a need for a dedicated tool. I collaborate on the development of the simulator Zombi, which allows us to simulate the vertical and horizontal components of the evolution of genomes along the branches of a species tree while considering ghost lineages. This tool is used for better characterizing the importance of considering ghost lineages when working on genomic flows. First, I confirm what was foreseen but overlooked, that ghost lineages can lead to the wrong identification of both the donor and the recipient of introgressions when using the popular D-statistic method, and prove that this effect is certainly massive. Second, I reanalyze three studies that used branch-lengths to deal with horizontal genomic flows and I show that their conclusions are reversed when considering that ghost lineages may be involved. Third, I provide the proof of concept that gene flows from ghost lineages are not solely a source of noise but can also be used to explore hidden diversity in phylogenies. I demonstrate that the detection of horizontal gene transfers contain information on the existence and the nature of ghost lineages and that detecting horizontal transfers may thus

be a way to explore hidden diversity when fossils are unavailable like in microbes.

This thesis contributes to both showing the utmost importance of taking into account ghost lineages in the study of gene flows and providing a new dedicated tool for simulating this ghost biversity. Thus it offers a new framework for future research on gene flow across all biodiversity.

Table des matières

Table des figures

1

Introduction

Contents

Dans son acception la plus large, la biodiversité désigne désigne l'ensemble des

espèces qui existent et ont existé sur Terre. Mais seule une infime proportion de cette diversité est connue. La majorité des espèces est maintenant éteinte et n'a laissé aucune trace, et on ne connaît qu'une fraction des espèces vivant sur Terre aujourd'hui. Par ailleurs, l'histoire de la vie sur terre, vieille de près de 4 milliards d'années, a été marquée par de nombreux échanges génétiques (ou flux de gènes) entre espèces distinctes, à toutes les époques et dans tous les groupes taxonomiques. Cette "histoire horizontale", par opposition à "l'histoire verticale" décrivant la transmission génétique classique d'ascendants à descendants, apparaît comme une force majeure en évolution. L'impact que peuvent avoir les espèces fantômes, *i.e.* la biodiversité éteinte, inconnue, et non considérée, sur l'identification et la caractérisation de ces flux de gènes a été largement sous-estimé, et donc sous-étudié, par la communauté scientifique. Cette thèse s'attelle à la tâche de montrer l'importance de prendre en compte les lignées fantômes lors de l'étude des flux de gènes.

Dans cette introduction, je présente tout d'abord l'ampleur de notre méconnaissance de la biodiversité. Je décris ensuite les différents mécanismes de flux de gènes ainsi que les méthodes qui permettent de les détecter. Enfin, j'explique pourquoi et comment les espèces fantômes peuvent influencer les résultats et les interprétations des méthodes de détection des flux de gènes.

1.1 Une biodiversité largement inconnue

Le terme "biodiversité" est imaginé par Wilson en 1988 (Wilson and Peter, 1988) et utilisé dans une publication scientifique pour la première fois un an plus tard (Wilson, 1989). Ce terme, qui provient de la contraction des termes "biological" (ou "biotic") et "diversity" désigne initialement la diversité du monde vivant à tous les niveaux : celui des gènes, des individus, des populations, des espèces et des écosystèmes. Il est communément employé pour parler de notions plus précises comme la richesse en espèces dans une niche écologique ou la diversité en espèces dans une communauté. Plusieurs autres termes sont utilisés pour parler de cette diversité du vivant, comme la "biodiversité globale", la "diversité absolue", ou encore la "richesse en espèces". Par commodité, dans ce manuscrit j'ai choisi d'utiliser le terme de "biodiversité" pour parler du nombre total d'espèces distinctes qui ont existé et qui existent actuellement sur Terre.

1.1.1 Combien d'espèces ?

En 2021, on comptabilise un peu plus de 2 millions d'espèces vivantes, identifiées et décrites, majoritairement des eucaryotes. Nos connaissances sont hétérogènes, avec ∼1 million d'espèces d'insectes (Froese *et al.*, 2019) et ∼400 000 espèces de plantes connues (Christenhusz and Byng, 2016), mais seulement quelques milliers de procaryotes[1] (Thomas and Nielsen, 2005, Figure 1.1).

Il existe plusieurs archives et catalogues qui tentent de répertorier nos connaissances sur la biodiversité actuelle. The Catalogue of life (Roskov *et al.*, 2000), Encyclopedia of Life (Parr *et al.*, 2014), GBIF[2] et Wikispecies[3] ont pour mission de créer des listes les plus exhaustives possibles des espèces connues. La taille de ces listes varie grandement du fait des différents critères utilisés pour l'ajout d'une nouvelle espèce, par exemple la définition utilisée pour une «espèce» ou la quantité de preuves nécessaires pour ajouter un nouvel organisme dans ces listes.

Quoiqu'il en soit, il est clair que le nombre d'espèces répertoriées dans ces bases de données est largement inférieur au nombre d'espèces existant aujourd'hui sur Terre (Mora *et al.*, 2011). Connaître le nombre d'espèces qui existent est d'une importance toute particulière pour l'étude de la vie sur Terre, pour comprendre les mécanismes de son apparition mais aussi ceux liés à sa disparition. Mesurer l'étendue de la biodiversité est d'autant plus important à l'heure où des changements environnementaux majeurs principalement dus aux activités humaines, tels que les perturbations des paysages, la destruction des habitats naturels et les changements climatiques, conduisent ou ont déjà conduit à des extinctions massives (McGill *et al.*, 2015; Ceballos *et al.*, 2017).

1. J'utilise dans cette introduction le terme "procaryote" en étant conscient qu'il est tombé en désuétude parce qu'il ne s'agit pas d'un groupe monophylétique et que son nom évoque un caractère primitif, deux choses à éviter quand on est évolutionniste. Malgré tout, l'état de la connaissance dans le domaine que je décris est différent pour les eucaryotes et les autres organismes, auxquels il peut donc être utile de donner un nom. Remplacer par "bactéries et archées" n'est plus vraiment pertinent car archée a les mêmes défauts que procaryote (probablement non monophylétique (Guy and Ettema, 2011; Williams *et al.*, 2012; Hug *et al.*, 2016; Williams *et al.*, 2020) et étymologie qui évoque un caractère primitif). Énumérer les clades monophylétiques dont il s'agit serait un peu fastidieux (bactéries et quelques clades d'archées) et n'aurait pas de sens si c'est pour désigner finalement le monde vivant non eucaryote par un autre nom. Donc procaryote est le plus commode pour mon introduction, et correspond à la littérature dont je fais la revue.
2. GBIF.org (2021), Home Page : https://www.gbif.org.
3. Wikispecies (2021), Home page : https://species.wikimedia.org/wiki/Main_Page

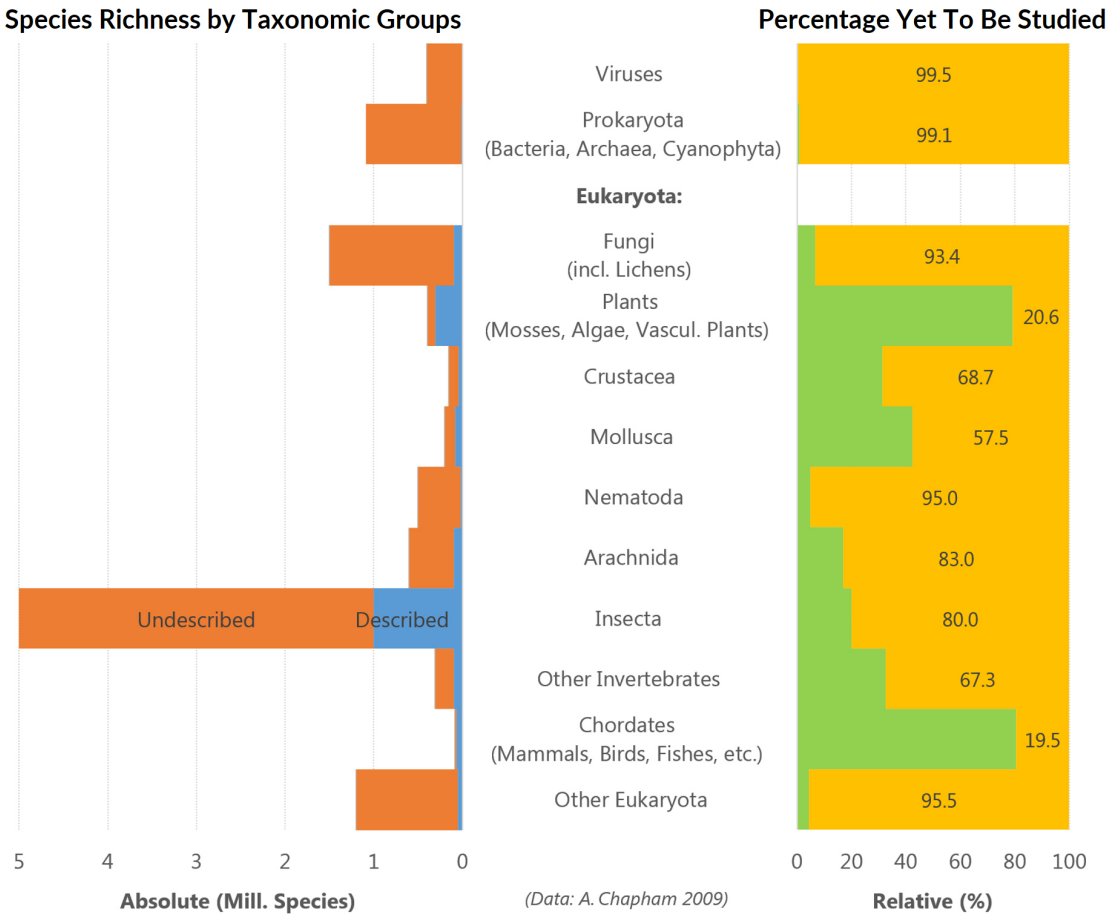


Figure *1.1* – *La distribution du nombre d'espèces connues et décrites comparé a la biodiversité vivante aujourd'hui sur Terre estimée, regroupées par grands groupes taxonomiques. Nombre absolu d'espèces sur la gauche (orange=nombre estimé d'espèces restant à décrire, bleu=déjà décrites). À droite : pourcentage d'espèces déjà décrites (vert) et estimées inconnues (jaune). (copié depuis Chapman, 2009)*

Selon certains auteurs, il existerait actuellement entre 3 et 100 millions d'espèces eucaryotes (May, 2010). Pour d'autres, c'est le chiffre de 8.7 millions (±1.3 million) d'espèces eucaryotes qui est retenu (Mora *et al.*, 2011), dont 2.2 millions (±0.18 million) d'espèces marines. Selon cette même étude, 86% des espèces sur Terre et 91% dans l'océan seraient encore non décrites. Ces chiffres sont en accord avec ceux publiés par Chapman (2009) dans un rapport mis à jour régulièrement et qui détaille, pour chaque grand groupe taxonomique, le nombre d'espèces décrites et le nombre d'espèces estimées (Figure 1.1 colonne de droite). Selon cette étude, 95% des fungi resteraient à décrire, ainsi que 80% des insectes et des arachnides et 20% des plantes ou des chordés, deux des groupes les plus étudiés par la communauté scientifique.

Chez les procaryotes, les estimations du nombre d'espèces sont encore plus variables et incertaines que chez les eucaryotes. Une des raisons est que la notion d'espèce est souvent plus difficile à définir. Une autre est qu'une majorité de ces organismes ne peuvent pas être cultivés, rendant leur description complexe (Lewis *et al.*, 2021, mais voir encadré 1 pour les apports de la métagenomique). Selon Chapman (2009), seules quelques dizaines de milliers d'espèces procaryotes sont décrites pour un peu moins de 1.5 millions d'espèces estimées (Figure 1.1). D'autres auteurs (Locey and Lennon, 2016) soupçonnent même l'existence de milliards (10^9) voire de billions (10^{12}) d'espèces procaryotes. Dans les deux cas, moins de 1% de la biodiversité procaryote a été décrite et elle pourrait donc être bien plus grande (de plusieurs ordres de grandeur) que celle des eucaryotes.

On notera des écarts importants dans les différentes estimations du nombre d'espèces présentées plus haut. Cela s'explique par les différentes méthodes utilisées pour ces estimations, qui font invariablement des approximations et ont des limites, parfois bien identifiées (Figure 1.2). Tous les chiffres que j'ai présentés jusqu'ici sont donc à prendre avec précaution.

Case Study	Limitations
Macroecological patterns	
Body size frequency distributions. By extrapolation from the frequency of large to small species, May [7] estimated 10 to 50 million species of animals.	May [7] suggested that there was no reason to expect a simple scaling law from large to small species. Further studies confirmed different modes of evolution among small species [4] and inconsistent body size frequency distributions among taxa [4].
Latitudinal gradients in species. By extrapolation from the better sampled temperate regions to the tropics, Raven [10] estimated 3 to 5 million species of large organisms.	May [2] questioned the assumption that temperate regions were better sampled than tropical ones; the approach also assumed consistent diversity gradients across taxa which is not factual [4].
Species-area relationships. By extrapolation from the number of species in deep-sea samples, Grassle & Maciolek [13] estimated that the world's deep seafloor could contain up to 10 million species.	Lambshead & Bouchet [12] questioned this estimation by showing that high local diversity in the deep sea does not necessarily reflect high global biodiversity given low species turnover.
Diversity ratios	
Ratios between taxa. By assuming a global 6:1 ratio of fungi to vascular plants and that there are ~270,000 species of vascular plants, Hawksworth [20] estimated 1.6 million fungi species.	Ratio-like approaches have been heavily critiqued because, given known patterns of species turnover, locally estimated ratios between taxa may or may not be consistent at the global scale [3,12] and because at least one group of organisms should be well known at the global scale, which may not always be true [15]. Bouchet [6] elegantly demonstrated the shortcomings of ratio-based approaches by showing how even for a well-inventoried marine region, the ratio of fishes to total multicellular organisms would yield ~0.5 million global marine species whereas the ratio of Brachyura to total multicellular organisms in the same sampled region would yield ~1.5 million species.
Host-specificity and spatial ratios. Given 50,000 known species of tropical trees and assuming a 5:1 ratio of host beetles to trees, that beetles represent 40% of the canopy arthropods, and that the canopy has twice the species of the ground, Erwin [9] estimated 30 million species of arthropods in the tropics.	
Known to unknown ratios. Hodkinson & Casson [18] estimated that 62.5% of the bug (Hemiptera) species in a sampled location were unknown; by assuming that 7.5%–10% of the global diversity of insects is bugs, they estimated between 1.84 and 2.57 million species of insects globally.	
Taxonomic patterns	
Time-species accumulation curves. By extrapolation from the discovery record it was estimated that there are ~19,800 species of marine fishes [23] and ~11,997 birds [22].	This approach is not widely applicable because it requires species accumulation curves to approach asymptotic levels, which is only true for a small number of well-described taxa [22–23].
Authors-species accumulation curves. Modeling the number of authors describing species over time allowed researchers to estimate that the proportion of flowering plants yet to be discovered is 13% to 18% [21].	This is a very recent method and the effect of a number of assumptions remains to be evaluated. One is the extent to which the description of new species is shifting from using taxonomic expertise alone to relying on molecular methods (particularly among small organisms [26]) and the other that not all authors listed on a manuscript are taxonomic experts, particularly in recent times when the number of coauthors per taxa described is increasing [21,38], which could be due to more collaborative research [38] and the acknowledgment of technicians, field assistants, specimen collectors, and so on as coauthors (Philippe Bouchet, personal communication).
Analysis of expert estimations. Estimates of ~5 million species of insects [15] and ~200,000 marine species [14] were arrived at by compiling opinion-based estimates from taxonomic experts. Robustness in the estimations is assumed from the consistency of responses among different experts.	Erwin [5] labeled this approach as "non-scientific" due to a lack of verification. Estimates can vary widely, even those of a single expert [5,6]. Bouchet [6] argues that expert estimations are often passed on from one expert to another and therefore a robust estimation could be the "same guess copied again and again".

doi:10.1371/journal.pbio.1001127.t001

Figure *1.2* – *Tableau des méthodes pour estimer le nombre global d'espèces et leurs limites (copié depuis Mora* et al., *2011).*

Box 1 : La métagénomique et l'exploration de la matière noire biologique

Récemment, l'exploration de la matière noire biologique, c'est à dire de l'ensemble de la diversité des micro-organismes non cultivables, est devenue en partie accessible grâce à l'avènement de la métagénomique qui exploite les technologies de séquençage de nouvelle génération (NGS). Cette méthode a permis entre autre la découverte de nombreux procaryotes jusque-là inconnus, parmi lesquels le groupe des Candidate Phyla Radiation (CPR, Figure 1.3) qui représenterait plus de 15% de la diversité bactérienne totale (Brown *et al.*, 2015; Hug *et al.*, 2016).

Chez les archées, les approches métagénomiques et le séquençage de cellules uniques ont permis d'obtenir le génome d'organismes auparavant non cultivés et inconnus, et de nombreuses nouvelles lignées d'archées ont ainsi été découvertes. Par exemple les *Nanohaloarchaea* (Narasingarao *et al.*, 2012), les **Aigarchaeota** (Nunoura *et al.*, 2011) ou les *Lokiarchaeota*. Ce dernier phylum a été découvert par séquençage métagénomique d'Archées des profondeurs marines non cultivables et il est le premier représentant du superphylum Asgard, groupe frère des eucaryotes (Spang *et al.*, 2015).

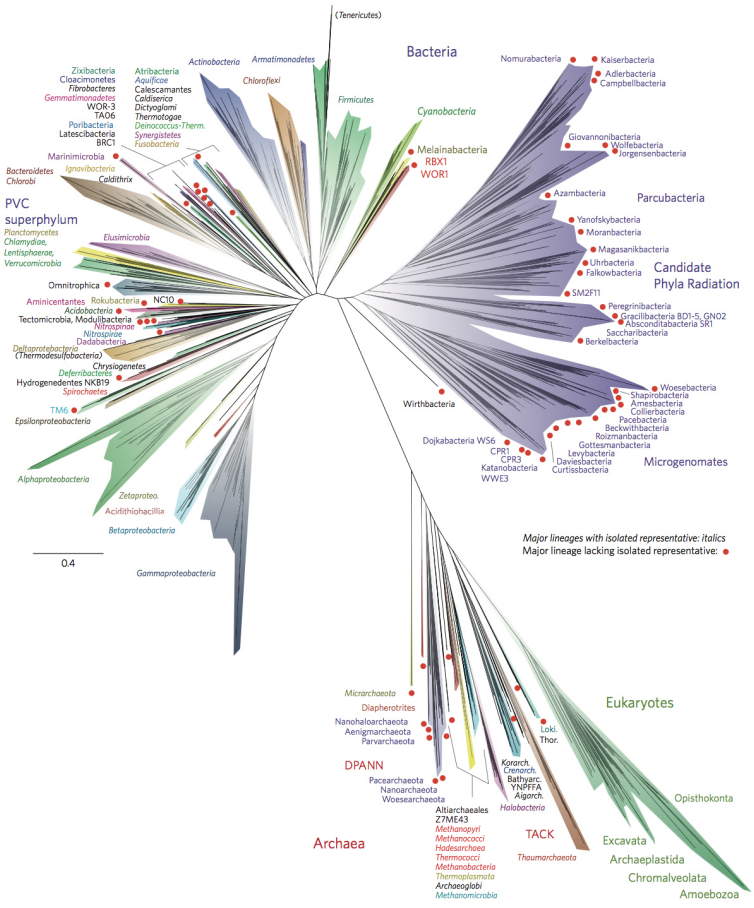


Figure *1.3* – *Arbre du vivant construit par Hug* et al. *(2016) contenant 92 phyla bactériens, 26 phyla archées, et tous les phyla eucaryotes. Le phylum des CPR est représenté en violet.*

Le manque de connaissances sur la biodiversité devient encore plus vaste si l'on prend en compte les espèces qui sont maintenant éteintes. Comme le remarquait Raup (1991) :

> *"To a first approximation, all species are extinct"*

Cette affirmation découle d'un calcul simple : si de façon régulière (background extinction) entre 0.01 et 1.0% d'espèces sont perdues par décennie (Costello *et al.*, 2013), sachant que l'apparition du premier organisme unicellulaire sur Terre remonterait à près de 4 milliards d'années (Dodd *et al.*, 2017), alors 99.99% des espèces ayant vécu sur Terre seraient désormais éteintes (Raup, 1986; Kunin and Gaston, 2012). Bien sûr, une espèce éteinte n'est pas forcément inconnue, car l'étude des restes fossiles permet sa détection et sa caractérisation. Au cours des 200 dernières années, des dizaines de milliers de fossiles de plus de 1000 espèces de dinosaures ont été trouvés, et en 2013, on comptabilisait presque 250 000 espèces fossiles décrites (Prothero, 2013). Cependant, le registre fossile est, lui aussi, très incomplet et fortement hétérogène dans les temps géologiques et à travers les clades et ne représenterait qu'une petite proportion de toutes les espèces qui ont existé. Enfin, une grande part des espèces fossiles identifiées seraient invalides ou redondantes. Par exemple, des 4861 fossiles de mammifères nord-américains décrits, entre 24 et 31 % seraient invalides (Alroy, 2002). Enfin, les procaryotes ne laissent pas de traces fossiles (Waggoner, 1996), à l'exception des cyanobactéries (Schopf, 2012) dont l'organisation en biofilm et l'activité métabolique entraîne, par une précipitation des carbonates, l'édification de structures que l'on appelle des stromatolithes. La plus ancienne a été datée à 3.7 Milliards d'années (Allen *et al.*, 2019).

En conclusion, nous ne connaissons qu'une qu'une infime fraction de la biodiversité actuelle, et une fraction encore plus infime de la biodiversité éteinte (Fig 1.4A). On pourrait donc compléter la citation de Raup (1991) ainsi :

> *"To a first approximation, all species are extinct",* ***and most of those that are not, are still unknown!***

On qualifie généralement de biodiversité fantôme (voir l'encadré 2 pour l'origine de ce terme) l'ensemble de la biodiversité inconnue, qu'elle soit éteinte ou non (Figure 1.4B), par opposition à la biodiversité connue, décrite plus haut. Dans ce manuscrit, je propose d'utiliser une définition légèrement différente de ce terme. En effet, du point de vue d'un groupe d'espèces étudiées (Figure 1.4A), toutes les autres espèces, qu'elles soient ou non connues, sont des fantômes. Je qualifierai donc de "biodiversité fantôme" l'ensemble des espèces, éteintes ou non, qui ne sont pas incluses dans une étude donnée (Figure 1.4 C).

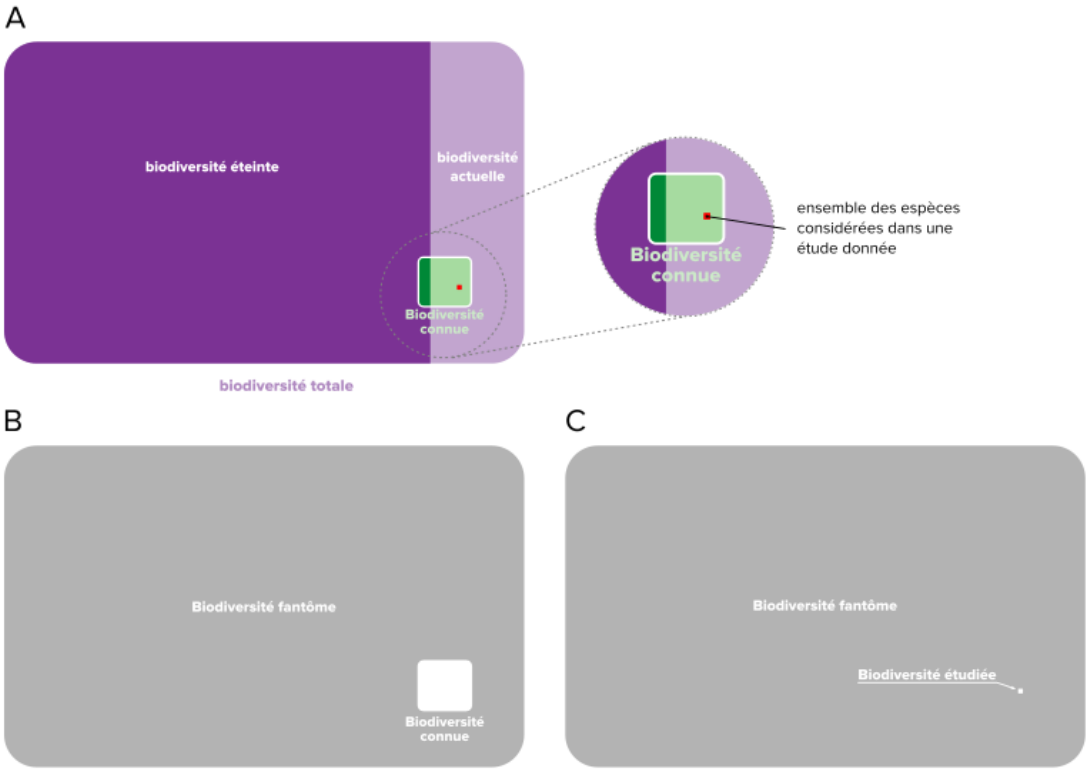


Figure *1.4* – *Catégorisation de la biodiversité et définitions de la biodiversité fantôme. A. La biodiversité totale contient l'ensemble des espèces éteintes et non éteintes. La biodiversité connue représente une faible part ce cette biodiversité. C'est dans cette biodiversité connue que sont échantillonnées les espèces étudiées par les scientifiques. On peut citer deux formes de biodiversité fantôme : celle qui représente tout ce qui est inconnu (B) ou celle qui représente tout ce qui n'est pas considéré dans le contexte d'une étude (C). C'est cette dernière définition qui a été utilisée dans ce manuscrit.*

Box 2 : L'origine du terme "espèces fantômes"

C'est du domaine de la paléontologie que vient le terme d'"espèces fantômes" (Ghost lineages en anglais, Norell MA, 1992). Cela représente le ou les chaînon(s) manquant(s) entre deux registres fossiles ou entre une observation fossile et l'observation des premiers descendants de ce clade fossile. De ce fait, il est important de noter la distinction qui existe entre le terme "ghost lineages" utilisé en paléontologie et celui que nous allons employer pour parler de l'ensemble de la biodiversité manquante et/ou inconnue dans un arbre phylogénétique, tel que l'arbre du vivant (Wills, 2007).

L'exemple classique utilisé en paléontologie pour illustrer cette notion d'espèces fantômes est celui des Cœlacanthes. Ce sont des poissons pulmonés apparentés aux tétrapodes primitifs. Pendant de nombreuses années, ce clade était considéré comme éteint. Les seules traces d'existence de ce taxon étaient des relevés fossiles datant du Dévonien (-420 à -360 millions d'années) jusqu'au Crétacé (-145 Ma à -66 Ma), aucun fossile n'ayant été retrouvé dans des sédiments plus jeunes. Cependant, en 1938 une espèce vivante apparentée aux Cœlacanthes fut découverte le long des côtes Africaines et plus récemment une population en Indonésie (Mahé *et al.*, 2021). Ce clade que l'on pensait éteint jusqu'alors, mais pour lequel des individus actuels ont été trouvés, est nommé Lazarus (Figure 1.5). Ainsi l'intervalle de 80 millions d'années séparant les relevés fossiles et le taxon Lazarus est forcément comblé par des espèces dont aucun fossile n'est connu et que l'on nomme alors des espèces "fantômes".

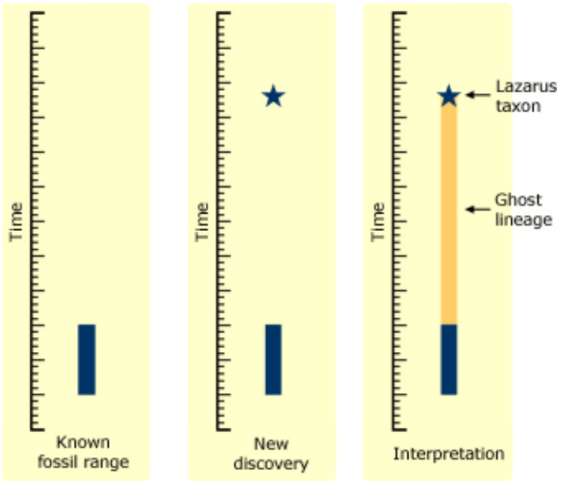


Figure *1.5* – *Les lignées fantômes en paléontologie. À gauche : Les fossiles d'un organisme se trouvent dans des roches d'une tranche d'âge limitée. Au milieu : ensuite, un fossile de l'organisme est ensuite découvert dans des roches des millions d'années plus jeunes. À droite : cela suggérerait que l'organisme ne s'est pas éteint mais qu'il y a une longue lacune dans le registre fossile de l'organisme. Figure tirée de Matt Wedel (2010)*

1.1.2 L'arbre du vivant.

La biodiversité dont nous avons parlé dans la section précédente est le résultat de l'évolution du vivant et de l'enchaînement d'événements de spéciations — processus par lequel une lignée ancestrale se sépare en deux lignées soeurs — et d'extinctions. Cette évolution buissonnante peut être représentée sous la forme d'un arbre phylogénétique. La racine de cet arbre représente l'ancêtre commun à tous les organismes, le plus récent dont seraient issues toutes les espèces connues, aussi nommé "dernier ancêtre commun universel" (ou LUCA pour Last Universal Common Ancestor). Chaque embranchement représente un événement de spéciation et chaque feuille correspond à une espèce. Les liens évolutifs connus ou suspectés de l'ensemble de la biodiversité peuvent être représentés dans ce qu'on appelle l'arbre du vivant (Figure 1.6). De la même façon qu'il existe plusieurs listes cataloguant la biodiversité, il existe plusieurs versions de cet arbre. La taxonomie du NCBI (National Center for Biotechnology Information) comprend plus de 1.4 millions d'espèces et permet de représenter un arbre du vivant consensus représentant les liens de parenté largement acceptés dans la littérature taxonomique actuelle. Cette version de l'arbre du vivant est visualisable avec l'outil Lifemap (de Vienne, 2016) (Figure 1.6). Une autre version de l'arbre portée par le projet Open Tree of Life (Hinchliff *et al.*, 2015) synthétise les arbres phylogénétiques de la littérature et y ajoute des données taxonomiques pour créer un arbre du vivant de plus de 2.3 millions de feuilles. Comme vu précédemment, notre connaissance actuelle de la biodiversité, et donc de son histoire évolutive, est très fragmentaire ; au point que l'arbre du vivant tel que nous le connaissons aujourd'hui est très majoritairement incomplet. Nous verrons dans la suite que cette vision incomplète de l'arbre du vivant, et des arbres phylogénétiques en général, a des conséquences importantes sur notre capacité à détecter des événements pourtant majeurs en évolution : les flux de gènes horizontaux.

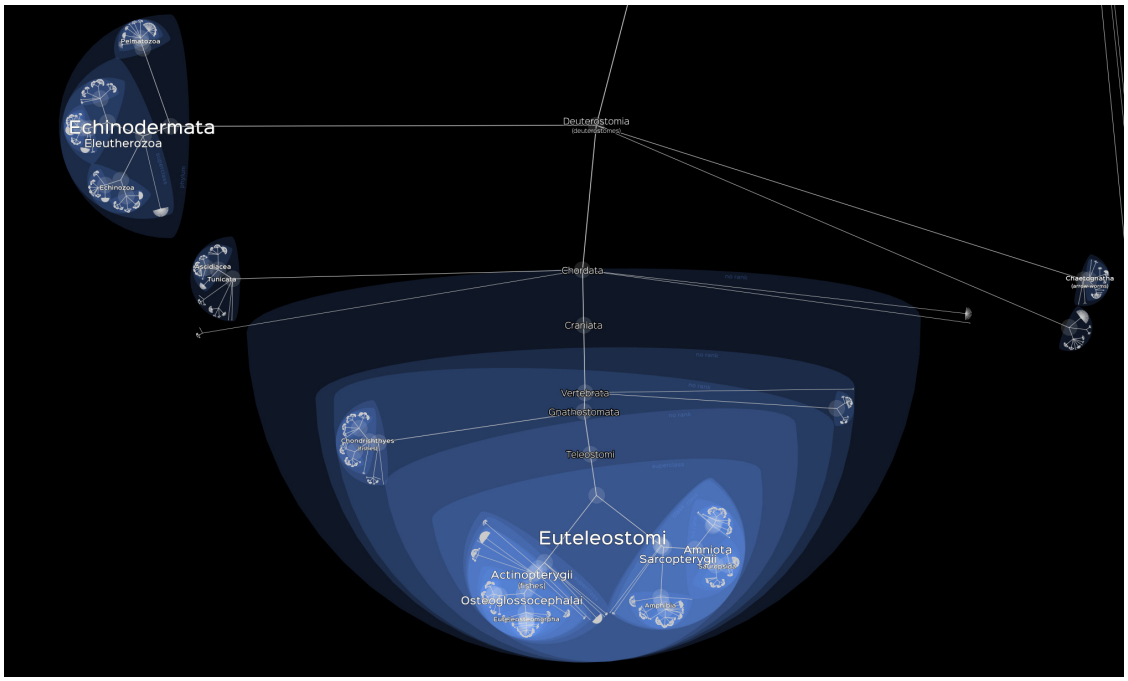


Figure *1.6* – *Représentation d'un morceau de l'arbre du vivant, ici les Deutérostomiens, avec l'outil Lifemap (de Vienne, 2016)*

1.2 Des flux de gènes omniprésents dans l'arbre du vivant

L'arbre du vivant, et les arbres phylogénétiques en général, représentent à la fois la diversité des espèces, une façon de les classer, et cette succession d'évènements de spéciation et d'extinction Cependant, les gènes portés par les organismes ne suivent pas toujours les branches (évolution verticale) de cet arbre et peuvent avoir des histoires évolutives plus complexes. Les phénomènes de flux de gènes (évolution horizontale) entre organismes sont un des mécanismes qui brise la barrière de l'espèce, donnant une forte porosité à travers l'ensemble du vivant et surtout entre la zone "fantôme" de la biodiversité et les zones étudiées (1.4C).

1.2.1 Des flux géniques à toutes les échelles

Les évènements de flux de gènes sont des phénomènes de transmission de matériel génétique à travers les frontières des espèces qui peuvent se produire à

toutes les échelles dans l'arbre du vivant. En outre, ces échanges sont un moteur majeur de l'évolution que ce soit chez les procaryotes (Syvanen, 1985; Koonin *et al.*, 2001; Boto, 2010; Hall *et al.*, 2017) ou chez les eucaryotes (Keeling and Palmer, 2008; Szöllősi *et al.*, 2015). On distingue trois grandes catégories de mécanismes moléculaires à l'origine de ces flux de gènes. **Les transferts horizontaux** de gènes correspondent au transfert et à l'intégration de matériel génétique d'un organisme à un autre, étudié principalement chez les procaryotes mais apparemment possible aussi chez les eucaryotes. **Les introgressions** sont dues à l'hybridation entre individus de deux populations plus ou moins éloignées génétiquement, ces hybridations sont courantes chez les organismes à reproduction sexuée. Enfin **l'endosymbiose** désigne l'acquisition de matériel génétique par une cellule hôte par la fusion de la cellule d'un autre organisme procaryote. Ces mécanismes sont présentés en détails dans les 3 parties suivantes.

Le transfert horizontal de gènes

La "transformation" des bactéries, dont la cause a été identifiée plus tard comme des transferts d'ADN, a été décrite en premier par Frederick Griffith en 1928 (Griffith, 1928). Il a montré que des bactéries pneumocoques non virulentes pouvaient devenir pathogènes par simple contact avec des bactéries virulentes. C'est bien plus tard, en 1944, que ce mécanisme a été identifié comme étant un échange d'ADN entre organismes (Avery *et al.*, 1944).

Il existe plusieurs mécanismes qui permettent le transfert de matériel génétique entre espèces (Thomas and Nielsen, 2005) :

1. Transformation : certaines espèces de procaryotes ont la capacité de siphonner l'ADN présent dans leur entourage et de l'incorporer à leur propre génome ;
2. Transduction : les virus peuvent parfois encapsuler des fragments d'ADN des cellules qu'elles infectent. Par la suite, ces mêmes virus vont pouvoir infecter d'autres organismes avec la possibilité de transmettre ces fragments d'ADN ;
3. Conjugaison : certaines bactéries peuvent s'échanger de l'ADN par le biais de "pili sexuels" qui permettent de créer un lien physique entre les deux organismes ;
4. "Gene transfer agent" : les bactéries ont la capacité de produire des capsides ressemblant à des virus qui contiennent de l'ADN. Ces capsides peuvent être transférées à d'autres bactéries.

Grâce à l'émergence du séquençage génomique, l'ampleur et l'importance de ces transferts pour l'évolution des procaryotes a pu être réalisé. Ces échanges sont une des forces majeures à l'origine de leur évolution. Ils sont fortement impliqués dans la propagation rapide des résistances aux antibiotiques, mais aussi dans l'adaptation à de nouveaux environnements (Ochman *et al.*, 2000). Les phénomènes de flux de gènes sont tellement fréquents, particulièrement les transferts horizontaux omniprésents chez les procaryotes, qu'ils ont conduit certains auteurs à questionner et remettre en cause la notion d'arbre pour décrire l'évolution des ces organismes (Woese, 2004).

Chez les eucaryotes l'importance des transferts horizontaux est souvent considéré comme limitée (Keeling and Palmer, 2008). Il existe des barrières capables de les empêcher, comme par exemple la séparation des cellules germinales et somatiques chez les animaux. Des événements d'échanges d'ADN provenant de bactéries ou de virus vers les plantes, les fungi et même les animaux ont néanmoins été identifiés (Husnik and McCutcheon, 2018). C'est évènement sont pourtant répandu chez les organismes unicellulaires comme les protistes (Huang *et al.*, 2004). Chez les eucaryotes multicellulaires, ces transferts sont plus rare mais pas inexistant. Chez certain animaux la prévalence de ces évènement de transfert est même très haute. Par exemple des centaines gènes chez les *Bdelloid rotifers* auraient été acquis horizontalement (Eyres *et al.*, 2015). Dans les deux sections suivant sont présenté deux mécanisme (que l'on distingue des transferts horizontaux) l'introgression et l'endosymbiose, qui permettent aussi des flux de gènes chez les eucaryotes multicellulaires.

L'introgression

Chez tous les organismes à reproduction sexuée, comme les plantes et les animaux, l'introgression est le résultat de l'hybridation de deux individus appartenant à deux populations distinctes plus ou moins distantes génétiquement (Harrison and Larson, 2014). Cette hybridation est suivie par de nombreux croisements entre les individus hybrides et non-hybrides d'un des parents de sorte qu'il ne reste que quelques traces du génome de l'autre parent dans la population après un certain temps. Bien que cette transmission de matériel génétique se fasse de manière verticale du point de vue des individus (par reproduction sexuée), *in fine* au niveau populationnel (et donc au niveau des espèces comme on travaille souvent avec un seul individu d'une population) elle se reflète par un transfert

d'ADN de l'espèce d'un des parents vers l'autre. Enfin, ces événements sont régulièrement liés à l'acquisition d'allèles et de traits avantageux (Huerta-Sánchez *et al.*, 2014) ou à des événements de radiation adaptative (Edelman *et al.*, 2019). Par exemple, ces mécanismes ont fortement influencé l'évolution des humains modernes. De nombreux événements d'hybridation et d'introgression ont été décrits (Dannemann and Racimo, 2018), le plus documenté étant l'introgression entre l'humain de Néandertal et l'humain moderne non-Africain. Les génomes des descendants de ces non-Africains contiendraient environ 2% d'ADN provenant de *Homo neanderthalensis* (Green *et al.*, 2010). L'ensemble de ces fragments d'ADN représenterait environ 20% du génome de Néandertal (Vernot and Akey, 2014).

L'endosymbiose

L'endosymbiose est un cas particulier de flux de gènes. On parle d'endosymbiose lorsqu'un organisme hôte "internalise" une autre cellule. La cellule ainsi engloutie est alors nommée endosymbionte. De fait les gènes portés par cet endosymbionte apparaissent comme ayant été acquis de façon horizontale comparés au reste du génome de l'hôte. L'apparition de la cellule eucaryote aurait pour origine une cellule hôte pré-eucaryote, ancêtre des cellules eucaryotes, qui aurait intégré par endosymbiose une organelle dérivée d'un endosymbionte Alphaproteobacterien, et aurait par la suite évolué pour former la mitochondrie. En effet, les gènes qui codent pour la mitochondrie sont plus proches de gènes bactériens que de n'importe quels gènes eucaryotes (Andersson *et al.*, 2003; Poole and Gribaldo, 2014; Pittis and Gabaldón, 2016). De façon similaire, la première acquisition des chloroplastes peut être attribuée à un événement d'endosymbiose d'une cyanobactérie. Ce plastide se serait par la suite transféré à travers plusieurs clades par l'enchaînement d'évènements d'endosymbiose (Bhattacharya and Medlin, 1995; Strassert *et al.*, 2021).

1.2.2 La détection des flux de gènes

Quel que soit l'événement considéré (transfert horizontal, introgression, endosymbiose), le résultat est le même : l'incorporation de fragments d'ADN dans le génome d'un organisme receveur, provenant d'une autre espèce ou d'une autre population. Ces fragments auront suivi une histoire évolutive différente de celle

du reste du génome. L'étude de la composition nucléotidique de ces séquences (leur taux de GC par exemple) ou de leur histoire évolutive (à travers la reconstruction de leurs arbres phylogénétiques) peut permettre de les identifier et de les étudier. Il existe une grande variété de méthodes pour détecter les flux de gènes, leurs principes généraux sont présentés dans les sections suivantes et sont illustrés dans la figure 1.7.

Notons que la détection des flux de gènes n'est pas une tâche aisée (Ravenhall *et al.*, 2015) et les méthodes développées dans ce but manquent souvent de précision. En règle générale, différentes méthodes infèrent des événements très différents pour un même jeu de données ce qui rend difficile l'évaluation de nos capacités réelles à détecter de façon exacte ces flux de gènes. Je reviendrai sur ce point dans la discussion.

Les méthodes paramétriques

Les méthodes dites paramétriques sont celles qui exploitent les variations de composition des séquences. Si la composition génétique de fragments d'ADN est éloignée de celle du reste du génome, il est possible que ces fragments aient été acquis de façon horizontale (Daubin *et al.*, 2003). Certaines de ces méthodes se basent sur la composition en nucléotides "GC" de l'ADN. Le contenu en bases G et C des génomes est très variable dans le monde vivant, celui-ci pouvant représenter entre 13% et 80% de l'ADN d'un organisme (Ravenhall *et al.*, 2015). Un fragment d'ADN avec un taux de GC qui diffère significativement du taux de GC du reste du génome (Figure 1.7 (1)) peut être considéré comme un bon candidat de gène acquis horizontalement (Guindon and Perriere, 2001). D'autres méthodes utilisent de la même façon les variations d'usage du code, *i.e.* le fait que certains codons soient préférentiellement utilisés, et que cette préférence ne soit pas la même pour les gènes horizontalement acquis que pour les autres gènes du génome de l'hôte (Becq *et al.*, 2010).

Les méthodes phylogénétiques

L'accumulation des données génomiques a permis le développement de méthodes qui utilisent les arbres phylogénétiques pour détecter les flux de gènes. Ces méthodes comparent les arbres phylogénétiques de gènes et d'espèces et

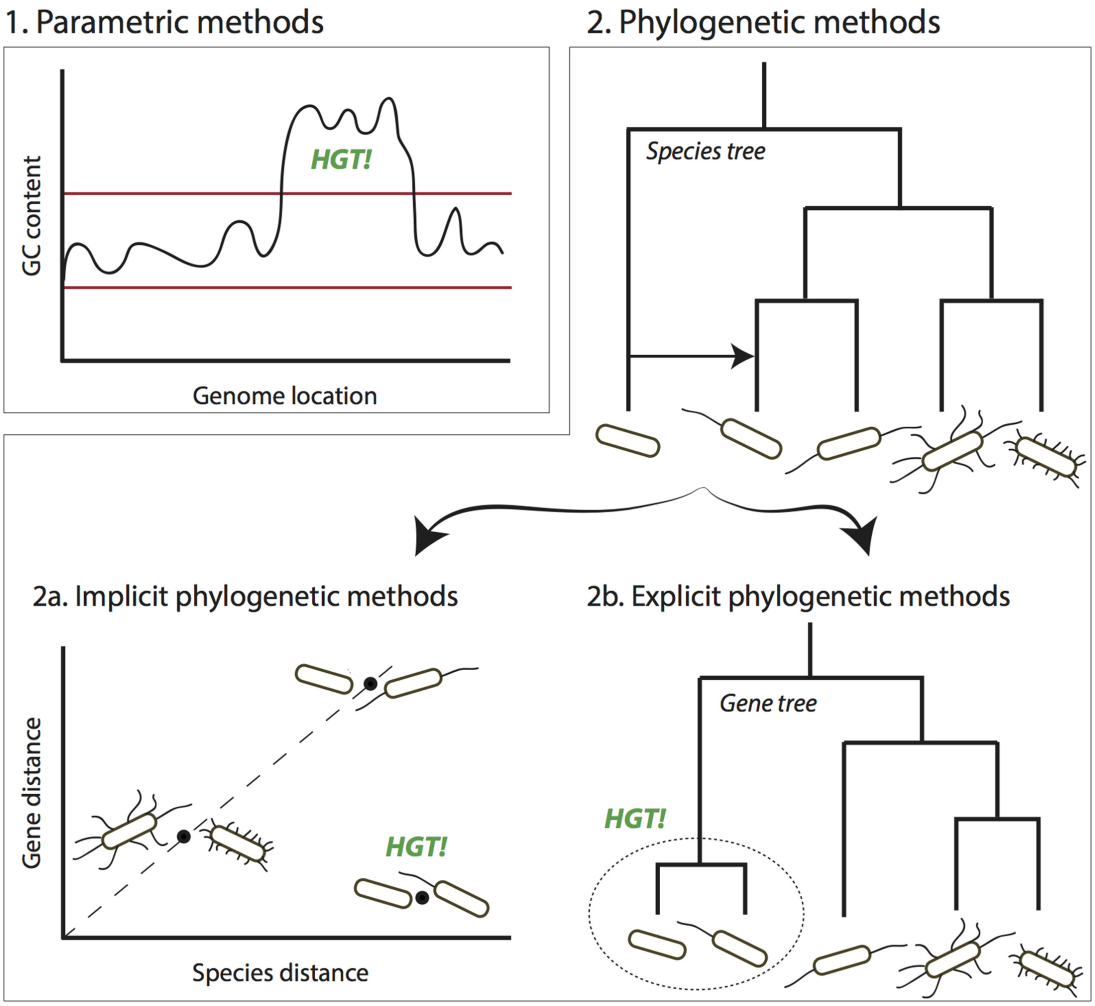


Figure *1.7* – *Aperçu des méthodes d'inférence des transferts horizontaux de gènes et d'autres formes de flux de gènes. (1) Les méthodes paramétriques déduisent les flux de gènes (ici présenté sous la forme d'un HGT) en calculant une statistique, ici le contenu GC, pour une fenêtre glissante et en la comparant à l'ensemble du génome, le contenu en GC moyen indiquée ici entre les deux lignes horizontales rouges. Les régions avec des valeurs atypiques sont déduites comme ayant été transférées horizontalement. (2) Les approches phylogénétiques reposent sur les discordances entre les arbres de gènes et les arbres d'espèces qui résultent des flux de gènes. a : Les méthodes phylogénétiques implicites contournent la reconstruction de l'arbre génétique, par exemple, en examinant les écarts entre les distances par paires entre les gènes et leurs espèces correspondantes. b : Les méthodes phylogénétiques explicites reconstruisent les arbres phylogénétiques et déduisent les événements de flux de gènes susceptibles d'expliquer les discordances observées (copié de Ravenhall et al., 2015).*

utilisent les discordances observées entre leurs topologie (Figure 1.7 (2)) ou les longueurs de branches pour détecter des transferts (et d'autres types d'événements évolutifs). Durant ma thèse, j'ai exclusivement utilisé ce type de méthodes, en particulier le test ABBA-BABA (Green *et al.*, 2010; Durand *et al.*, 2011), une méthode dite de "réconciliation" (ALE, Szöllősi *et al.* (2015)), ainsi que des méthodes utilisant les longueurs de branche des arbres. Je décris ci-dessous le principe général de ces trois approches de détection des flux de gènes. Les méthodes paramétriques, décrites au-dessus, basées sur les variations de la composition des séquences sont implicitement incluses dans les méthodes phylogénétiques puisque les variations de contenu en GC des séquences transférées (par exemple), vont se refléter dans les arbres phylogénétiques des gènes (voir Figure 1.7).

Le test ABBA-BABA Le test ABBA-BABA est devenu populaire suite à son utilisation pour identifier l'évènement d'introgression entre l'Humain de Néandertal et l'Humain moderne non-Africain (Green *et al.*, 2010; Durand *et al.*, 2011). Sa facilité d'utilisation et sa simplicité d'interprétation en font un test couramment utilisé pour inférer des flux de gènes chez les eucaryotes (les papiers de Durand *et al.* (2011) et Patterson *et al.* (2012), dans lesquels le test est formalisé, sont respectivement cités plus de 800 et 1200 fois).

Le test d'introgression ABBA-BABA est aussi appelé statistique-D (D-statistic). Il considère 3 populations composant un groupe interne (P1, P2, P3) et un groupe (ou population) externe (O) avec une topologie en "échelle" ((((P1),P2),P3),O) (Figure 1.8, haut). L'idée est de rechercher deux motifs de polymorphisme nucléotidique (ou SNP pour Single-Nucleotide Polymorphism) biallélique notés "ABBA" et "BABA", avec "A" l'état ancestral (toujours porté par le groupe externe O) et "B" l'allèle dérivé. L'hypothèse nulle que fait ce test est que dans un scénario d'évolution strictement vertical et sans flux de gènes, les deux motifs apparaîtront seulement par tri de lignées incomplet (ILS pour Incomplete Lineage Sorting) et seront donc observés en fréquences égales. Dévier de cet attendu, *i.e.* l'excès d'un des deux motifs, sera interprété comme une introgression entre deux lignées du groupe interne. Pour tester l'existence d'un excès d'un des motifs, on calcule la statistique-D (équation au centre de la Figure 1.8). Si un excès du motif ABBA est observé, alors D est positif, et l'interprétation du test est qu'une introgression a eu lieu entre P2 et P3 (Figure 1.8, en bas à gauche). Le test ne permet pas d'identifier le sens d'introgression.

Inversement, si le motif BABA est en excès, alors D est négatif et l'interprétation est qu'une introgression a eu lieu entre P3 et P1 (Figure 1.8, en bas à droite).

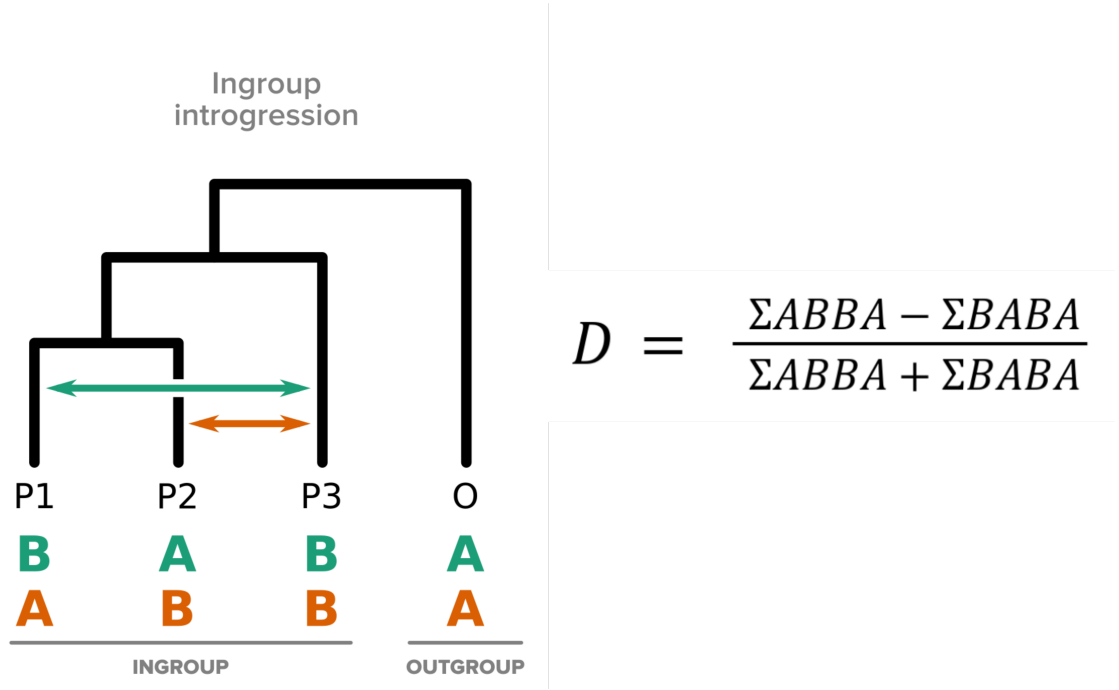


Figure *1.8* – *Description du test ABBA-BABA. Gauche, étant données quatre populations P1,P2,P3,O avec la topologie ((((P1),P2),P3),O) et les deux motifs de SNPs ABBA et BABA. Droite, on calcul la statistique-D pour tester si un événement d'introgression existe entre P1 ou P2, et P3. Une valeur de statistique-D positive est due à un excès du motif ABBA et est interprétée comme une introgression entre P3 et P2 (flèche orange). À l'inverse, une statistique-D négative due à un excès du motif BABA est interprétée comme une introgression entre P1 et P3 (flèche verte) (Tricou et al., 2021).*

La réconciliation d'arbres de gènes et d'arbre d'espèces Un des outils de détection des transferts horizontaux utilisé dans cette thèse est la "réconciliation" d'arbres. Le principe de cette méthode est de comparer l'histoire évolutive des espèces considérées, sous la forme d'un *arbre d'espèces*, aux histoires évolutives des gènes portés par ces espèces, les *arbres de gènes* (Figure 1.9). Si des différences sont observées entre un arbre d'espèces et un arbre de gènes, les méthodes de réconciliation vont tenter de proposer des événements évolutifs permettant d'expliquer ces différences, et donc de "réconcilier" ces deux arbres. Les événements en question sont des duplications de gènes, des pertes de gènes et des transferts horizontaux de gènes (Menet *et al.*, 2021).

Dans mes travaux de thèse la majorité des réconciliations ont été calculées avec l'outil de réconciliation ALE (Amalgamated Likelihood Estimation) (Szöllősi

et al., 2015). Ce programme présente plusieurs avantages comparé à d'autres outils disponibles et fonctionnant sur le même principe, comme ecceTera (Jacox *et al.*, 2016), Notung (Chen *et al.*, 2000), RANGER-DTL (Bansal *et al.*, 2018) ou AnGST (David and Alm, 2011). Premièrement, ALE a été développé par Gergely J. Szöllősi, un ancien membre du laboratoire où j'ai effectué ma thèse, et co-développé par certains membres encore présents dans mon équipe. Deuxièmement, là où la majorité des autres outils de réconciliation nécessitent de donner comme information les taux (ou coûts) des différents événements évolutifs considérés (Duplication, Transfert, Perte), taux qui nous sont très généralement inconnus, ALE est capable de les estimer par une approche probabiliste. Cela permet aussi d'avoir des scores pour les transferts qui sont inférés, ces scores représentant les probabilités *a posteriori* de chaque évènement.

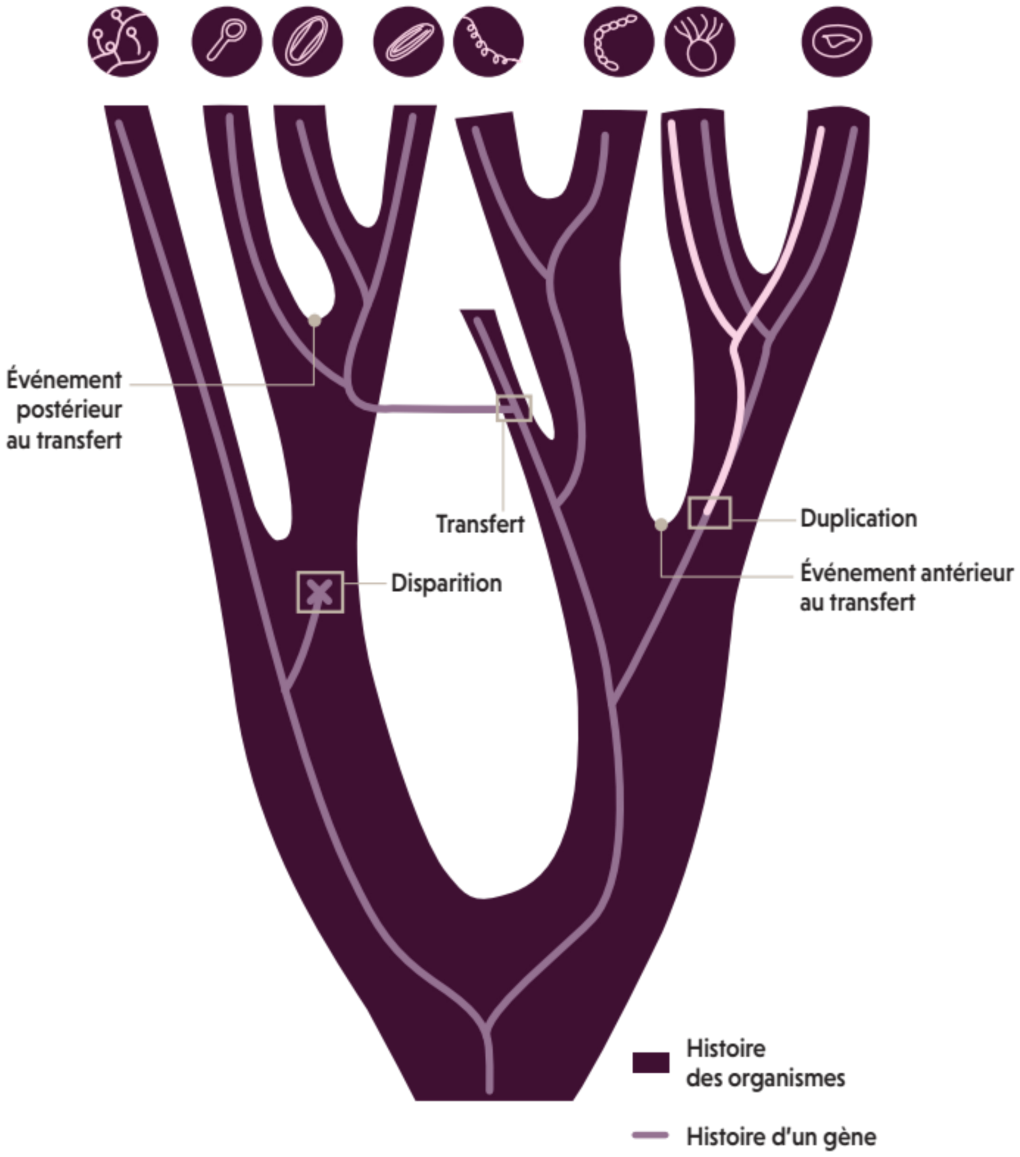


Figure *1.9* – *Dans cet exemple fictif de réconciliation phylogénétique, le tube épais est l'arbre d'espèces. Les lignes à l'intérieur représentent l'histoire évolutive d'un gène. Ce gène ne suit pas partout l'histoire des espèces échantillonnées. En particulier, il est porté un moment dans une espèce inconnue, puis se transfère (modifié de Tannier* et al. *(2019)).*

L'utilisation des tailles des branches dans les arbres phylogénétiques

Une autre façon de détecter les flux de gènes est de comparer les distances évolutives ou similarités entres séquences représentées dans des arbres par les longueurs des branches. Une conséquence des flux de gènes, autre que le changement de la topologie de l'arbre de gènes, est la diminution des longueurs des branches qui les séparent les espèces porteuse des gènes échangés. Les méthodes comme le D3 (Hahn and Hibbins, 2019), DIP (Forsythe *et al.*, 2020) ou

encore le DCT et BLT (Suvorov *et al.*, 2020) utilisent cette réduction des longueurs des branches dans les arbres de gènes qui sont transférés pour identifier les flux de gènes.

Par exemple, la méthode du D3 (une méthode dérivée du test ABBA-BABA présenté plus haut, Figure 1.10) compare la longueur des branches de toutes les familles de gènes sur 3 espèces ((A,B),C) pour détecter de l'introgression. Deux topologies alternatives, ne suivant pas celle de l'arbre d'espèces, peuvent survenir à cause de l'ILS ou de flux de gènes : ((B,C),A) ou ((A,C),B) (Figure 1.10 a et b). Une introgression entre deux lignées va réduire la distance qui les sépare dans les arbres des gènes transmis. Le D3 se calcule de la même façon que la statistique-D, mais au lieu de compter le nombre de motifs de SNP ici on compare la somme des distances séparant B et C à celles séparant A et C à travers l'ensemble des arbres de gènes. Plusieurs autres méthodes exploitent le signal des distances évolutives et des longueurs de branches pour détecter les flux de gènes (Novichkov *et al.*, 2004; Huang *et al.*, 2014; Adato *et al.*, 2015; Suvorov *et al.*, 2020; Susko *et al.*, 2021).

Cet attendu est aussi utilisé pour résoudre certaines questions évolutives en présence de flux de gènes. Dans (Fontaine *et al.*, 2015), les auteurs tentent de résoudre la phylogénie des espèces du complexe *Anopheles gambiae* face à de nombreux évènements d'introgression à travers tout le clade. Leur hypothèse est que la vraie topologie de l'arbre d'espèces correspond à celle qui a les longueurs de branches les plus longues et donc non affectée par les flux de gènes. Flux de gènes qui auraient autrement diminué la longueur des branches. Une autre façon d'utiliser les longueurs de branches est décrite par Pittis and Gabaldón (2016). Celles-ci sont utilisées pour ordonner chronologiquement les différents évènements de flux de gènes reçus par un organisme. Pour un emsemble de gènes qui a été acquis horizontalement par le même organisme, plus les branches séparant donneur et receveur sont longues plus l'évènement de flux de gènes est ancien, inversement plus elles sont courtes plus l'événement est récent. Je reviendrai sur ces trois méthodes plus en détails dans le chapitre 4.

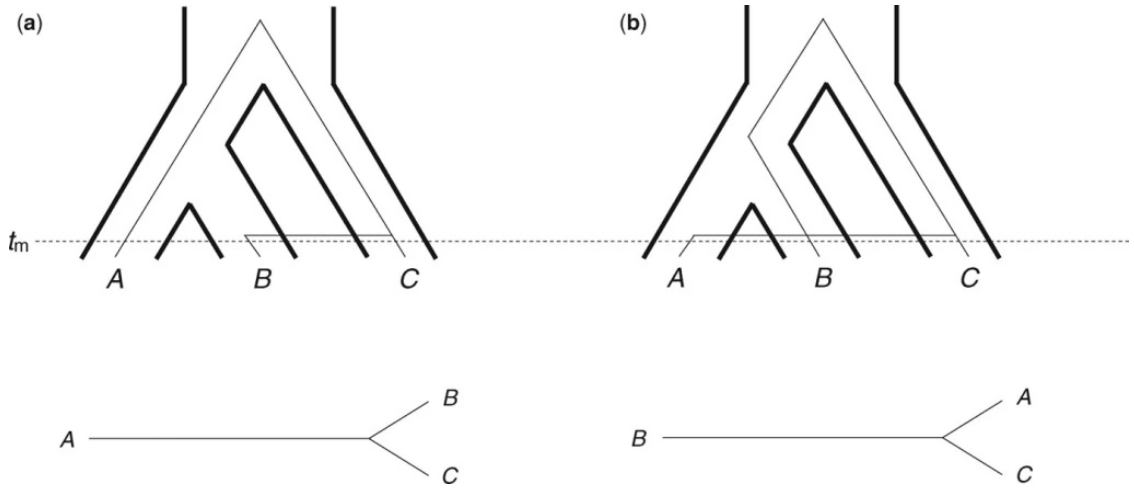


Figure *1.10* – *Utiliser la longueur des branches pour détecter les flux de gènes, le test D3. Ce test considère trois espèces A, B et C avec la topologie ((A,B),C). Deux topologies alternatives peuvent survenir à cause de l'ILS ou de flux de gènes (a et b). Dans un scénario d'évolution avec seulement de l'ILS et pas de flux de gènes, les valeurs attendues des sommes des distance séparant B à C et A à C sont exactement les mêmes. Si l'une des valeurs est supérieure à l'autre, alors il y a suspicion d'introgression. On calcule le D3 pour tester cette hypothèse. Un D3 positif est interprété comme une introgression entre B et C et un D3 négatif comme une introgression entre A et C. Tout comme le test ABBA-BABA, ce test ne permet pas d'évaluer le sens de l'introgression (modifié de Hahn and Hibbins (2019)).*

1.3 Pourquoi les lignées fantômes pourraient entraver notre capacité à détecter les flux de gènes

En règle générale, les méthodes phylogénétiques pour la détection des flux de gènes se basent sur l'hypothèse que, pour un gène transféré horizontalement, les versions portées par le donneur et le receveur vont être plus proches l'une de l'autre qu'attendu par leur lien de parenté dans l'arbre d'espèces. Ce rapprochement se reflète dans un arbre de gènes par le branchement de l'espèce porteuse du gène reçu à côté de l'espèce porteuse du gène donné (Figure 1.7 2b). Ce rapprochement est également visible via une diminution des longueurs des branches qui les séparent puisque la distance évolutive qui sépare les deux gènes est plus courte, leurs séquences étant d'autant plus similaires que le transfert est récent. Cependant, cette hypothèse ne peut être vérifiée que si l'on dispose d'une connaissance exhaustive de la biodiversité pour faire ce genre de comparaisons.

La justification est illustrée dans la figure 1.11. Prenons d'abord les cas A et B qui illustrent respectivement des transferts entre Sp3 et Sp2 ou entre Sp3 et Sp1 (Figure 1.11, gauche). Dans les deux cas on observe dans les arbres de gènes (Figure 1.11, centre) un rapprochement phylogénétique du donneur et du receveur. Dans le cas A, le gène porté pas Sp3 est le frère du gène porté par Sp2. Dans le cas B, le gène porté pas Sp3 est le frère du gène porté par Sp1. De plus on observe une diminution des longueurs de branches (illustré par les T1 et T2 dans les quatre arbres de la figure 1.11) qui séparent Sp3 et Sp2 (cas A) et Sp1 (cas B).

Maintenant prenons le cas C (Figure 1.11, droite), d'un gène transféré provenant d'une espèce fantôme X et reçu par Sp2. On observe que les gènes portés par Sp3 et Sp1 sont frères dans cet arbre de gènes, on retrouve alors la même topologie que l'arbre de gènes du cas B. Dans ce contexte la topologie observée n'est pas due à un rapprochement de Sp3 et Sp1 mais est due au fait que le gène porté par Sp2 est plus profond dans l'arbre, là où brancherait le gène porté par la lignée donneuse X. De plus, on va observer une augmentation des distances évolutives qui séparent Sp2 à Sp3 et Sp2 à Sp1, ce qui se reflète par de plus longues branches dans l'arbre de gènes. En effet, T1 et T2 dans l'arbre C sont plus grands que T1 et T2 dans l'arbre d'espèces à gauche.

Les espèces fantômes et les transferts provenant de celles-ci inversent donc les hypothèses à la base de la détection des flux de gènes. Dans le cas C, ce que l'on va observer, si l'on n'a pas la connaissance de la présence d'espèces fantômes, est la proximité phylogénétique des gènes portés par Sp1 et Sp3 comparée à la topologie de l'arbre d'espèces décrite par l'arbre de gauche. Sans *a priori* la décision la plus parcimonieuse pour ce scénario est d'inférer un transfert entre Sp1 et Sp3.

Il est raisonnable de penser que les flux de gènes ont toujours existé (Woese, 1998; Gogarten *et al.*, 2008). Ainsi, plus un flux de gènes est profond dans un arbre phylogénétique, et donc ancien, plus il y a de chance que la lignée donneuse soit désormais éteinte et n'ait laissé aucun descendant. Une observation similaire peut être faite si on considère les nombreuses espèces inconnues et non-échantillonnées dans les analyses phylogénétiques comme potentielles origines de transfert. Plus la quantité de lignées absentes d'une analyse est élevée, ce qui semble être la règle plutôt que l'exception, plus il y a de chances que la lignée donneuse soit elle aussi absente de l'arbre phylogénétique considéré tout comme ses descendants. On peut donc légitimement se poser la question de l'impact réel des lignées fantômes sur notre capacité à détecter des flux de gènes.

Dans cette thèse, je montre que la non prise en compte de lignées fantômes conduit plus souvent à la détection de transferts de gènes artefactuels qu'à l'identification de vrais évènements évolutifs. Je montre aussi qu'une conscience accrue de l'impact des lignées fantômes sur la détection des flux de gènes pourrait permettre de détecter des goupes fantômes et de les placer dans l'arbre du vivant.

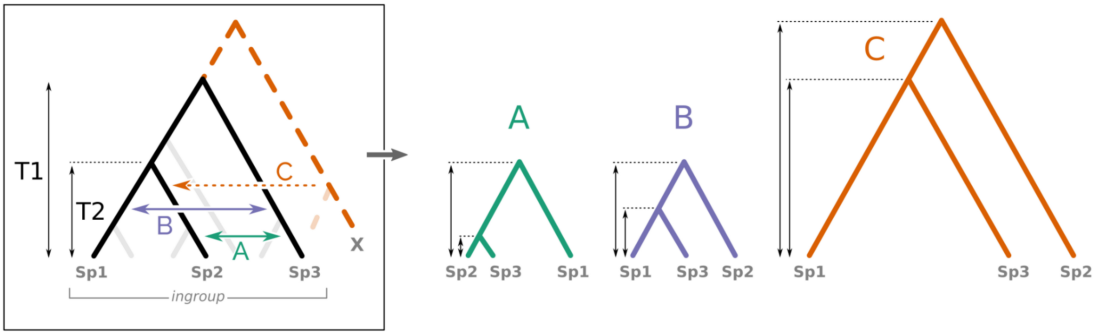


Figure *1.11* – *Effet des flux de gènes provenant d'espèces connues et fantômes sur la topologie et la longueur des branches des arbres phylogénétiques des régions correspondant à des fragments d'ADN transférés. Un transfert provenant d'une lignée fantôme (X et scénario C) produit une phylogénie avec des tailles de branches plus longues comparée à l'arbre d'espèces. Alors que des transferts entre espèces échantillonnées produisent le résultat opposé (scénarios A et B). De plus, les topologies des arbres de gènes résultant des scénarios C et B sont identiques alors que les événements de transferts impliquent un duo de donneur et receveur complètement différent (X et Sp2 dans le cas C et Sp3 et Sp1 dans le cas B).*

1.4 Motivations et Objectifs

Les flux de gènes, des phénomènes omniprésents dans l'arbre du vivant, constituent une force évolutive majeure qui façonne l'évolution d'une grande part de la biodiversité (Syvanen, 1985; Koonin *et al.*, 2001; Keeling and Palmer, 2008; Boto, 2010; Szöllősi *et al.*, 2015; Hall *et al.*, 2017). Ils ont longtemps été considérés comme une source d'erreur en phylogénétique et phylogénomique car ils compliquent la reconstruction des arbres phylogénétiques. Depuis le début des années 2000, cependant, plusieurs auteurs et autrices ont réalisé qu'ils étaient en réalité une riche mine d'information pour reconstruire l'histoire évolutive des organismes. Ils ont ainsi été utilisés pour raciner l'arbre du vivant (Abby *et al.*, 2012), pour améliorer la reconstruction d'arbres de gènes et d'espèces (Szöllősi *et al.*, 2015; Morel *et al.*, 2019) ou encore pour dater les arbres d'espèces (Davín *et al.*, 2018).

Un aspect largement négligé dans l'étude des flux de gènes, cependant, est l'impact que les espèces éteintes ou non encore connues, qui représentent la majeure part de la biodiversité, ont sur leur détection. le rôle que jouent ces lignées fantômes dans l'étude des flux de gènes a été largement négligé, par omission ou parce qu'elles ont été considérées, sans justification, comme une source de bruit de fond qui pouvait être ignoré.

Pourtant, et comme remarqué précédemment par Maddison (1997a) et Galtier and Daubin (2008), la surabondance des espèces fantômes comparée à la diversité connue et la très grande fréquence à laquelle se font les flux de gènes dans l'arbre du vivant impliquent que certains gènes, présents dans des espèces observables aujourd'hui, seraient apparus ou auraient évolué pendant un certain temps dans des lignées qui sont maintenant éteintes, ou nous sont simplement inconnues. Ce constat est à l'origine de mes travaux et de cette thèse. Il a aussi plusieurs implications : premièrement, que détecter les flux de gènes sans considérer la participation complète des espèces fantômes peut être source d'erreurs lorsqu'on cherche à les détecter. Deuxièmement, que les fragments d'ADN transférés horizontalement pourraient apporter des indices sur la nature et l'existence de la biodiversité fantôme.

L'absence de considération des espèces fantômes pour l'étude des flux de gènes et l'impact que cette biodiversité inconnue peut avoir sur les méthodes classiques utilisées en phylogénétique sont des problèmes que l'on commence tout juste à identifier et explorer. La maigre quantité de données génomiques et l'absence de méthodes de détection des flux de gènes suffisamment développées ne permettaient jusqu'alors pas d'évaluer ou de quantifier de tels effets. De plus, les méthodes d'inférence, comme celles qui permettent de détecter les flux de gènes, sont testées et validées sur des données simulées (Biller *et al.*, 2016). Or l'absence de méthode permettant la simulation *in silico* de génomes et de flux de gènes avec la participation intégrante des espèces fantômes explique que le problème qu'elles peuvent représenter n'ait que récemment été identifié. Le développement de l'outil de simulation Zombi auquel j'ai participé a contribué à mettre en avant ce problème. Ainsi, la communauté scientifique commence seulement à réaliser l'impact et l'importance de l'étude des lignées fantômes (Martin Kuhlwilm *et al.*, 2019; Ottenburghs, 2020).

Mon travail pendant cette thèse propose une exploration de l'impact des espèces fantômes sur la détection des flux de gènes. Je présente également une

méthode exploitant les traces d'ADN transféré par des espèces fantômes dans des espèces vivantes comme un signal permettant de détecter la diversité inconnue. Les résultats présentés sont rédigés sous la forme d'articles en anglais car pour au moins trois d'entre eux ils sont destinés à être - ou sont déjà - publiés dans des revues scientifiques. Le premier article (Davín *et al.*, 2020), présente *Zombi*, un outil bioinformatique que j'ai participé à développer et qui permet la simulation d'arbres d'espèces, de génomes et de séquences tout en prenant en compte les espèces fantômes (Chapitre 2). Le second manuscrit, soumis à *Systematic Biology* (2^{e} révision soumise), examine la robustesse de l'un des tests les plus courants de détection de l'introgression, le test ABBA-BABA (ou statistique-D), à la présence d'espèces fantômes (Chapitre 3). Le troisième manuscrit revient sur trois études qui se basent sur les tailles de branches dans les arbres phylogénétiques pour détecter de l'introgression, trouver de "bons" marqueurs pour la reconstruction phylogénétique et dater des évènements de flux de gènes. Les trois méthodes utilisées sont réévaluées à la lumière de l'existence probable d'espèces fantômes et les conclusions de ces études sont réévaluées (Chapitre 4). Le quatrième manuscrit, de format court, présente une démonstration théorique de la possibilité d'exploiter la détection des gènes transférés depuis des espèces fantômes pour détecter des clades encore inconnus (Chapitre 5).

Pour finir, au cours de ma thèse j'ai eu l'occasion de collaborer sur plusieurs autres projets. Le premier porte sur l'évolution de la taille des génomes des drosophiles, le contenu en éléments transposables et la taille efficace des organismes (Annexe 1). Le second s'inscrit dans le projet de thèse de Djivan Prentout et porte sur la description d'une paire de chromosomes sexuels homologues entre deux espèces, *Cannabis sativa* et *Humulus lupulus* (Annexe 2). Le troisième projet vise à détecter des échanges de traits linguistiques entre différentes langues en utilisant les outils de détection des flux de gènes de la biologie évolutive, comme le test ABBA-BABA (Annexe 3). Enfin, le quatrième projet s'inscrit dans le projet de thèse de Julien Joseph et porte sur l'évolution de la recombinaison chez les canidae (Annexe 4).

2

Simuler des espèces éteintes avec *Zombi*

Les outils de simulation de l'évolution *in silico* sont fréquemment utilisés pour l'évaluation de la robustesse des méthodes d'inférence statistique, fournissant des scénarios contrôlés dans lesquels des informations complètes sur les modèles et les processus d'évolution sont disponibles. Il est courant d'utiliser des simulations, en phylogénie ainsi que dans l'étude de la diversité biologique et ses origines, pour avoir une meilleure intuition et compréhension du vivant. Ces deux dernières décennies, un grand nombre de simulateurs d'arbres d'espèces et d'arbres de gènes ont été développés. Une grande partie de ces outils ont été développés pour répondre à un besoin particulier (Table 2.1) ou pour répondre à une question biologique précise (Mallo *et al.*, 2016). Une des limitations majeures de ces outils est qu'aucun ne considère les espèces fantômes, ils ne simulent que des lignées vivantes.

Nom	Arbre d'espèces	Arbres de gène	Séquences	Échantillonnage	Espèces éteintes	Régions intergénique	Arbres réconcilié	Fusion de gènes	ILS
Zombi	•	•	•	•	•	•	•		
ALF	•	•	•					•	
SimPhy	•		•	•			•		•
EvolSimulator	•		•						
GenPhyloData	•		•						
SaGePhy	•	•	•	•				•	

Table *2.1* – *Comparaison des fonctionnalités disponibles dans les principaux simulateurs d'évolution. Zombi (Davín* et al.*, 2020), ALF (Dalquen* et al.*, 2012), SimPhy (Mallo* et al.*, 2016), EvolSimulator (Beiko and Charlebois, 2007), GenPhyloData (Sjöstrand* et al.*, 2013) et SaGePhy (Kundu and Bansal, 2019). Il est reporté si les outils possèdent ou non les fonctionnalités suivantes : simuler des arbres d'espèces, des arbres de gènes (niveau du génome, ce qui signifie qu'il considère la structure du génome, et les relations de contiguïté physique des gènes dans un génome), des séquences (niveau des séquences ADN et ou ARN), la présence de lignées éteintes (Espèces éteintes), la possibilité d'échantillonner des espèces intégrées dans le simulateur et d'élaguer les arbres génétiques selon les espèces échantillonnées (Échantillonnage), la simulation de régions intergéniques, la sortie d'arbres réconciliés, simule la fusion et la fission des gènes (Fusion de gènes) ou simule du trie de lignées incomplet (ILS).*

Dans ce chapitre est présenté *Zombi* (Davín *et al.*, 2020), un simulateur d'arbres d'espèces, de génomes et de séquences qui prend explicitement en compte l'évolution des génomes dans les espèces fantômes, éteintes ou non-échantillonnées. Les simulateurs comme SimPhy (Mallo *et al.*, 2016) ou SaGePhy (Kundu and Bansal, 2019) font l'hypothèse implicite que lors d'un transfert de gènes, le donneur du transfert laisse toujours un descendant survivant parmi les espèces échantillonnées, ils ont une connaissance implicite du futur. *Zombi*, au contraire, ne prend pas en compte le futur, ou l'information de quelles lignées vont disparaître ou apparaître, pour déterminer qui sont les donneurs et receveurs d'un événement de transfert de gènes. Faire cette hypothèse d'absence d'espèces cachées peut pourtant grandement entraver la simulation de scénarios d'évolution plus proche de l'état de la biodiversité actuelle. Le développement de méthodes pour la simulation des espèces fantômes est un atout pour l'étude de cette diversité cachée et de son impact sur la détection des flux de gènes et d'autres processus évolutifs.

Bioinformatics, 36(4), 2020, 1286–1288
doi: 10.1093/bioinformatics/btz710
Advance Access Publication Date: 30 September 2019
Applications Note

Phylogenetics

Zombi: a phylogenetic simulator of trees, genomes and sequences that accounts for dead linages

Adrián A. Davín[1,2,*], Théo Tricou[3], Eric Tannier[3,4], Damien M. de Vienne[3,†] and Gergely J. Szöllősi[1,2,5,†]

[1]MTA-ELTE Lendület Evolutionary Genomics Research Group, Budapest, Hungary, [2]Department of Biological Physics, Eötvös Loránd, Budapest, Hungary, [3]Univ Lyon, Université Lyon 1, CNRS, Laboratoire de Biométrie et Biologie Évolutive UMR5558, Villeurbanne F-69622, France, [4]INRIA Grenoble Rhône-Alpes, Montbonnot-Saint-Martin F-38334, France and [5]Evolutionary Systems Research Group, Centre for Ecological Research, Hungarian Academy of Sciences, Tihany H-8237, Hungary

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

Received on April 11, 2019; revised on September 9, 2019; editorial decision on September 11, 2019; accepted on September 26, 2019

Abstract

Summary: Here we present Zombi, a tool to simulate the evolution of species, genomes and sequences in silico, that considers for the first time the evolution of genomes in extinct lineages. It also incorporates various features that have not to date been combined in a single simulator, such as the possibility of generating species trees with a pre-defined variation of speciation and extinction rates through time, simulating explicitly intergenic sequences of variable length and outputting gene tree—species tree reconciliations.
Availability and implementation: Source code and manual are freely available in https://github.com/AADavin/ZOMBI/.
Contact: aaredav@gmail.com
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Reconstructing the pattern of horizontal gene transfers between species can help us date the origin of different taxa (Davín *et al.*, 2018; Wolfe and Fournier, 2018), understand the spread of genes of clinical importance (Lerminiaux and Cameron, 2019) and resolve difficult phylogenetic questions, such as inferring the rooting point of prokaryotic trees (Abby *et al.*, 2012; Szöllősi *et al.*, 2012; Williams *et al.*, 2017) or the evolutionary position of certain lineages of unclear origin (Boussau *et al.*, 2008). In the last decades, a large number of simulators have been developed to model a wide range of evolutionary scenarios (Beiko and Charlebois, 2007; Carvajal-Rodríguez, 2008; Dalquen *et al.*, 2012; Kundu and Bansal, 2019; Mallo *et al.*, 2016; Sjöstrand *et al.*, 2013) but none so far have considered the existence of extinct lineages and the horizontal transmission of genes (by lateral gene transfers) involving species that are not represented in the phylogeny (Fournier *et al.*, 2009; Szöllősi *et al.*, 2013; Zhaxybayeva and Peter Gogarten, 2004). Zombi simulates explicitly the genome evolution taking place in these extinct lineages, which is expected to have an impact in extant lineages by means of Lateral Gene Transfers (Szöllősi *et al.*, 2013). By not considering extinct lineages, other simulators make the implicit assumption that the transfer donor always leaves a surviving descendant among sampled species, while we know that this is most often not true (Szöllősi *et al.*, 2013). Making this assumption may potentially hamper our ability to simulate realistic scenarios of evolution. In addition to considering evolution along extinct lineages, Zombi includes several features hitherto not found together in any other simulator (Supplementary Table S1).

2 Basic features of Zombi

Zombi is a multilevel simulator, where a species tree is first simulated, then genomes evolve along the branches of this species tree, and finally, sequences are generated for each genome. These three steps, depicted in Figure 1 and detailed hereafter, are controlled by three main 'modes', named T, G and S, for species Tree, Genome and Sequence, respectively.

The T mode simulates a species tree under the birth-death model (Kendall, 1948), using the Gillespie algorithm (Gillespie, 1977), which is the standard method for simulating arbitrarily complex continuous time Markov processes (Supplementary Fig. S1). While more efficient and accurate methods exist to simulate the reconstructed tree (Hartmann *et al.*, 2010), taking into consideration unrepresented (extinct and unsampled) species requires simulating the complete species tree, which includes all extinct and unsampled branches of the phylogeny (Szöllősi *et al.*, 2013). This tree is subsequently pruned to obtain the reconstructed tree, by removing all the

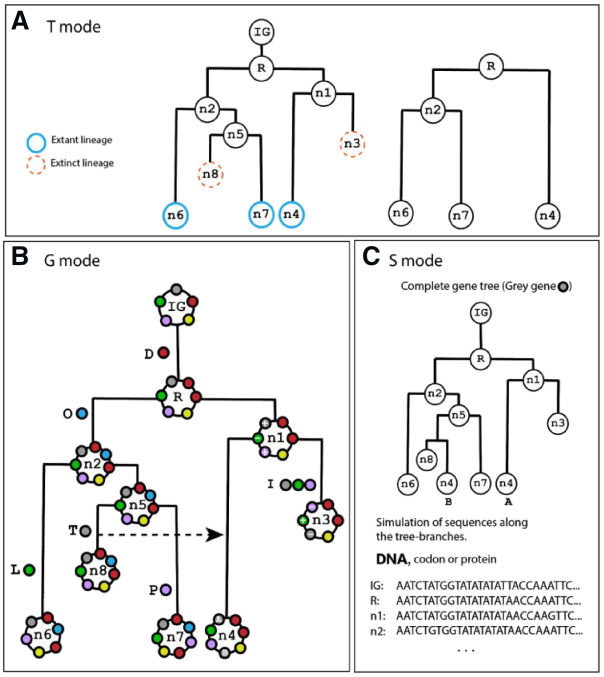


Fig. 1. Overview of the three steps of the Zombi simulator. (**A**) In T mode, Zombi simulates a species tree using a birth-death process and outputs the pruned version of it by removing extinct lineages. In this example, lineages n3 and n8 go extinct before the simulation ends. (**B**) in **G** mode, a circular genome evolves within the branches of the complete species tree obtained with the **T** mode by Duplications (D), Originations (O), Inversions (I), Transpositions (P), Losses (L) and Transfers (T) of genes. The simulation starts with the initial genome (IG) containing a number of genes determined by the user (5 in this example, represented by the coloured circles). Each gene has an orientation (+ or -) that is determined randomly and represents the direction of the gene in the coding strand. Several events affecting different genes and their impact on the genome structure are indicated next to the branches where they occur. The inversion events not only modify the positions of the genes but also change their orientation. (**C**) In **S** mode, Zombi can be used to simulate codon, nucleotides and amino acids along the branches of the gene family trees. Here, the gene tree of the grey coloured gene family from B has been depicted

lineages that did not survive until the end of the simulation (Fig. 1A).

The **G** mode simulates the evolution of genomes within the branches of the complete species tree (Fig. 1B) using also the Gillespie algorithm (Supplementary Fig. S2) to account for six possible genome-level events: duplications, losses, inversions, transpositions, transfers and originations. Each of the first five events is characterized by two parameters: the first one is the effective rate, that controls the frequency and fixation probability; the second one controls the extension, i.e. the number of contiguous genes simultaneously affected by the event. Originations of new genes occurs one by one and therefore only a single effective rate parameter is needed. When a Transfer event occurs, the recipient lineage is randomly chosen from all the lineages alive at that time. The user can make the frequence of transfers to be higher between closely related lineages (Ochman *et al.*, 2000) (Supplementary Fig. S3). Once the simulation reaches the end, Zombi outputs a list containing each event that has occurred in the simulation for every gene family (all genes that share a common origin). Besides, the gene trees of each family are reconstructed by combining both species-level events (Speciations and Extinctions) and genome-level events (Duplications, Transfers and Losses). Inversions and transpositions do not modify the topology of the tree but add an extra layer of complexity by changing the neighborhood of genes, which is especially relevant when genome-level events affect more than one gene at a time (Supplementary Fig. S4). The gene family trees are also pruned to present the user the trees that can be expected to be recovered from most real-data analyses, removing all extinct lineages and gene branches that do not arrive until the present time. The **S** mode, finally, simulates gene sequences (at either the codon, nucleotide or protein level) along the gene family trees (Fig. 1C). The user can modify the scaling of the tree to better control the number of substitutions that take place per unit of time, and thus simulate fast or slow-evolving genes.

3 Advanced features

In addition to the basic features presented above, 'advanced' modes of Zombi (listed in Supplementary Table S2) can be used to obtain richer and more realistic evolutionary scenarios. For example, it is possible to use a species tree input by the user, to generate species trees with variable extinction and speciation rates, or to control the number of living lineages at each unit of time (Supplementary Fig. S5). At the genome level, Zombi can simulate genomes using branch-specific rates (Gu mode, allowing the user to simulate very specific scenarios such as one in which a certain lineage experiences a massive loss of genes), gene-family specific rates (Gm mode, which makes easier the process of using rates estimated from real datasets) and genomes accounting for intergenic regions (Gf mode) of variable length [drawn from a flat Dirichlet distribution (Biller *et al.*, 2016)]. At the sequence level the user can fine-tune the substitution rates to make them branch specific. Zombi provides the user with a clear and detailed output of the complete evolutionary process simulated, including the reconciled gene trees with the species tree in the RecPhyloXML reconciliation standard (Duchemin *et al.*, 2018).

4 Performance and validation

Simulations with Zombi are fast: with a starting genome of 500 genes and a species tree of 2000 taxa (extinct + extant), it takes around 1 min on a 3.4Ghz laptop to simulate all the genomes (Supplementary Fig. S6).

We validated that the distribution of waiting times between successive events was following an exponential distribution (Supplementary Figs S7 and S8), that the distribution of intergene sizes at equilibrium was following a flat Dirichlet distribution, as expected from Biller *et al.* (2016) (Supplementary Fig. S9), that the number of events and their extension occur with a frequency according to their respective rates (Supplementary Fig. S10) and that the gene family size distribution followed a power-law when duplication rates are higher than loss rates and stretched-exponential in the opposite case (Reed and Hughes, 2003; Szöllosi and Daubin, 2011) (Supplementary Fig. S11). We also checked by hand the validity of many simple scenarios to detect possible inconsistencies in the algorithm.

5 Implementation

Zombi is implemented in Python 3.6. It relies on the ETE 3 toolkit (Huerta-Cepas *et al.*, 2016) and the Pyvolve package (Spielman and Wilke, 2015). It is freely available at https://github.com/AADavin/ZOMBI along with detailed documentation and two tutorials in a wiki page.

Acknowledgements

We thank Vincent Daubin, Wandrille Duchemin, Nicolas Lartillot and Thibault Latrille for insightful discussions during the preparation of this manuscript.

Funding

A.A.D. and G.J.Sz. received funding from the European Research Council under the European Union's Horizon 2020 research and innovation programme under grant agreement no. 714774., in addition, G.J.Sz. was supported by the grant GINOP-2.3.2.-15-2016-00057. T.T and D.M.d.V received funding from grant ANR-18-CE02-0007-01 ('STHORIZ').

Conflict of Interest: none declared.

References

Abby,S.S. *et al*. (2012) Lateral gene transfer as a support for the tree of life. *Proc. Natl. Acad. Sci. USA*, **109**, 4962.

Beiko,R.G. and Charlebois,R.L. (2007) A simulation test bed for hypotheses of genome evolution. *Bioinformatics*, **23**, 825–831.

Biller,P. *et al*. (2016) Breaking good: accounting for fragility of genomic regions in rearrangement distance estimation. *Genome Biol. Evol*., **8**, 1427–1439.

Boussau,B. *et al*. (2008) Accounting for horizontal gene transfers explains conflicting hypotheses regarding the position of aquificales in the phylogeny of bacteria. *BMC Evol. Biol*., **8**, 272.

Carvajal-Rodríguez,A. (2008) Simulation of genomes: a review. *Curr. Genomics*, **9**, 155–159.

Dalquen,D.A. *et al*. (2012) ALF—a simulation framework for genome evolution. *Mol. Biol. Evol*., **29**, 1115–1123.

Davín,A.A. *et al*. (2018) Gene transfers can date the tree of life. *Nat. Ecol. Evol*., **2**, 904–909.

Duchemin,W. *et al*. (2018) RecPhyloXML – a format for reconciled gene trees. *Bioinformatics*, **34**, 3646.

Fournier,G.P. *et al*. (2009) Horizontal gene transfer from extinct and extant lineages: biological innovation and the coral of life. *Philos. Trans. R. Soc. Lond. Ser. B Biol. Sci*., **364**, 2229–2239.

Gillespie,D.T. (1977) Exact stochastic simulation of coupled chemical reactions. *J. Phys. Chem*., **81**, 2340–2361.

Hartmann,K. *et al*. (2010) Sampling trees from evolutionary models. *Syst. Biol*., **59**, 465.

Huerta-Cepas,J. *et al*. (2016) ETE 3: reconstruction, analysis, and visualization of phylogenomic data. *Mol. Biol. Evol*., **33**, 1635–1638.

Kendall,D.G. (1948) On the generalized 'Birth-and-Death' process. *Ann. Math. Stat*., **19**, 1–15.

Kundu,S. and Bansal,M.S. (2019) SaGePhy: an improved phylogenetic simulation framework for gene and subgene evolution. *Bioinformatics*, **35**, 3496–3498.

Lerminiaux,N.A. and Cameron,A.D.S. (2019) Horizontal transfer of antibiotic resistance genes in clinical environments. *Can. J. Microbiol*., **65**, 34–44.

Mallo,D. *et al*. (2016) SimPhy: phylogenomic simulation of gene, locus, and species trees. *Syst. Biol*., **65**, 334–344.

Ochman,H. *et al*. (2000) Lateral gene transfer and the nature of bacterial innovation. *Nature*, **405**, 299–304.

Reed,W.J. and Hughes,B.D. (2003) Power-law distribution from exponential processes: an explanation for the occurrence of long-tailed distributions in biology and elsewhere. *Sci. Math. Jpn*. http://www.math.uvic.ca/faculty/reed/JAMS_sub.pdf.

Sjöstrand,J. *et al*. (2013) GenPhyloData: realistic simulation of gene family evolution. *BMC Bioinformatics*, **14**, 209.

Spielman,S.J. and Wilke,C.O. (2015) Pyvolve: a flexible python module for simulating sequences along phylogenies. *PLoS One*, **10**, e0139047.

Szöllősi,G.J. and Daubin,V. (2011) The pattern and process of gene family evolution. *arXiv preprint arXiv:1102.2331*.

Szöllősi,G.J. *et al*. (2012) Phylogenetic modeling of lateral gene transfer reconstructs the pattern and relative timing of speciations. *Proc. Natl. Acad. Sci. USA*, **109**, 17513–17518.

Szöllősi,G.J. *et al*. (2013) Lateral gene transfer from the dead. *Syst. Biol*., **62**, 386–397.

Williams,T.A. *et al*. (2017) Integrative modeling of gene and genome evolution roots the archaeal tree of life. *Proc. Natl. Acad. Sci. USA*, **114**, E4602.

Wolfe,J.M. and Fournier,G.P. (2018) Horizontal gene transfer constrains the timing of methanogen evolution. *Nat. Ecol. Evol*., **2**, 897–903.

Zhaxybayeva,O. and Peter Gogarten,J. (2004) Cladogenesis, coalescence and the evolution of the three domains of life. *Trends Genet*., **20**, 182–187.

3

Les lignées fantômes influencent l'interprétation du test d'introgression ABBA-BABA

Les phénomènes d'introgression sont courants chez les eucaryotes. Par exemple de nombreux événements de flux de gènes ont été identifiés entre les populations humaines archaïques et d'humains modernes (Dannemann and Racimo, 2018), dans le groupe des *Lepidoptera* (genre *Heliconius*, les papillons) (Edelman *et al.*, 2019) ou encore chez des espèces domestiquées (Wu *et al.*, 2018; Rouard *et al.*, 2018), de plantes (Burgarella *et al.*, 2019) comme d'animaux (Wu *et al.*, 2018). Ces évènements peuvent parfois être liés à des processus macro-évolutifs, d'adaptation évolutive ou de radiation évolutive (Ottenburghs, 2020).

Un grand nombre d'études ont utilisé le test ABBA-BABA et la statistique-D (Meyer *et al.*, 2012; Eaton and Ree, 2013; Martin *et al.*, 2013; Schumer *et al.*, 2016; Zhang *et al.*, 2018) pour identifier des introgressions entre des lignées existantes, présentes dans les arbres phylogénétiques considérés. L'interprétation de ce test se fait généralement sans prendre en compte les espèces fantômes, qu'elles soient éteintes, inconnues ou non-échantillonnées. Pourtant le test s'applique seulement sur un sous-échantillon de 4 espèces, donc les espèces absentes de l'arbre à 4 lignées utilisées pour le test sont légion en comparaison. La possibilité d'introgression impliquant une espèce ou une population non visible

par le test est rarement considérée alors même que leur absence peut affecter l'interprétation du test et les conclusions qui en sont tirées.

Ghost lineages highly influence the interpretation of introgression tests

Théo Tricou[1,*], Eric Tannier[1,2] and Damien M. de Vienne[1,*]

[1] Univ Lyon, Université Lyon 1, CNRS, Laboratoire de Biométrie et Biologie Évolutive UMR5558, F-69622 Villeurbanne, France

[2] INRIA Grenoble Rhône-Alpes, F-38334, France

***Corresponding authors: E-mails:** *theo.tricou@univ-lyon1.fr*, *damien.de-vienne@univ-lyon1.fr*

Abstract

Most species are extinct; those that are not are often unknown. Sequenced and sampled species are often a minority of known ones. Past evolutionary events involving horizontal gene flow, such as horizontal gene transfer, hybridization, introgression and admixture, are therefore likely to involve "ghosts", *i.e.* extinct, unknown or unsampled lineages. The existence of these ghost lineages is widely acknowledged, but their possible impact on the detection of gene flow and on the identification of the species involved is largely overlooked. It is generally considered as a possible source of error that, with reasonable approximation, can be ignored. We explore the possible influence of absent species on an evolutionary study by quantifying the effect of ghost lineages on introgression as detected by the popular D-statistic method. We show from simulated data that under certain frequently encountered conditions, the donors and recipients of horizontal gene flow can be wrongly identified if ghost lineages are not taken into account. In particular, having a distant outgroup, which is usually recommended, leads to an increase in the error probability and to false interpretations in most cases. We conclude that introgression from ghost lineages should be systematically considered as an alternative possible, even probable, scenario.

Key words: ghost lineage, D-statistic, ABBA-BABA, introgression, gene flow, simulation

1. Introduction

Evolutionary studies are always restricted to a subset of species, populations or individuals. This is by choice, because only a fraction of the data is relevant to the question being addressed, and by necessity, because the approaches used have methodological and technical limitations. Another reason is that most lineages are simply unknown. More than 99.9% of all species that have ever lived are now extinct (Raup 1991) and only a small fraction of extant species have been described. The number of extant eukaryote species that are still uncatalogued is almost an order of magnitude higher than the number of those reported (~1.3 million species have been catalogued, Mora et al. 2011), and is many orders of magnitudes higher if we consider Bacteria and Archaea diversity (Locey and Lennon 2016).

Taking these extinct, unknown or unsampled "ghost" lineages into account is particularly important when studying introgression, *i.e.* the integration of genetic material from one lineage to another *via* hybridization and subsequent backcrossing. This mode of gene flow across species boundaries appears to be common in the Eukaryotic domain and has been shown to be adaptive in some cases (see for example Hedrick 2013 for a review). Introgression has been reported in such diverse lineages as humans (Green et al. 2010; Meyer et al. 2012), boars (Liu et al. 2019), butterflies (Martin et al. 2013; Smith and Kronforst 2013; Massardo et al. 2020), fishes (Schumer et al. 2016; Meier et al. 2017), plants (Eaton and Ree 2013; Zhang et al. 2019) and fungi (Zhang et al. 2018; Keuler et al. 2020), to name but a few.

Since ghost lineages are probably massively present around any phylogeny of extant species, many gene flow events that are detectable now are likely to have involved a ghost lineage. This has been repeatedly acknowledged (Maddison 1997; Galtier and Daubin 2008; Green et al. 2010; Eaton and Ree 2013; Szöllősi et al. 2013, 2015), especially in studies of introgression between populations, but it was considered either a source of noise (Pease and Hahn 2015), or a problem that could be resolved by adding new species as they become available, or by combining the results of multiple detection tests (Eaton et al. 2015; Kumar et al. 2017; Barlow et al. 2018). Recently, Hibbins and Hahn (2021) advised bearing ghost lineages "in mind" when investigating gene flow but, as far as we know, the real impact of ghost lineages on the ability of different methods to detect gene flow and correctly identify involved lineages has not been properly evaluated and quantified.

Over the past few years, the ever-growing number of sequenced genomes and the development of new methods have improved the detection of introgression. One of the most widely used methods for inferring introgression is the D-statistic (or Patterson's D), also known as the ABBA-BABA test (Kulathinal et al. 2009; Green et al. 2010; Durand et al. 2011; Patterson et al. 2012). There are many reasons for its success. The D-statistic is easy to understand and implement, quick to compute and easy to interpret. This method is based on phylogenetic discordance and can discriminate incongruence caused by Incomplete Lineage Sorting (ILS) from incongruence caused by gene flow (Kulathinal et al. 2009; Green et al. 2010; Durand et al. 2011; Patterson et al. 2012). The ABBA-BABA test considers four taxa: three ingroup taxa and one outgroup, with a ladder-like phylogenetic relationship (Fig. 1). The test relies on counts of the number of sites that support a discordant topology. Two biallelic SNP patterns are considered, ABBA and BABA, depending on which allele (A: ancestral, B: derived) is present in each taxon. The D-statistic is computed using the classic formula from Durand et al. (2011):

$$D = \frac{\Sigma ABBA - \Sigma BABA}{\Sigma ABBA + \Sigma BABA} \quad (1)$$

The null hypothesis states that under a scenario with no gene flow, both ABBA and BABA patterns can be attributed to ILS and thus should be observed in equal numbers. Significant deviation from this expectation, resulting in a D-statistic significantly different from zero, is usually interpreted as introgression between two of the three lineages forming the ingroup (Fig. 1, left panel). The outgroup should be distant enough from the ingroup such that it is not involved in an introgression with any of the ingroup lineages (Green et al. 2010; Osborne et al. 2016; Irwin et al. 2018).

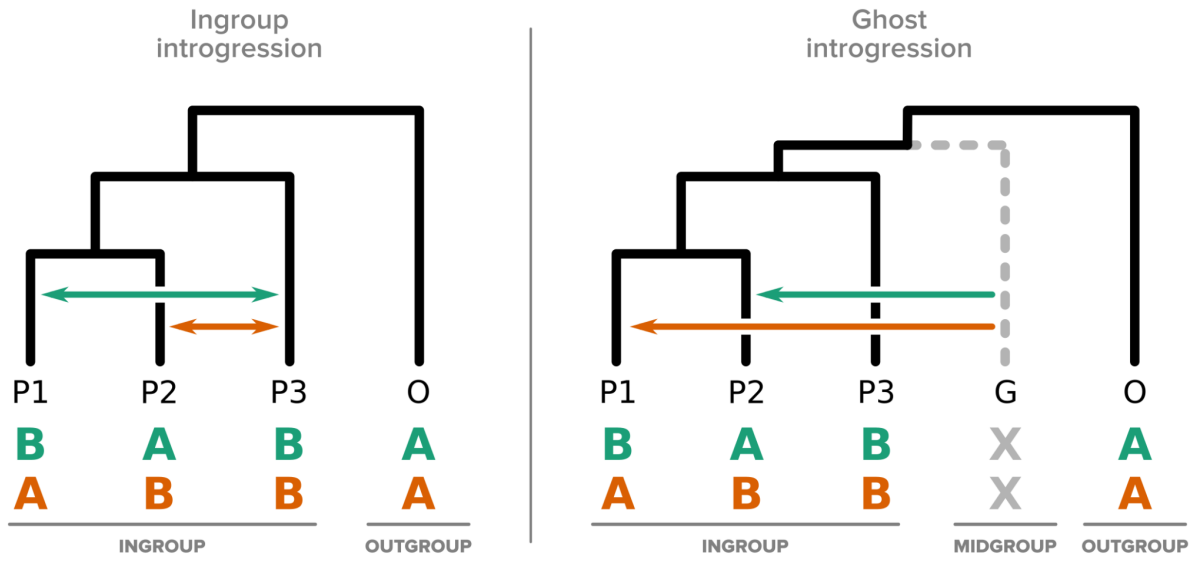


Figure 1. Introgression events that can result in a significant excess of ABBA or BABA patterns according to the D-statistic. The usual interpretation of this excess is the hypothesis of "ingroup" introgression (left panel). However "ghost" (or "midgroup") introgression (right panel) from ghost lineages (G) can produce similar patterns.

Undersampling is known to be one of the factors that can possibly confound the D-statistic (Martin et al. 2015; Zheng and Janke 2018), and affect the detection of introgression. This is because using a subset logically leads to an underestimation of the true frequency of introgression and thus inflates the role of ILS (Maddison and Knowles 2006). It has been clearly stated that the donor genome could easily be misidentified because introgression from a sampled lineage (*e.g.* P3) or from one sister ghost lineage to the same recipient lineage would produce the same signal and result in indistinguishable D-statistic results (Eaton and Ree 2013; Eaton et al. 2015; Pease and Hahn 2015; Zheng and Janke 2018). Another stronger impact of ghost lineages, however, was foreseen early on in the history of the test (by Durand et al. (2011), in their first description of the test), but has been largely overlooked afterwards: introgression from a ghost lineage between the ingroup and the outgroup (the "midgroup", see Fig. 1) could lead to the wrong identification of both the donor ***and*** the recipient genomes (Fig. 1, right panel). Under this scenario, none of the species thought to be involved in the introgression event are correctly identified. This possible source of error in the interpretation of the D-statistic has often been acknowledged (Durand et al. 2011; Ottenburghs et al. 2017;

Zheng and Janke 2018; Hibbins and Hahn 2021) but surprisingly, it does not seem to have changed the way the test is commonly interpreted, perhaps because it is thought that the impact of this possibility is low, even though it has not been formally quantified. This is the goal of this study.

We begin by an illustration of the possibility of misinterpreting the ABBA-BABA test when some species are unknown or not included using a previously published bear phylogeny, recurrently used later on to estimate parameters on realistic situations. We then quantify the effect of ghost lineages on the misidentification of the donor and the recipient lineage using simulations. We explore the impact of outgroup choice, number of unsampled species and genetic divergence between introgressed taxa on the probability of misinterpreting introgression events.

We show that under the realistic assumption that there are many ghost lineages branches in the tree, and assuming a simple demographic history of the populations considered, most significant D-statistics are attributable to ghost lineages. This suggests that most of the lineages involved (donors and recipients) are incorrectly identified by the usual interpretation of D-statistics. The error rate increases with the distance between ingroup and outgroup, even though the outgroup is usually chosen so that its distance from the ingroup is sufficient to avoid any introgression between the two (Green et al. 2010; Osborne et al. 2016; Irwin et al. 2018). This observation, that a close outgroup as well as a distant outgroup is a source of interpretation error, hampers the delimitation of a safe zone for the interpretation of the D-statistic.

These results call for a new way of interpreting D-statistics, and more generally call into question established methods of introgression detection. Our results illustrate the recent statement by Ottenburghs (2020) that “the presence of ghost introgression has important consequences for the study of evolutionary processes”, and provide a demonstration of this importance.

2 Materials and Methods

2.1 Bear genomic dataset

We use the dataset from Barlow et al. (2018) to illustrate the possibility of misinterpreting significant results of an ABBA-BABA test. This dataset has the advantage of being easily

available and to support, according to the authors, a simple introgression scenario, *i.e.* a documented introgression between polar bears and brown bears from the ABC Islands in Alaska. We downloaded the genome sequences (Barlow et al. 2019) of three brown bears (*Ursus arctos*) from Alaska (id: Adm1), Russia (id: 235) and Slovenia (id: 191Y), one polar bear (*U. maritimus*; id: NB) and one American black bear (*U. americanus*; id: Uamericanus), all aligned against the panda reference genome (Li et al. 2010). Their relationship is ((((Alaska = P1, Russia = P2), Slovenia = P3), Polar bear = P4), Black bear = O). Using scripts available from the GitHub repository of Barlow et al. (2018) https://github.com/jacahill/Admixture, we computed the D-statistic for two quartets: (((Alaska,Russia),Polar bear),Black bear) and (((Alaska,Russia),Slovenia),Black bear). We used the script for all sites (not transversions only like in Barlow et al, 2018) as we do not have any archaic species in the dataset. The weighted block jackknife from the script was used to compute the Z-score in non-overlapping 1 Mb windows with a script available on the GitHub repository mentioned above. We considered the result to be significant if it was more than three standard deviations from zero (Z > 3 or Z < -3), as *per* Green et al. (2010).

2.2 Species tree and gene tree simulation

Species trees were simulated using the birth-death simulator implemented in the R function *rphylo* from the *ape* package (Paradis and Schliep 2019). Speciation rate was fixed at 1, extinction rate at 0.9 and the simulation was stopped when N extant lineages, varying in {20, 40, 60, 80, 100}, were present in the species tree (step 1 in Figure 2). Then 20 taxa were uniformly sampled from the N taxa. An introgression event was chosen in the species tree (including unsampled lineages), with a donor and a recipient. The donor branch was selected among the branches of the species tree with a probability proportional to its length, and the time of introgression uniformly at random on this branch. The recipient was randomly sampled among the lineages present at the time of the introgression, with a probability that decreases exponentially with the phylogenetic distance from the donor (step 2 in Figure 2):

$$p_i = e^{-\alpha D_i} \quad (2)$$

Where D_i is the distance from the *i*-th recipient, normalized by the distances from all possible recipients, and α is a parameter in {0, 1, 10, 100, 1000}. With $\alpha = 0$, introgressions occur

between any contemporaneous branches on a tree with equal probability. When α increases, introgressions are more likely to occur between closely related taxa. If the recipient is a branch with no extant offspring, then the introgression cannot be detected, so only introgressions such that the recipient has descendants among the 20 remaining species were kept. After setting the donor and recipient, a gene tree was generated with subtree pruning and regrafting (SPR) (Bordewich and Semple 2005), simulating the introgression event (step 3 in Figure 2). All unsampled lineages are then pruned from species and gene trees.

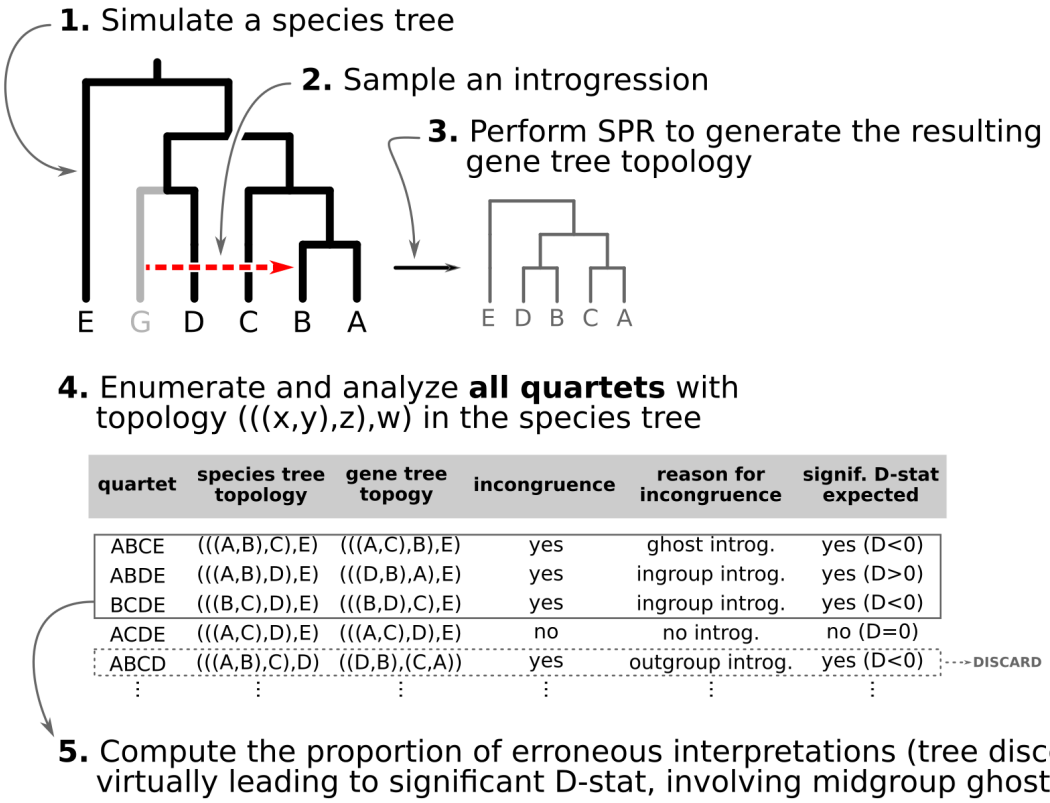


Figure 2. Species tree/Gene tree simulation: (1) a species tree is generated under a birth death model and 20 taxa are sampled from it; (2) an introgression event is picked from a random donor and recipient; (3) an introgressed gene tree is constructed from the species tree by SPR; (4) for each quartet with a ladder-like topology (((x,y),z),w) in the species tree, species tree and gene tree topologies are compared to determine if there is an incongruence caused by the introgression; (5) the proportion of erroneous interpretation of the D-statistic across the species tree is computed by the sum of all introgressions with a midgroup ghost donor over all introgressions detected, outgroup introgressions excluded.

2.3 Comparison of the species tree and the gene tree as proxy for the D-statistic

For each gene tree/species tree pair, we counted all species quartets with a ladder-like topology (((P1,P2),P3),P4) in the species tree and (((P1,P3),P2),P4) or (((P2,P3),P1),P4) in the gene tree. These configurations were interpreted as yielding significant D-statistics. This avoids the computational burden of simulating sequences for a high number (250 000, see Section 2.5) of cases and gives reasonably equivalent results (Supplementary Material Section 1).

We then counted the number of situations where the result was correctly interpreted (the simulated introgression is between extant lineages in this quartet) or misinterpreted (the simulated introgression is from ghost lineage) (step 5 in Figure 2).

2.4 Measuring the distance to the outgroup

For each species quartet we computed the distance between outgroup and ingroup using the ratio $R=t1/t2$, where $t1$ is the distance (sum of branch lengths) between the most recent common ancestor of the ingroup and the most recent common ancestor of all four taxa (see $t1$ in Fig. 5), and $t2$ is the total height of the four-taxon tree (see $t2$ in Fig. 5). To correlate this distance with the rate of interpretation error of D-statistics, we identified ten intersecting subsets of quartets according to their R value: subsets for which R is higher than a threshold "x", with "x" varying from 0 to 0.9 with a step of 0.1. We computed the rate of interpretation error for each of the 10 subsets.

2.5 Simulation dataset

We simulated, for each value of N in {20, 40, 60, 80, 100} and α in {0, 1, 10, 100, 1000}, 100 species trees with N species, and for each species tree, 100 independent gene trees with independent introgression events. For each gene tree/species tree pair, 20 species were uniformly sampled from N (extant species), and the rest were pruned, resulting in 250,000 pairs of trees each with 20 leaves.

3 Results

3.1 Bear phylogeny exemplifies the problem of interpreting the D-statistic without taking unsampled lineages into account

Using genomic data, we show how the presence or absence of one lineage, in this case the polar bear, can lead to opposite interpretations of the D-statistic if interpreted without considering ghost lineages. From the bear phylogeny (Material and Methods; phylogeny shown in Figure 3A), we removed either the Slovenian bear (Figure 3B, subset 1), which is not thought to have introgressed with other bear species from the ingroup, or the polar bear (subset 2), which is suspected to have introgressed with brown bears from Alaska (Cahill et al. 2013; Liu et al. 2014; Lan et al. 2016; Kumar et al. 2017; Barlow et al. 2018). In the first subset, we identified 175,413 ABBA patterns and 226,992 BABA patterns resulting in a significant negative D-score (D = -0.128, Z = -11.71), which is congruent with introgression between polar bears and Alaskan bears (Fig. 3B) in the usual interpretation of the test. In the second subset, we identified 266,173 ABBA patterns and 213,830 BABA patterns resulting in a significant positive D-score (D = 0.109, Z = 14.24).

If we were to interpret this second result as evidence of introgression between the lineages sampled here (Alaskan, Russian, Slovenian and black bears), we would conclude that there was introgression between bears from Slovenia and Russia (Figure 3B), even though this significant positive D-statistic could also be attributed to introgression between polar bears (not sampled here) and Alaskan bears. The latter attribution, however, relies on our knowledge of the existence of polar bears, considered a ghost lineage. This hypothesis could similarly be called into question if we knew the existence of another lineage, because we can never assume that we know all the lineages leading to extant or extinct species. Thus, even with good taxonomic sampling, there is a real chance that an interpretation based only on known lineages wrongly infers introgression events.

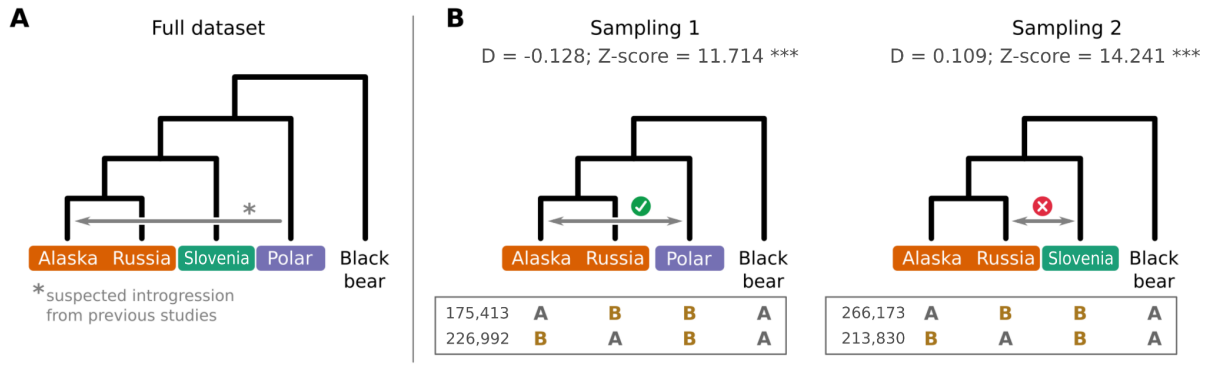


Figure 3. The effect of sampling on the interpretation of the D-statistic, using bear genomic

data as an example. **A**. Phylogenetic relationship of the five bear taxa sampled. The grey arrow shows the introgression inferred from previous studies. **B**. D-statistic calculated from two four-taxon subsets. The number of ABBA and BABA patterns is given below the trees. In subset 1, the Slovenian bear, a lineage that is not thought to be involved in introgression, is removed. In subset 2, the donor of the introgression shown in 3A (*i.e.* the polar bear) is removed. Introgressions were inferred from the D-statistic (grey arrows), and their congruence with other studies (green tick = congruent, red cross = not congruent) is indicated above the arrow.

3.2 Significant D-statistics are often due to introgressions from ghost lineages

Using simulated datasets (Material and Methods), we estimated the frequency of misinterpreting introgression events. We counted the number of D-statistics due to midgroup ghost introgressions (corresponding to the *proportion of erroneous interpretations*). We observed between 15% and 100% of erroneous interpretations, the frequency of which increased with (i) the proportion of unsampled lineages, (ii) the distance between ingroup (P1, P2, P3) and outgroup O, and (iii) the probability of introgression between distantly related lineages (Supplementary Material Section 2 for a complete summary). We describe these three trends in detail in the following sections and relate the range of each parameter we used to biological data such as the bear genomes described above.

3.2.1. The proportion of unsampled species

The effect of absent lineages on the interpretation of the D-statistic was investigated using simulated species trees with N = {20, 40, 60, 80, 100} from which 20 species were randomly sampled. This corresponds to a sampling effort ranging from 100% (20 species out of 20) to 20% (20 species out of 100). We observed that low sampling contributes to an increase in the number of misinterpreted D-statistics due to ghost introgression (Fig. 4). While the mean proportion of erroneous interpretations is ~25% when 100% of extant lineages are sampled, it is close to 60% when only 20% of extant lineages are sampled.

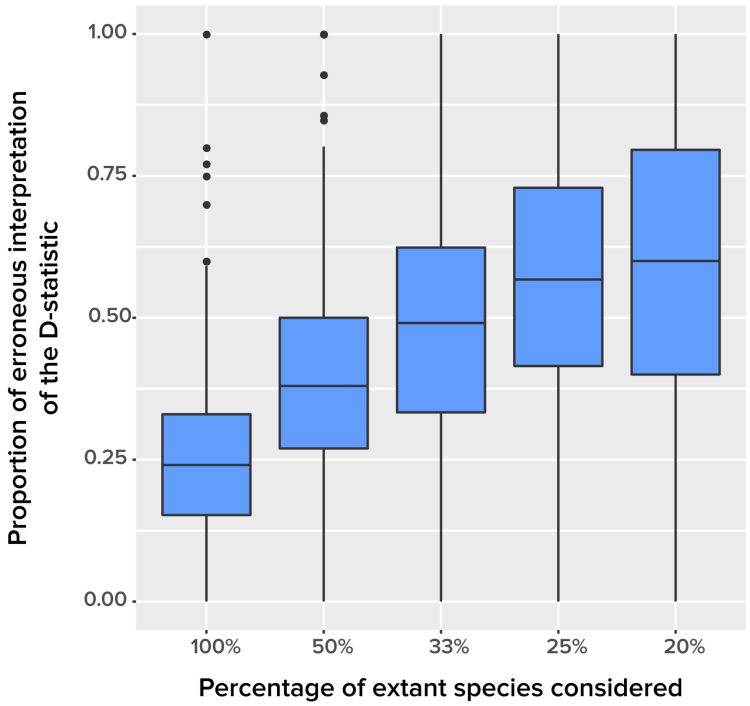


Figure 4. The effect of taxonomic sampling (x-axis) on the proportion of erroneous interpretations of the D-statistic (y-axis). The error rate is increasing with the amount of unknown.

For example, to study introgression in bears, Barlow et al. (2018) sampled 13 *Ursus* species and one *Ailuropoda* species, both members of the family Ursidae, based on the availability of genomic data. However, the Global Biodiversity Information Facility (gbif.org) reports 140 species in Ursidae, 100 of them belonging to the genus *Ursus*, which is close to the highest error rate in the simulations.

By contrast, and unlike we initially expected, we found no correlation between the number of extinct lineages in the tree and the proportion of erroneous interpretations of the D-statistic (Supplementary Material Section 3). Our interpretation is that increasing the number of extinct lineages is achieved by increasing the probability of extinction in the birth-death process, which also increases the probability that mid-group lineages, the possible source of ghost introgressions, become extinct before having the opportunity to introgress. Further investigations are needed to better characterize this effect.

3.2.2 The distance between outgroup and ingroup

In ABBA-BABA tests, the outgroup is usually chosen so that its distance from the ingroup is sufficient to minimize the chance of introgression between the two (Green et al. 2010; Osborne et al. 2016; Irwin et al. 2018). Zheng and Janke (2018) stated that the distance between outgroup and ingroup had little to no impact on the sensitivity of the D-statistic. However, they focused on evaluating the effect of saturation of sequence substitutions in the outgroup and did not consider possible introgressions from mid- or outgroups.

From our simulations, we observed that the proportion of ghost introgressions (leading to erroneous interpretations) increased with *R*, the relative distance to the outgroup (see Materials and Methods). On average, when $R > 0.3$, more than 50% of the significant D-statistics are associated with ghost introgressions (Fig. 5). We found that, when $R > 0.7$, a median of 100% of D-statistics resulted from ghost introgressions.

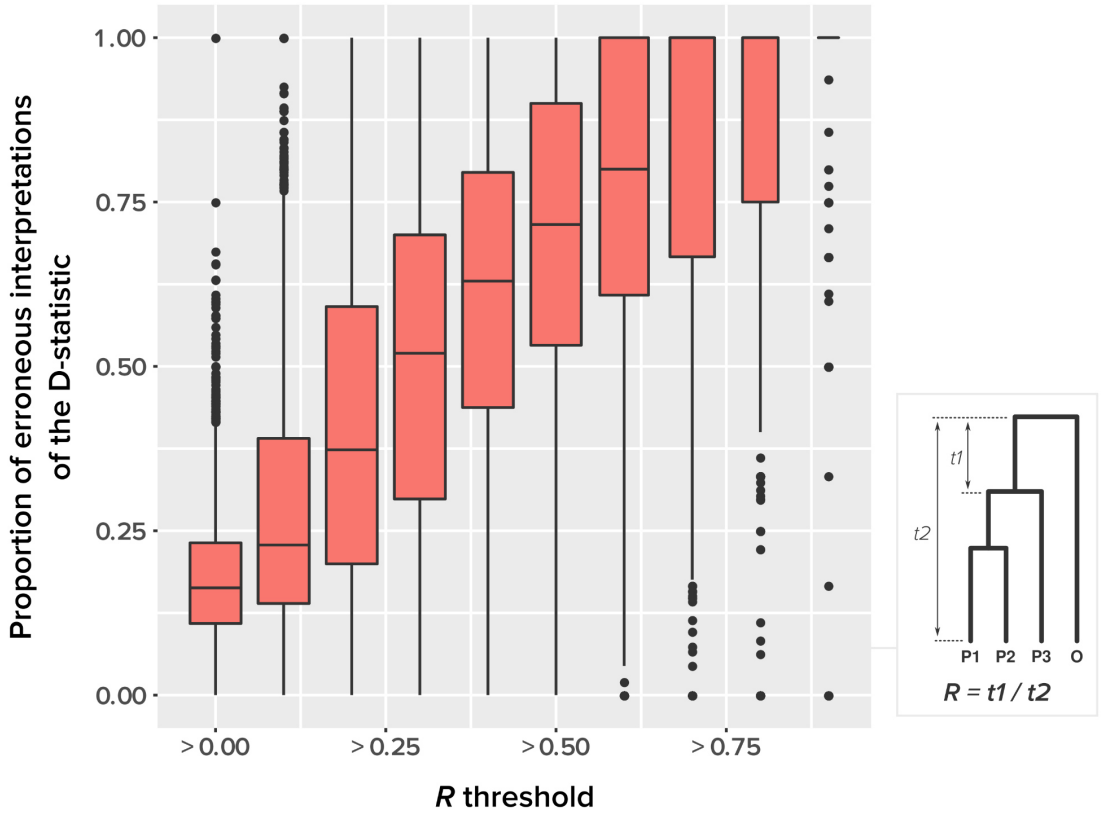


Figure 5. The relationship between outgroup distance (*R*) and the proportion of erroneous interpretations of the D-statistic for different thresholds of *R* (x-axis). The distances *t1* and *t2* used to calculate the relative distance to the outgroup (*R*) are described on the right box.

To relate our findings to different biological data, Green et al. 2010 used the D-statistic to detect introgression between Neanderthal and modern humans and had a *R* value equal to 0.873: 825,000 years separated modern humans from Neanderthal, while 6.5 million years separated humans from chimps (the outgroup) (Green et al. 2010). The study of bears of Barlow et al. (2018) used two different outgroups, black bears and pandas, with *R* values *ca*. 0.4 and 0.9, respectively. According to our simulations, all these values fall within the range of high probability of erroneous interpretation.

3.2.3 The distance between donor and recipient

Species that are genetically close have a higher chance to introgress, which could mitigate the previous result. Indeed, if the distance between outgroup and ingroup is sufficient to prevent introgression between the two, then putative midgroups ghost lineages may also be too distant. It is well known that the probability of hybridization, and consequently of introgression, decreases as genetic distance between species increases (Edmands 2002; Mallet 2005; Chapman and Burke 2007; Montanari et al. 2014).

To test whether this observation mitigates the importance of ghost lineages when detecting introgression, we used different values of α, a parameter that lets the probability of introgression vary with the phylogenetic distance between donor and recipient (see Material and Methods). We used $\alpha = \{0, 1, 10, 100, 1000\}$. When $\alpha = 0$, introgressions occur uniformly at random; when $\alpha = 1000$, introgressions occur almost exclusively between sister taxa (Figure 6A).

We observed that the impact of outgroup distance on the proportion of erroneous interpretations decreased with increasing values of α (Figure 6B). As expected, in simulations where alpha is maximum ($\alpha = 1000$), the proportion of significant D-statistics due to ghost introgressions is not affected by the distance separating ingroup and outgroup. Nevertheless, this proportion remains quite high, and its median value does not fall below 25% under our settings. For other values of α, this proportion is higher and increases with distance to the outgroup.

We investigated what could be a realistic value of alpha in biological data. We used phylogenetic trees from several studies where one or several introgression paths were identified. We counted the number of internal nodes between donor and recipient. Then, we randomly simulated the same number of introgressions with different values of alpha and

calculated the average number of internal nodes between donor and recipient (removing introgressions between sister branches as this cannot be observed when using D-statistic). We retained the value of alpha giving the average number of internal nodes that was closest to what is observed in the biological data (Supplementary Material Section 4). We analyzed the bear phylogeny of Hailer et al. (2012) and found that $\alpha = 0$ gives the closest result (actual number of nodes = 5; with $\alpha = 0$, the average number of nodes from the simulations was 4.5). In the phylogeny of the *Bos* species complex of Wu et al. (2018), the actual number of nodes is higher than in our simulations even with $\alpha = 0$. The same result was found with the phylogeny of the *Anopheles gambiae* species complex of Fontaine et al. (2015). By contrast, for the woodcreeper phylogeny of Pulido-Santacruz et al. (2020), we found that the value closest to the biological dataset was obtained with $\alpha = 100$. Lastly, it was estimated that values of α between 100 and 1000 best fit the spider phylogeny of Leduc-Robert and Maddison (2018). These results are described in full in Supplementary Materials section 3. These examples, taken from very diverse organisms, tend to show that the higher probability of introgression between closely related species is not sufficient to secure a safe zone for the ABBA-BABA test.

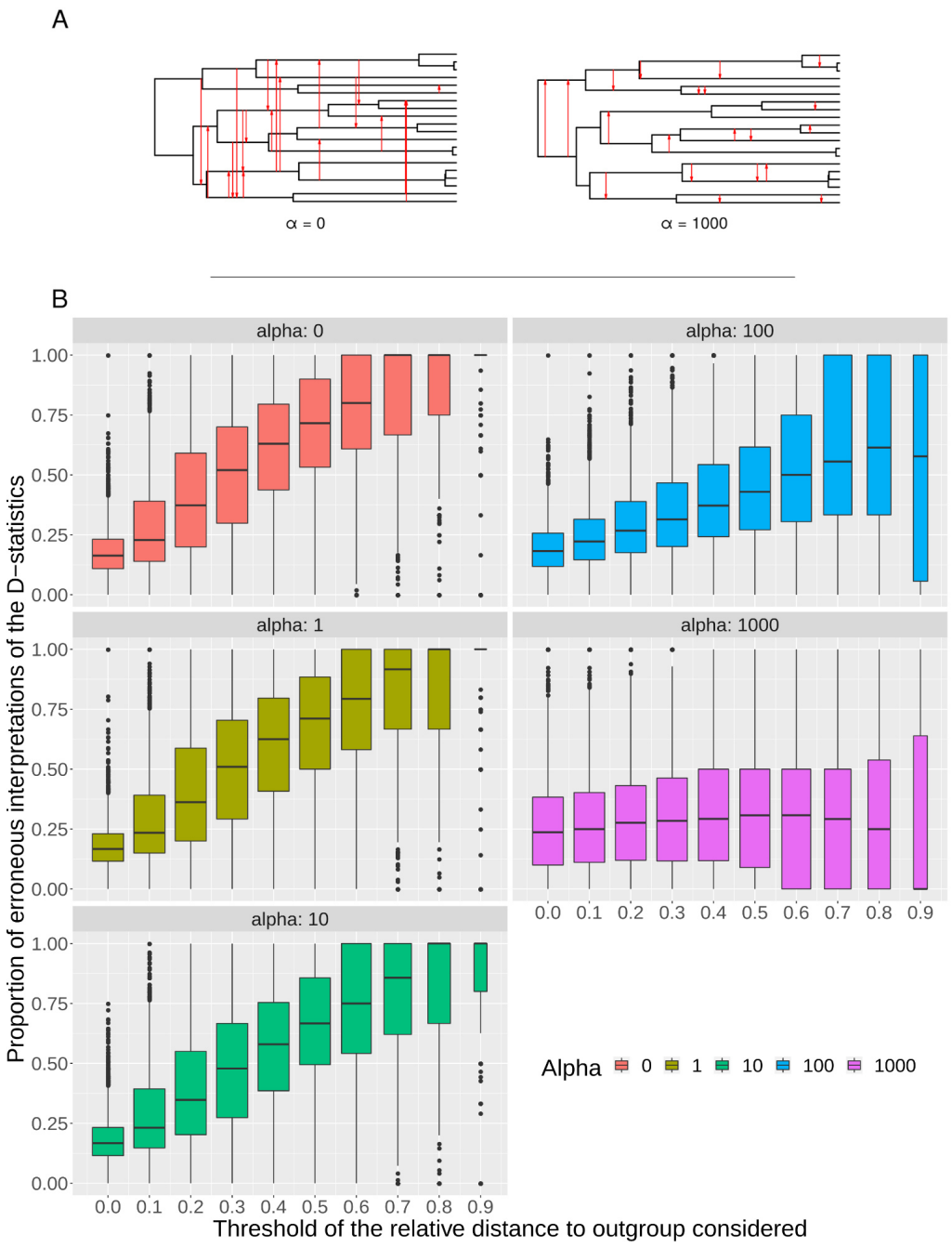


Figure 6. The effect of the probability of introgression on the proportion of erroneous interpretations of the D-statistic. **A.** Illustration of the effect of the α parameter, which imposes constraints on introgression in relation to phylogenetic distance. **B**. Relationship between relative outgroup distance (x-axis) and the proportion of erroneous interpretations (y-axis) for different levels of constraint on introgression related to phylogenetic distance, determined by $\alpha = \{0, 1, 10, 100, 1000\}$.

4 Discussion

4.1 Ghost lineages: An important factor affecting introgression tests

Different parameters are known to affect the robustness or sensitivity of the D-statistic. For instance, variations in population size (Eriksson and Manica 2012; Lohse and Frantz 2014; Martin et al. 2015; Zheng and Janke 2018) and/or ancestral population structure (Durand et al. 2011; Lohse and Frantz 2014; Martin et al. 2015) have been shown to produce significant D-statistics in the absence of introgression. Applying the D-statistic to smaller genomic windows, rather than over the entire genome, gives very variable estimates of D (Martin et al. 2015). Complex introgression scenarios, with more than one introgression in the quartet, are another source of error (Rogers and Bohlender 2015; Elworth et al. 2018). Our findings suggest that, in addition to the variables listed above, the interpretation of the D-statistic should systematically, and maybe primarily, take into account ghost lineages.

Recently, Hibbins and Hahn (2021) published results that are in line with our findings. Their simulation study confirms that introgression from a midgroup ghost lineage can result in a significant D-statistic, which may lead to the misidentification of the identity of the lineages involved in the introgression if only known lineages are considered. Similar results have been observed with the D3 statistic (Hahn and Hibbins 2019), a test for detecting introgression that uses only three lineages and the branch lengths of the phylogeny. While the study of Hibbins and Hahn (2021) confirms that ghost introgressions may lead to erroneous interpretations, they do not quantify the extent to which this factor affects the interpretation of the D-statistic.

4.2. An intractable incompleteness

Although some families are believed to be extensively described (Chapman and Burke 2007), it is not possible and probably will never be possible to assume that we work with an exhaustive taxon set. A study from 2011 estimated that 8.7 million of eukaryote species are alive today (Mora et al. 2011), and a study from 2016 estimated that there are 1 trillion species on Earth (Locey and Lennon 2016). By contrast, 2.5 million species have been described and catalogued in The Catalogue of Life (CoL). The kingdom Animalia has the most descriptions with 1.4M catalogued species, while Plantae, Fungi and other kingdoms have 375K, 145K and 81K catalogued species, respectively. This means that, at best, we know 25% of the biodiversity that is alive today. In practice, there is a strong disparity in the

percentage of undescribed species among taxonomic groups. Chapman estimated in 2009 that less than 20% of the phylum Chordata is yet to be described while 80% of the class Insecta is still unknown to us (Chapman et al. 2009). The proportion of undescribed species is even higher in prokaryotes with less than 1% of species described for viruses, Archaea and Bacteria, and because microbial biodiversity is harder to estimate than that of macro-organisms, these numbers could be orders of magnitude lower. Thus, the effects shown here cannot be circumvented by adding or expecting more species and improving computational techniques to handle larger datasets. We are bound to work with a very small fraction of what exists. According to our simulations, with 25% of species sampled, on average more than 50% of introgressions could be due to ghost lineages and subsequently be misinterpreted by a D-statistic. This implies that, with the exception of some well described eukaryote groups such as the genus *Homo*, our lack of taxonomic knowledge will greatly impact the reliability of the D-statistic.

4.3 Other introgression detection methods

Several other methods have been developed to mitigate some of the limitations of the D-statistic, but their robustness to ghost lineages has not yet been explored.

It is possible to apply the D-statistic test in datasets with more than four species by performing multiple tests on different quartets. The D-statistics are then analyzed together, using the interpretation of each individual test as a constraint for the interpretation of the other tests. This enables a finer detection of introgression, the identification of donors and recipients (while a single test cannot distinguish the donor from the recipient), and possibly assigning introgression events to groups of taxa instead of single taxa (Pease et al. 2016; Rouard et al. 2018; Suvorov et al. 2020). However, if each individual test is interpreted, as it is usually done, without considering the possibility of ghost introgressions, the joint interpretation of multiple tests will miss a high number of scenarios. Moreover, there is no method that formalizes the constraints from multiple tests and no guarantee that the result is correct or unique, and that the order in which single tests are analysed does not matter.

Extensions of the D-statistic, namely the partitioned-D (Eaton and Ree 2013) and D_{FOIL} (Pease and Hahn 2015) tests, have been proposed to infer introgression in 5-taxon phylogenies (instead of four) and to polarize (in some cases a direction to the introgression can be assessed) the introgression. Although D_{FOIL} can detect more introgression patterns than

ABBA-BABA, it is still blind to ghost lineages, and thus presents similar theoretical possibilities of misinterpretation as the ABBA-BABA test (see Supplementary Material Section 5 for a listing of the patterns that lead to misinterpretation). This possibility is mentioned by Pease and Hahn (2015) but the proportion of misinterpretations remains to be quantified, and alternative interpretations handling ghost lineages has not been written.

Soraggi et al. (2018) proposed an extension of the D-statistic (D_{ext}) to study introgression events among non-African human populations using Africans as the outgroup. This test is robust to introgression from an external group which is not part of the analysed populations. However, the use of this version of D-statistic is restricted to extinct clades for which introgression events with extant species have been already identified as the Neanderthal introgression. This precludes its use in cases where there is no *a priori* knowledge on the existence of ghost lineages.

Other more global methods, such as STRUCTURE/ADMIXTURE (Pritchard et al. 2000; Tang et al. 2005), Treemix (Pickrell and Pritchard 2012) and Phylonet (Than et al. 2008; Wen et al. 2018), have been designed to detect introgression across an entire phylogeny. These tools do not consider ghost lineages and their potential effect on the detected signal. Thus, introgression events are only inferred between known lineages and branches in a tree. It is interesting to note that two of these tools, STRUCTURE and ADMIXTURE, which are popular choices for reconstructing genetic history and testing admixture scenarios, have recently been shown to be subject to misinterpretation due to ghost introgressions (Lawson et al. 2018). The impact of ghost lineages on the detection of introgression is therefore not just a question of using the right tool.

4.4. Application of the D-statistic in the light of ghost introgressions.

Now that the importance of the possibility of ghost introgression is recognized, and that no current method is able to handle the effects of ghost lineages, it is time to adapt all methods or develop new methods to take this factor into account.

For the single D-statistic test (four-taxon quartet), the solution is simply to take this uncertainty into account by considering alternative scenarios with at least equal probability. In phylogenies with more than four taxa, the D-statistics from all quartets could be analyzed using an algorithmic method that would combine a set of scenarios. This will require formalizing the objective (*i.e.* minimizing the number of incoherences between quartet

results), choosing a set of quartets according to this objective and devising a combinatorial algorithm that handles, for each quartet, information from several possible scenarios including ghost lineages.

Note that this approach will not only avoid interpretation errors, but will possibly point to the existence of unknown lineages that have contributed through introgression to the genomes of known lineages. Therefore, this approach would combine the detection of introgression and the detection of unsampled or extinct taxa. This has already been achieved with *ad hoc* methods for human (Prüfer et al. 2014; Dannemann and Racimo 2018) and whale lineages (Foote et al. 2019) and could be generalized to enrich the phylogeny of known species with unknown species, for which we have no trace other than the genes that have lived, for a while, in them. This is a promising route for future work.

5. Conclusion

The D-statistic is a key tool for studying introgression as it provides, in specific cases, a robust test for detecting gene flow. However, our results show that one important caveat of this test is its lack of consideration of ghost lineages, which can lead to the misinterpretation of a significant result. Thus, the *bona fide* interpretation of a single significant D-statistic should be a set of possible scenarios that include the possibility of ghost introgressions, which are equally likely in the absence of other information. Based on our simulations, we have suggested that ghost introgressions are often the most likely scenario. It is possible that in the future, the usual interpretation of a significant D-statistic , *i.e.* ingroup introgression, becomes the exception rather than the rule.

Acknowledgments

This work was supported by the French National Research Agency (Grants ANR-18-CE02-0007-01 and ANR-19-CE45-0010). Simulations were performed using the computing facilities of the CC LBBE/PRABI. We thank Gergely Szöllosi for useful discussions

Software Availability

All codes used to generate and analyze the simulations performed in this study are available at: https://github.com/theotricou/Ghost_abba_baba.

References

Barlow A., Cahill J.A., Hartmann S., Theunert C., Xenikoudakis G., Fortes G.G., Paijmans J.L.A., Rabeder G., Frischauf C., Grandal-d'Anglade A., García-Vázquez A., Murtskhvaladze M., Saarma U., Anijalg P., Skrbinšek T., Bertorelle G., Gasparian B., Bar-Oz G., Pinhasi R., Slatkin M., Dalén L., Shapiro B., Hofreiter M. 2018. Partial genomic survival of cave bears in living brown bears. Nat. Ecol. Evol. 2:1563–1570.

Barlow A., Cahill J.A., Hartmann S., Theunert C., Xenikoudakis G., Fortes G.G., Paijmans J.L.A., Rabeder G., Frischauf C., Grandal-D'Anglade A., García-Vázquez A., Murtskhvaladze M., Saarma U., Anijalg P., Skrbinšek T., Bertorelle G., Gasparian B., Bar-Oz G., Pinhasi R., Slatkin M., Dalén L., Shapiro B., Hofreiter M. 2019. Data from: Partial genomic survival of cave bears in living brown bears. :6331386741 bytes.

Bordewich M., Semple C. 2005. On the Computational Complexity of the Rooted Subtree Prune and Regraft Distance. Ann. Comb. 8:409–423.

Cahill J.A., Green R.E., Fulton T.L., Stiller M., Jay F., Ovsyanikov N., Salamzade R., John J.S., Stirling I., Slatkin M., Shapiro B. 2013. Genomic Evidence for Island Population Conversion Resolves Conflicting Theories of Polar Bear Evolution. PLOS Genet. 9:e1003345.

Chapman A.D., Australia, Department of the Environment W. Heritage, and the Arts, Australian Biological Resources Study. 2009. Numbers of living species in Australia and the world. Canberra, A.C.T.: Department of the Environment, Water, Heritage and the Arts.

Chapman M.A., Burke J.M. 2007. Genetic divergence and hybrid speciation. Evol. Int. J. Org. Evol. 61:1773–1780.

Dannemann M., Racimo F. 2018. Something old, something borrowed: admixture and adaptation in human evolution. Curr. Opin. Genet. Dev. 53:1–8.

Durand E.Y., Patterson N., Reich D., Slatkin M. 2011. Testing for Ancient Admixture between Closely Related Populations. Mol. Biol. Evol. 28:2239–2252.

Eaton D.A.R., Hipp A.L., González‐Rodríguez A., Cavender‐Bares J. 2015. Historical introgression among the American live oaks and the comparative nature of tests for introgression. Evolution. 69:2587–2601.

Eaton D.A.R., Ree R.H. 2013. Inferring Phylogeny and Introgression using RADseq Data: An Example from Flowering Plants (Pedicularis: Orobanchaceae). Syst. Biol. 62:689–706.

Edmands S. 2002. Does parental divergence predict reproductive compatibility? Trends Ecol. Evol. 17:520–527.

Elworth R.A.L., Allen C., Benedict T., Dulworth P., Nakhleh L. 2018. DGEN: A Test Statistic for Detection of General Introgression Scenarios. bioRxiv.:348649.

Eriksson A., Manica A. 2012. Effect of ancient population structure on the degree of polymorphism shared between modern human populations and ancient hominins. Proc. Natl. Acad. Sci. 109:13956–13960.

Fontaine M.C., Pease J.B., Steele A., Waterhouse R.M., Neafsey D.E., Sharakhov I.V., Jiang X., Hall A.B., Catteruccia F., Kakani E., Mitchell S.N., Wu Y.-C., Smith H.A., Love R.R., Lawniczak M.K., Slotman M.A., Emrich S.J., Hahn M.W., Besansky N.J. 2015. Extensive introgression in a malaria vector species complex revealed by phylogenomics. Science. 347:1258524.

Foote A.D., Martin M.D., Louis M., Pacheco G., Robertson K.M., Sinding M.-H.S., Amaral A.R., Baird R.W., Baker C.S., Ballance L., Barlow J., Brownlow A., Collins T., Constantine R., Dabin W., Rosa L.D., Davison N.J., Durban J.W., Esteban R., Ferguson S.H., Gerrodette T., Guinet C., Hanson M.B., Hoggard W., Matthews C.J.D., Samarra F.I.P., Stephanis R. de, Tavares S.B., Tixier P., Totterdell J.A., Wade P., Excoffier L., Gilbert M.T.P., Wolf J.B.W., Morin P.A. 2019. Killer whale genomes reveal a complex history of recurrent admixture and vicariance. Mol. Ecol. 28:3427–3444.

Galtier N., Daubin V. 2008. Dealing with incongruence in phylogenomic analyses. Philos. Trans. R. Soc. B Biol. Sci. 363:4023–4029.

Green R.E., Krause J., Briggs A.W., Maricic T., Stenzel U., Kircher M., Patterson N., Li H., Zhai W., Fritz M.H.-Y., Hansen N.F., Durand E.Y., Malaspinas A.-S., Jensen J.D., Marques-Bonet T., Alkan C., Prüfer K., Meyer M., Burbano H.A., Good J.M., Schultz R., Aximu-Petri A., Butthof A., Höber B., Höffner B., Siegemund M., Weihmann A., Nusbaum C., Lander E.S., Russ C., Novod N., Affourtit J., Egholm M., Verna C., Rudan P., Brajkovic D., Kucan Ž., Gušic I., Doronichev V.B., Golovanova L.V.,

Lalueza-Fox C., Rasilla M. de la, Fortea J., Rosas A., Schmitz R.W., Johnson P.L.F., Eichler E.E., Falush D., Birney E., Mullikin J.C., Slatkin M., Nielsen R., Kelso J., Lachmann M., Reich D., Pääbo S. 2010. A Draft Sequence of the Neandertal Genome. Science. 328:710–722.
Hahn M.W., Hibbins M.S. 2019. A Three-Sample Test for Introgression. Mol. Biol. Evol. 36:2878–2882.
Hailer F., Kutschera V.E., Hallström B.M., Klassert D., Fain S.R., Leonard J.A., Arnason U., Janke A. 2012. Nuclear Genomic Sequences Reveal that Polar Bears Are an Old and Distinct Bear Lineage. Science. 336:344–347.
Hedrick P.W. 2013. Adaptive introgression in animals: examples and comparison to new mutation and standing variation as sources of adaptive variation. Mol. Ecol. 22:4606–4618.
Hibbins M., Hahn M. 2021. Phylogenomic approaches to detecting and characterizing introgression. .
Irwin D.E., Milá B., Toews D.P.L., Brelsford A., Kenyon H.L., Porter A.N., Grossen C., Delmore K.E., Alcaide M., Irwin J.H. 2018. A comparison of genomic islands of differentiation across three young avian species pairs. Mol. Ecol. 27:4839–4855.
Keuler R., Garretson A., Saunders T., Erickson R.J., Andre N.S., Grewe F., Smith H., Lumbsch H.T., Huang J.-P., Clair L.L.S., Leavitt S.D. 2020. Genome-scale data reveal the role of hybridization in lichen-forming fungi. Sci. Rep. 10:1497.
Kulathinal R.J., Stevison L.S., Noor M.A.F. 2009. The Genomics of Speciation in Drosophila: Diversity, Divergence, and Introgression Estimated Using Low-Coverage Genome Sequencing. PLOS Genet. 5:e1000550.
Kumar V., Lammers F., Bidon T., Pfenninger M., Kolter L., Nilsson M.A., Janke A. 2017. The evolutionary history of bears is characterized by gene flow across species. Sci. Rep. 7:46487.
Lan T., Cheng J., Ratan A., Miller W., Schuster S.C., Farley S., Shideler R.T., Mailund T., Lindqvist C. 2016. Genome-wide evidence for a hybrid origin of modern polar bears. bioRxiv.:047498.
Lawson D.J., van Dorp L., Falush D. 2018. A tutorial on how not to over-interpret STRUCTURE and ADMIXTURE bar plots. Nat. Commun. 9:3258.
Leduc-Robert G., Maddison W.P. 2018. Phylogeny with introgression in Habronattus jumping spiders (Araneae: Salticidae). BMC Evol. Biol. 18:24.
Li R., Fan W., Tian G., Zhu H., He L., Cai J., Huang Q., Cai Q., Li B., Bai Y., Zhang Z., Zhang Y., Wang W., Li J., Wei F., Li H., Jian M., Li J., Zhang Z., Nielsen R., Li D., Gu W., Yang Z., Xuan Z., Ryder O.A., Leung F.C.-C., Zhou Y., Cao J., Sun X., Fu Y., Fang X., Guo X., Wang B., Hou R., Shen F., Mu B., Ni P., Lin R., Qian W., Wang G., Yu C., Nie W., Wang J., Wu Z., Liang H., Min J., Wu Q., Cheng S., Ruan J., Wang M., Shi Z., Wen M., Liu B., Ren X., Zheng H., Dong D., Cook K., Shan G., Zhang H., Kosiol C., Xie X., Lu Z., Zheng H., Li Y., Steiner C.C., Lam T.T.-Y., Lin S., Zhang Q., Li G., Tian J., Gong T., Liu H., Zhang D., Fang L., Ye C., Zhang J., Hu W., Xu A., Ren Y., Zhang G., Bruford M.W., Li Q., Ma L., Guo Y., An N., Hu Y., Zheng Y., Shi Y., Li Z., Liu Q., Chen Y., Zhao J., Qu N., Zhao S., Tian F., Wang X., Wang H., Xu L., Liu X., Vinar T., Wang Y., Lam T.-W., Yiu S.-M., Liu S., Zhang H., Li D., Huang Y., Wang X., Yang G., Jiang Z., Wang J., Qin N., Li L., Li J., Bolund L., Kristiansen K., Wong G.K.-S., Olson M., Zhang X., Li S., Yang H., Wang J., Wang J. 2010. The sequence and de novo assembly of the giant panda genome. Nature. 463:311–317.
Liu L., Bosse M., Megens H.-J., Frantz L.A.F., Lee Y.-L., Irving-Pease E.K., Narayan G., Groenen M.A.M., Madsen O. 2019. Genomic analysis on pygmy hog reveals extensive interbreeding during wild boar expansion. Nat. Commun. 10:1992.
Liu S., Lorenzen E.D., Fumagalli M., Li B., Harris K., Xiong Z., Zhou L., Korneliussen T.S., Somel M., Babbitt C., Wray G., Li J., He W., Wang Z., Fu W., Xiang X., Morgan C.C., Doherty A., O'Connell M.J., McInerney J.O., Born E.W., Dalén L., Dietz R., Orlando L., Sonne C., Zhang G., Nielsen R., Willerslev E., Wang J. 2014. Population Genomics Reveal Recent Speciation and Rapid Evolutionary Adaptation in Polar Bears. Cell. 157:785–794.
Locey K.J., Lennon J.T. 2016. Scaling laws predict global microbial diversity. Proc. Natl. Acad. Sci. 113:5970–5975.
Lohse K., Frantz L.A.F. 2014. Neandertal Admixture in Eurasia Confirmed by Maximum-Likelihood Analysis of Three Genomes. Genetics. 196:1241–1251.
Maddison W.P. 1997. Gene Trees in Species Trees. Syst. Biol. 46:523–536.
Maddison W.P., Knowles L.L. 2006. Inferring Phylogeny Despite Incomplete Lineage Sorting. Syst. Biol. 55:21–30.
Mallet J. 2005. Hybridization as an invasion of the genome. Trends Ecol. Evol. 20:229–237.
Martin S.H., Dasmahapatra K.K., Nadeau N.J., Salazar C., Walters J.R., Simpson F., Blaxter M., Manica A., Mallet J., Jiggins C.D. 2013. Genome-wide evidence for speciation with gene flow in Heliconius butterflies. Genome Res. 23:1817–1828.

Martin S.H., Davey J.W., Jiggins C.D. 2015. Evaluating the Use of ABBA–BABA Statistics to Locate Introgressed Loci. Mol. Biol. Evol. 32:244–257.
Massardo D., VanKuren N.W., Nallu S., Ramos R.R., Ribeiro P.G., Silva-Brandão K.L., Brandão M.M., Lion M.B., Freitas A.V.L., Cardoso M.Z., Kronforst M.R. 2020. The roles of hybridization and habitat fragmentation in the evolution of Brazil's enigmatic longwing butterflies, Heliconius nattereri and H. hermathena. BMC Biol. 18:84.
Meier J.I., Marques D.A., Mwaiko S., Wagner C.E., Excoffier L., Seehausen O. 2017. Ancient hybridization fuels rapid cichlid fish adaptive radiations. Nat. Commun. 8:14363.
Meyer M., Kircher M., Gansauge M.-T., Li H., Racimo F., Mallick S., Schraiber J.G., Jay F., Prüfer K., Filippo C. de, Sudmant P.H., Alkan C., Fu Q., Do R., Rohland N., Tandon A., Siebauer M., Green R.E., Bryc K., Briggs A.W., Stenzel U., Dabney J., Shendure J., Kitzman J., Hammer M.F., Shunkov M.V., Derevianko A.P., Patterson N., Andrés A.M., Eichler E.E., Slatkin M., Reich D., Kelso J., Pääbo S. 2012. A High-Coverage Genome Sequence from an Archaic Denisovan Individual. Science. 338:222–226.
Montanari S.R., Hobbs J.-P.A., Pratchett M.S., Bay L.K., Van Herwerden L. 2014. Does genetic distance between parental species influence outcomes of hybridization among coral reef butterflyfishes? Mol. Ecol. 23:2757–2770.
Mora C., Tittensor D.P., Adl S., Simpson A.G.B., Worm B. 2011. How Many Species Are There on Earth and in the Ocean? PLoS Biol. 9:e1001127.
Osborne O.G., Chapman M.A., Nevado B., Filatov D.A. 2016. Maintenance of Species Boundaries Despite Ongoing Gene Flow in Ragworts. Genome Biol. Evol. 8:1038–1047.
Ottenburghs J. 2020. Ghost Introgression: Spooky Gene Flow in the Distant Past. BioEssays. 42:2000012.
Ottenburghs J., Kraus R.H.S., van Hooft P., van Wieren S.E., Ydenberg R.C., Prins H.H.T. 2017. Avian introgression in the genomic era. Avian Res. 8:30.
Paradis E., Schliep K. 2019. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. Bioinformatics. 35:526–528.
Patterson N., Moorjani P., Luo Y., Mallick S., Rohland N., Zhan Y., Genschoreck T., Webster T., Reich D. 2012. Ancient Admixture in Human History. Genetics. 192:1065–1093.
Pease J.B., Haak D.C., Hahn M.W., Moyle L.C. 2016. Phylogenomics Reveals Three Sources of Adaptive Variation during a Rapid Radiation. PLOS Biol. 14:e1002379.
Pease J.B., Hahn M.W. 2015. Detection and Polarization of Introgression in a Five-Taxon Phylogeny. Syst. Biol. 64:651–662.
Pickrell J.K., Pritchard J.K. 2012. Inference of Population Splits and Mixtures from Genome-Wide Allele Frequency Data. PLOS Genet. 8:e1002967.
Pritchard J.K., Stephens M., Donnelly P. 2000. Inference of Population Structure Using Multilocus Genotype Data. Genetics. 155:945–959.
Prüfer K., Racimo F., Patterson N., Jay F., Sankararaman S., Sawyer S., Heinze A., Renaud G., Sudmant P.H., de Filippo C., Li H., Mallick S., Dannemann M., Fu Q., Kircher M., Kuhlwilm M., Lachmann M., Meyer M., Ongyerth M., Siebauer M., Theunert C., Tandon A., Moorjani P., Pickrell J., Mullikin J.C., Vohr S.H., Green R.E., Hellmann I., Johnson P.L.F., Blanche H., Cann H., Kitzman J.O., Shendure J., Eichler E.E., Lein E.S., Bakken T.E., Golovanova L.V., Doronichev V.B., Shunkov M.V., Derevianko A.P., Viola B., Slatkin M., Reich D., Kelso J., Pääbo S. 2014. The complete genome sequence of a Neanderthal from the Altai Mountains. Nature. 505:43–49.
Pulido-Santacruz P., Aleixo A., Weir J.T. 2020. Genomic data reveal a protracted window of introgression during the diversification of a neotropical woodcreeper radiation*. Evolution. 74:842–858.
Raup D.M. 1991. Extinction: bad genes or bad luck? New York: W.W. Norton.
Rogers A.R., Bohlender R.J. 2015. Bias in estimators of archaic admixture. Theor. Popul. Biol. 100:63–78.
Rouard M., Droc G., Martin G., Sardos J., Hueber Y., Guignon V., Cenci A., Geigle B., Hibbins M.S., Yahiaoui N., Baurens F.-C., Berry V., Hahn M.W., D'Hont A., Roux N. 2018. Three New Genome Assemblies Support a Rapid Radiation in Musa acuminata (Wild Banana). Genome Biol. Evol. 10:3129–3140.
Schumer M., Cui R., Powell D.L., Rosenthal G.G., Andolfatto P. 2016. Ancient hybridization and genomic stabilization in a swordtail fish. Mol. Ecol. 25:2661–2679.
Smith J., Kronforst M.R. 2013. Do Heliconius butterfly species exchange mimicry alleles? Biol. Lett. 9:20130503.
Soraggi S., Wiuf C., Albrechtsen A. 2018. Powerful Inference with the D-Statistic on Low-Coverage Whole-Genome Data. G3 GenesGenomesGenetics. 8:551–566.
Suvorov A., Kim B.Y., Wang J., Armstrong E.E., Peede D., D'Agostino E.R.R., Price D.K., Wadell P., Lang M., Courtier-Orgogozo V., David J.R., Petrov D., Matute D.R., Schrider D.R., Comeault A.A. 2020.

Widespread introgression across a phylogeny of 155 Drosophila genomes. bioRxiv.:2020.12.14.422758.
Szöllősi G.J., Davín A.A., Tannier E., Daubin V., Boussau B. 2015. Genome-scale phylogenetic analysis finds extensive gene transfer among fungi. Philos. Trans. R. Soc. B Biol. Sci. 370:20140335.
Szöllősi G.J., Tannier E., Lartillot N., Daubin V. 2013. Lateral Gene Transfer from the Dead. Syst. Biol. 62:386–397.
Tang H., Peng J., Wang P., Risch N.J. 2005. Estimation of individual admixture: Analytical and study design considerations. Genet. Epidemiol. 28:289–301.
Than C., Ruths D., Nakhleh L. 2008. PhyloNet: a software package for analyzing and reconstructing reticulate evolutionary relationships. BMC Bioinformatics. 9:322.
Wen D., Yu Y., Zhu J., Nakhleh L. 2018. Inferring Phylogenetic Networks Using PhyloNet. Syst. Biol. 67:735–740.
Wu D.-D., Ding X.-D., Wang S., Wójcik J.M., Zhang Y., Tokarska M., Li Y., Wang M.-S., Faruque O., Nielsen R., Zhang Q., Zhang Y.-P. 2018. Pervasive introgression facilitated domestication and adaptation in the Bos species complex. Nat. Ecol. Evol. 2:1139–1145.
Zhang B.-W., Xu L.-L., Li N., Yan P.-C., Jiang X.-H., Woeste K.E., Lin K., Renner S.S., Zhang D.-Y., Bai W.-N. 2019. Phylogenomics Reveals an Ancient Hybrid Origin of the Persian Walnut. Mol. Biol. Evol. 36:2451–2461.
Zhang W., Zhang X., Li K., Wang C., Cai L., Zhuang W., Xiang M., Liu X. 2018. Introgression and gene family contraction drive the evolution of lifestyle and host shifts of hypocrealean fungi. Mycology. 9:176–188.
Zheng Y., Janke A. 2018. Gene flow analysis method, the D-statistic, is robust in a wide parameter space. BMC Bioinformatics. 19:10.

Ghost lineages highly influence the interpretation of introgression tests

Théo Tricou[1,*], Eric Tannier[1,2] and Damien M. de Vienne[1,*]

[1] Univ Lyon, Université Lyon 1, CNRS, Laboratoire de Biométrie et Biologie Évolutive UMR5558, F-69622 Villeurbanne, France
[2] INRIA Grenoble Rhône-Alpes, F-38334, France

***Corresponding authors: E-mails:** theo.tricou@univ-lyon1.fr, damien.de-vienne@univ-lyon1.fr

Supplementary Material

1. *Equivalence between D-statistics on sequence simulations and gene tree species tree comparisons*

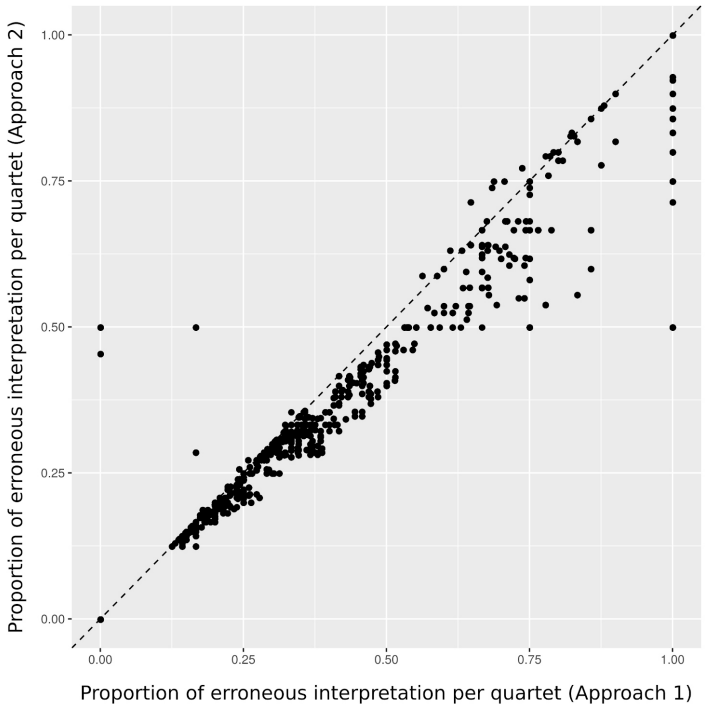
The standard simulation procedure to evaluate gene flow inference methods is to draw a species tree, choose an introgression between two of its branches and then simulate sequences along this reticulated species tree. ABBA-BABA patterns are read on these sequences and the tests are realized from these patterns.

We explore here a way to bypass the sequence simulation, which is the most computationally demanding. We construct a *gene tree* from the species tree and the introgression by performing a subtree prune and regraft (SPR). Then for each quartet of species (((P1,P2),P3),P4) we check if the gene tree has this topology on this quartet or an alternative one. If the gene tree has (((P1,P3),P2),P4), we identify this situation with an excess of BABA patterns when simulating sequences, and if the gene tree has (((P2,P3),P1),P4), with an excess of ABBA patterns. We discard the cases of introgressions involving P4.

We hypothesize that each time the gene tree shows one of these patterns, we would have obtained a significant D-statistic test if we had simulated sequences, with a stochastic uncertainty due to the random process and the simulation parameters, such as population size and sequence size. In order to test this statement we simulated sequences on a control dataset. For this we used *ms* (Hudson 2002) to simulate 10^6 independent loci with a single mutation each (resulting in 10^6 SNP pattern), evolving in populations of fixed size (Ne) of 100,000 individuals over 10^6 generations. For each quartet with topology (((P1,P2),P3),O) in the species tree and for each introgression simulated, we counted the number of loci with ABBA and BABA patterns. We then computed the D-statistics. The R package *Coala* (Staab and Metzler 2016) was used to execute *ms* and to analyse the resulting data.

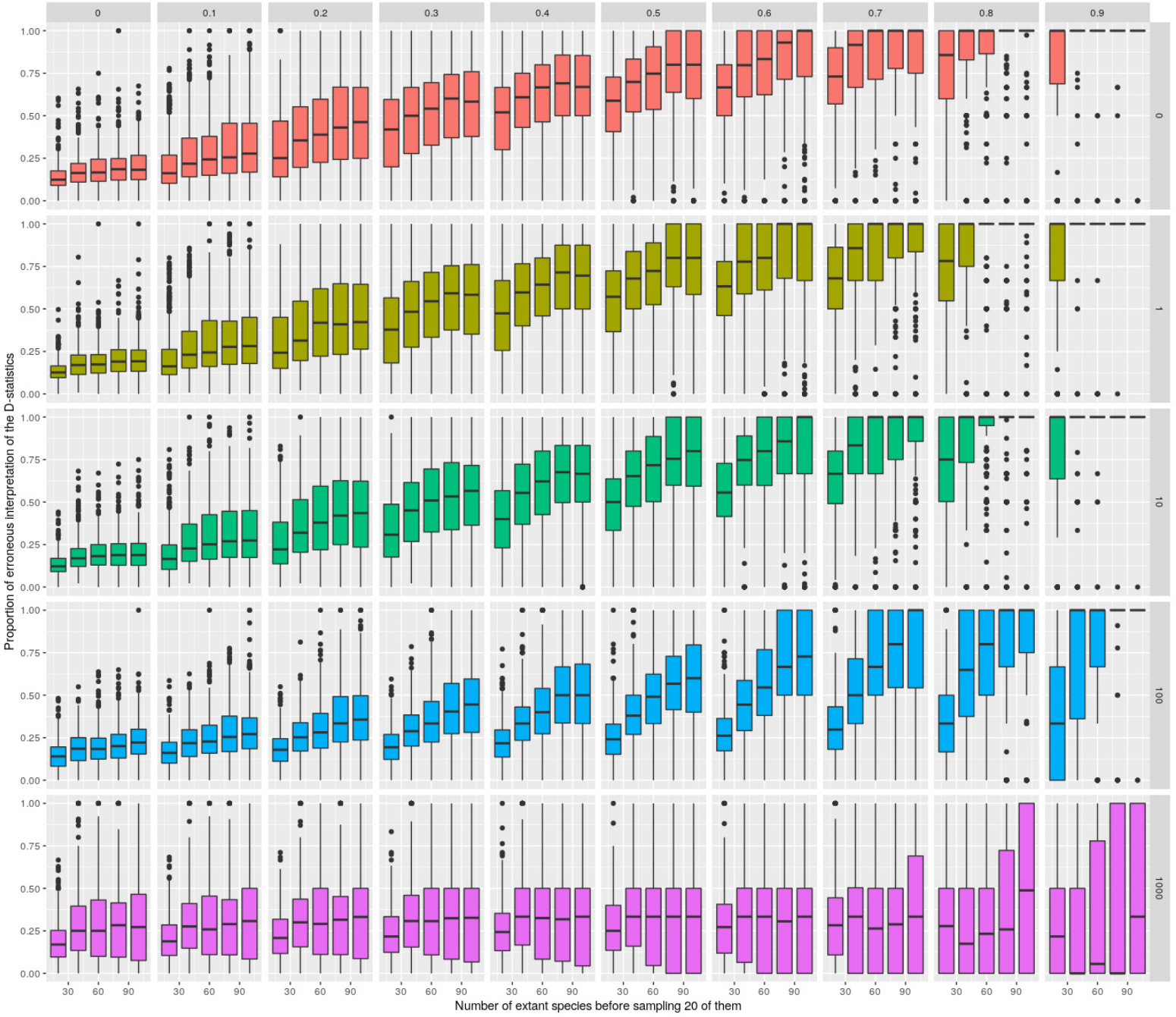
In Supplementary Figure 1 we show that the proportion of erroneous interpretation is highly correlated for both approaches ($R^2 > 0.96$).

In consequence we kept only the computationally less intensive approach for the study with several parameters to explore.



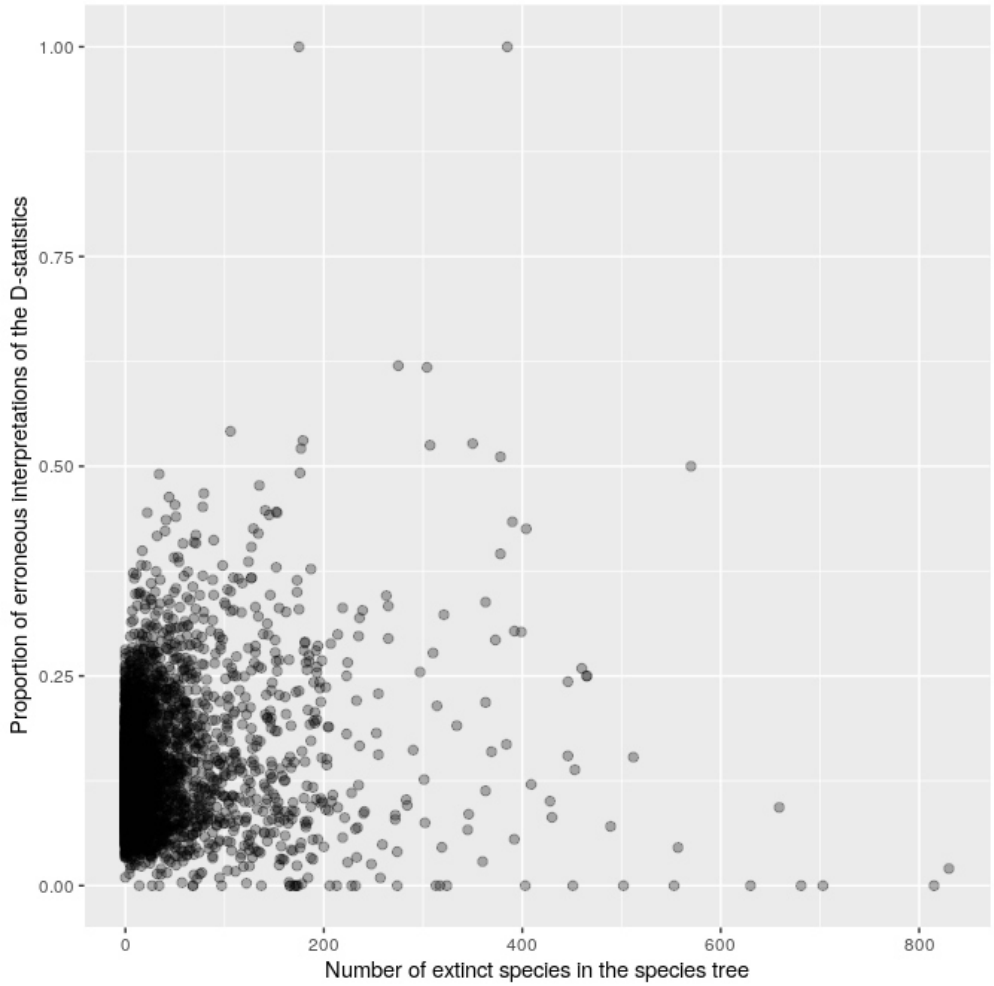
Supplementary Figure 1. Comparison of the simulation procedure we used and a more complete and more computationally demanding one: computation of the D-statistics based on simulated SNPs (x-axis) or inference based on the topology of the trees after introgression (y-axis). The proportion of erroneous interpretation over 500 introgressions for a unique species tree (N=40, α=0, extinction probability pex=0.5) was computed for each quartet (black dots). The dashed line represents the first diagonal. The squared Fisher correlation coefficient is $R^2 = 0.9618662$, providing a robust basis for using one approach as a proxy for the other

2. *All results of varying parameters*



Supplementary Figure 2. Proportion of erroneous interpretation of D-statistics for all subsets of parameters tested. Mean proportion of erroneous interpretations observed (**y-axis**) as a function of the taxonomic sampling effort (x-axis) 20 species sampled for N extent species simulated (N = 20, 40, 60, 80, 100) by the distance to the outgroup (**columns**) for different strengths of the phylogenetic distance effect (**lines**) as controlled by α (α = 0, 1, 10, 100 and 1000).

3. No effect of the number of extinct species



Supplementary Figure 2. Effect of the number of extinct species on the proportion of erroneous interpretation of the D-statistics. Proportion of erroneous interpretations observed (y-axis), function of the number of extinct lineages in the species tree (x-axis). Species trees were simulated using 4 extinction rates, pext (pext = 0, 0.3, 0.6, 0.9). For each parameter pext, 1000 species trees were simulated with 20 extant species at the end and with 100 introgressions sampled.

4. Estimating alpha on biological data

In the five following figures, five different phylogenies were taken from the literature :

(A) the bear phylogeny from Hailer et al. (2012), Figure 1A in this paper

(B) the bos phylogeny from Wu et al. (2018), Figure 1 and Supplementary Figure 13 in this paper

(C) the mosquito phylogeny from Fontaine et al. (2015), Figure 1C in this paper

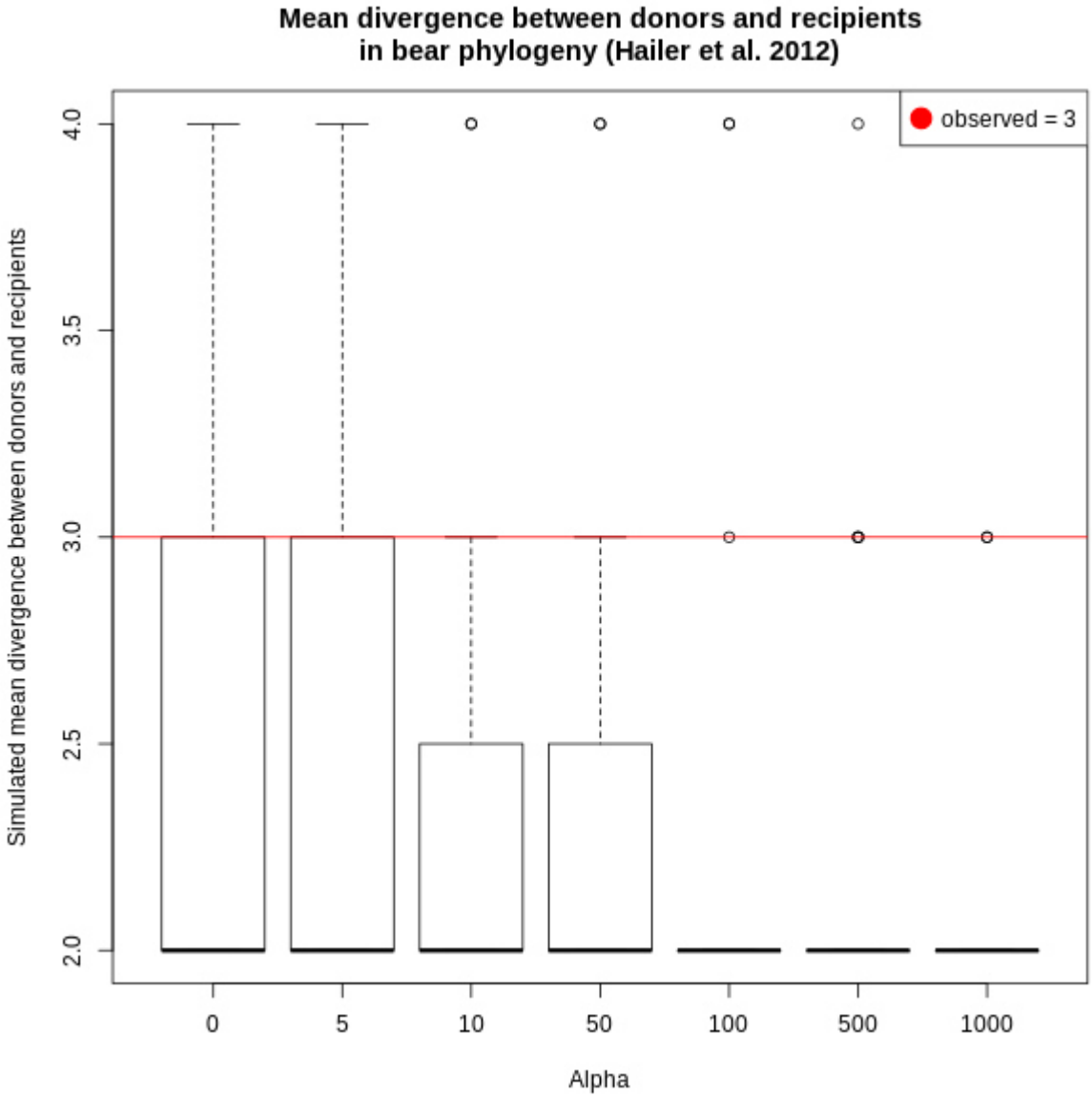
(D) the woodcreeper phylogeny from Pulido-Santacruz, Aleixo, and Weir (2020), Figure 5 in this paper

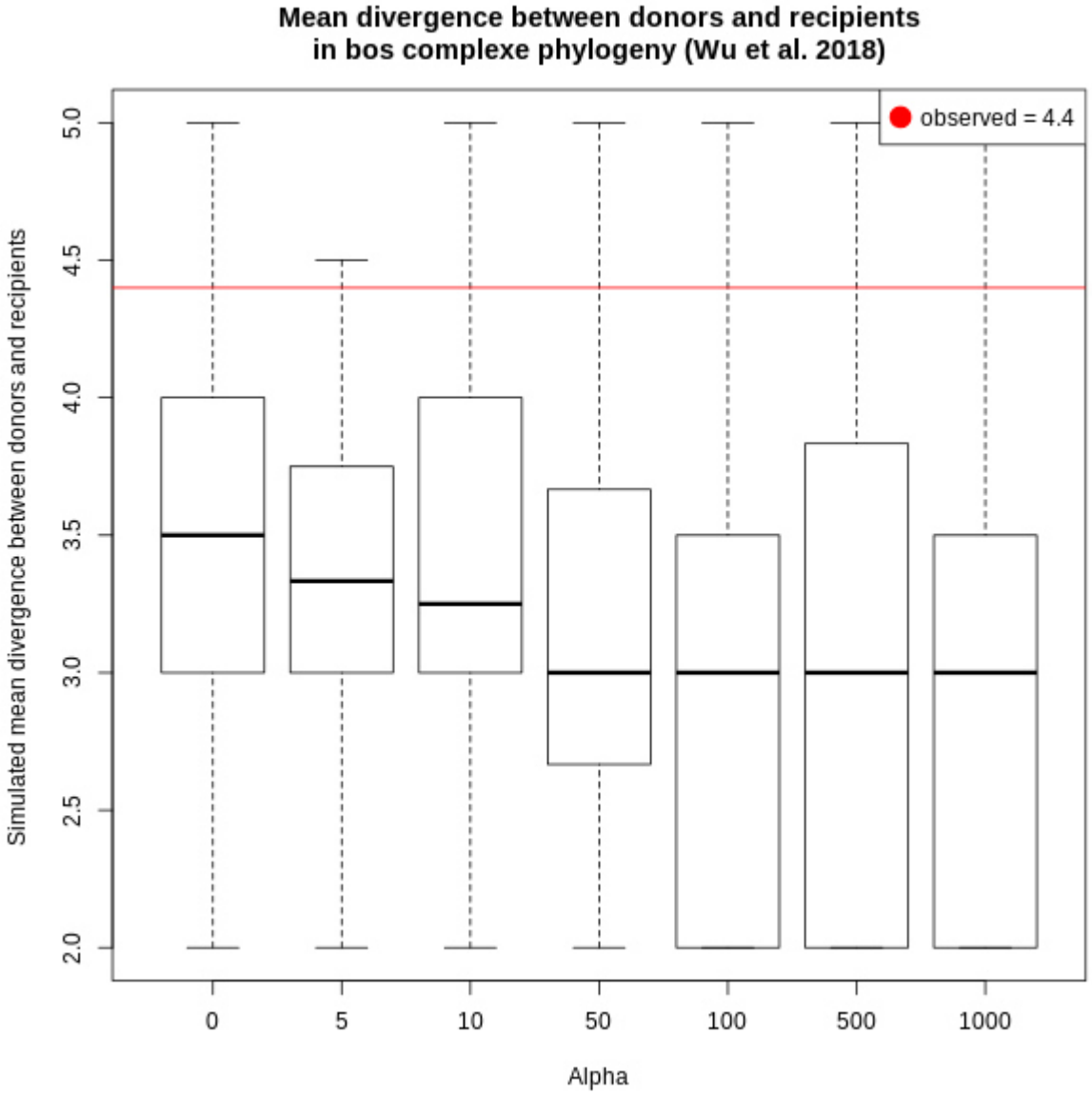
(E) the spider phylogeny from Leduc-Robert and Maddison (2018) Figure 1 in this paper

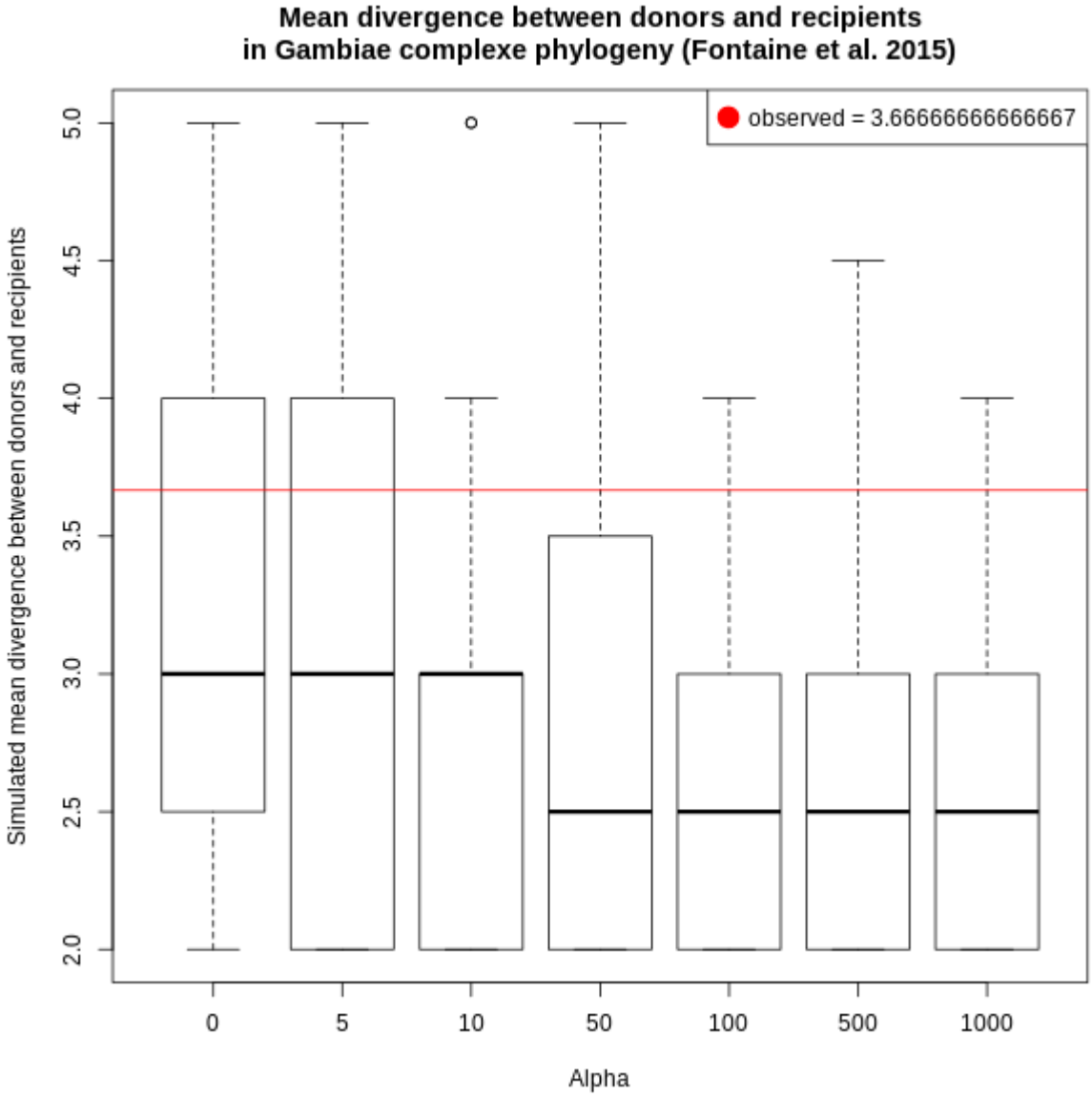
For each phylogeny, there are a certain number of introgressions documented. We simulated the same number of introgressions with alpha = {0,1,10,100,1000}. For the observed and simulated situations we computed the mean number of nodes in the phylogeny separating the donors and recipients of introgressions, excluding introgressions between sister branches in the simulations (because they cannot be detected in the biological data). This mean number is drawn on the five figures, in function of alpha, and the observed number is added on the figure by an horizontal line.

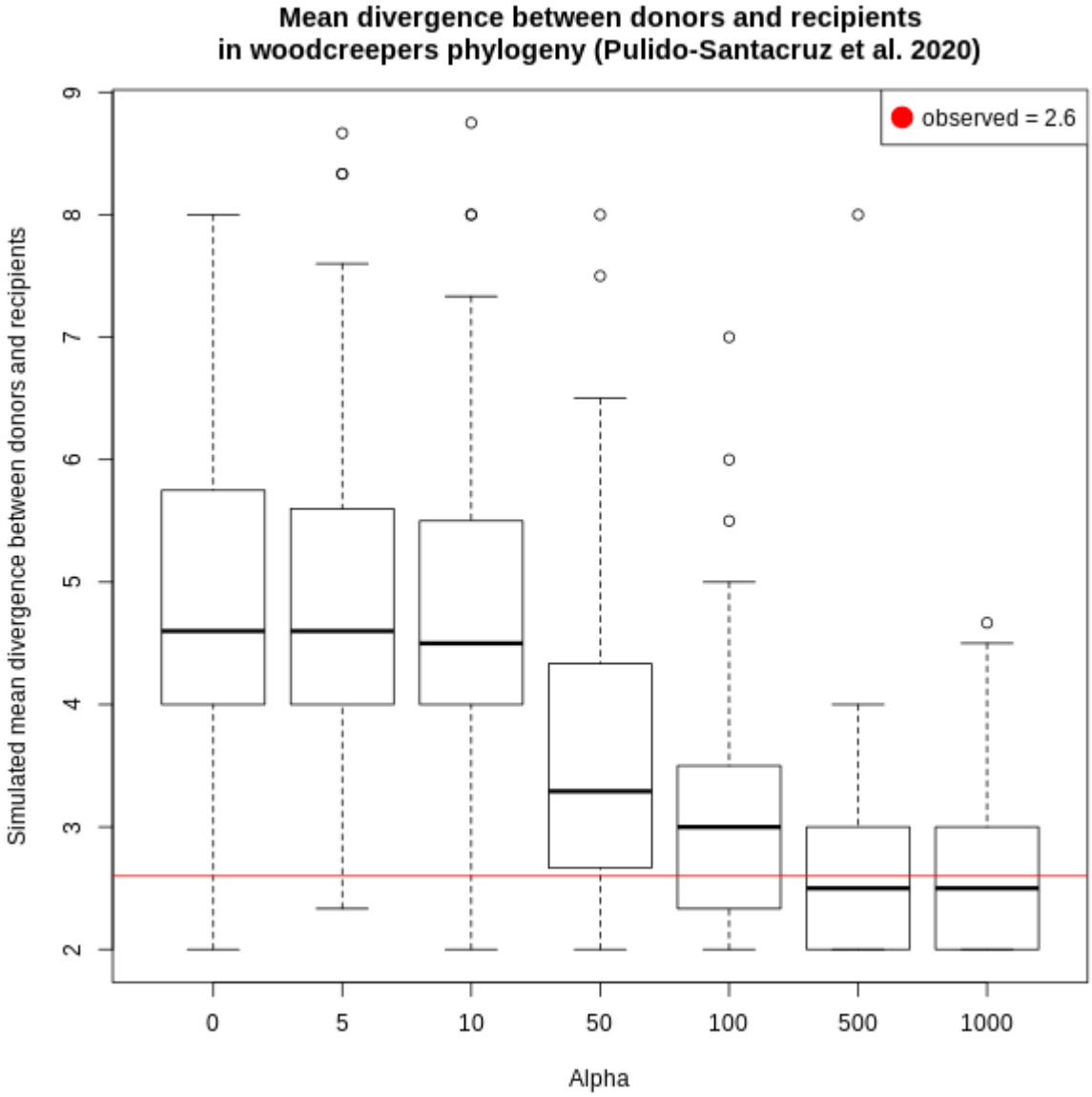
This shows ranges of alpha parameters that can best explain the observations. This range highly depends on the dataset, and shows that there is no unique possible choice of alpha. In consequence the range in which we make this parameter vary is not unrealistic, and no extremal value can be a priori prefered.

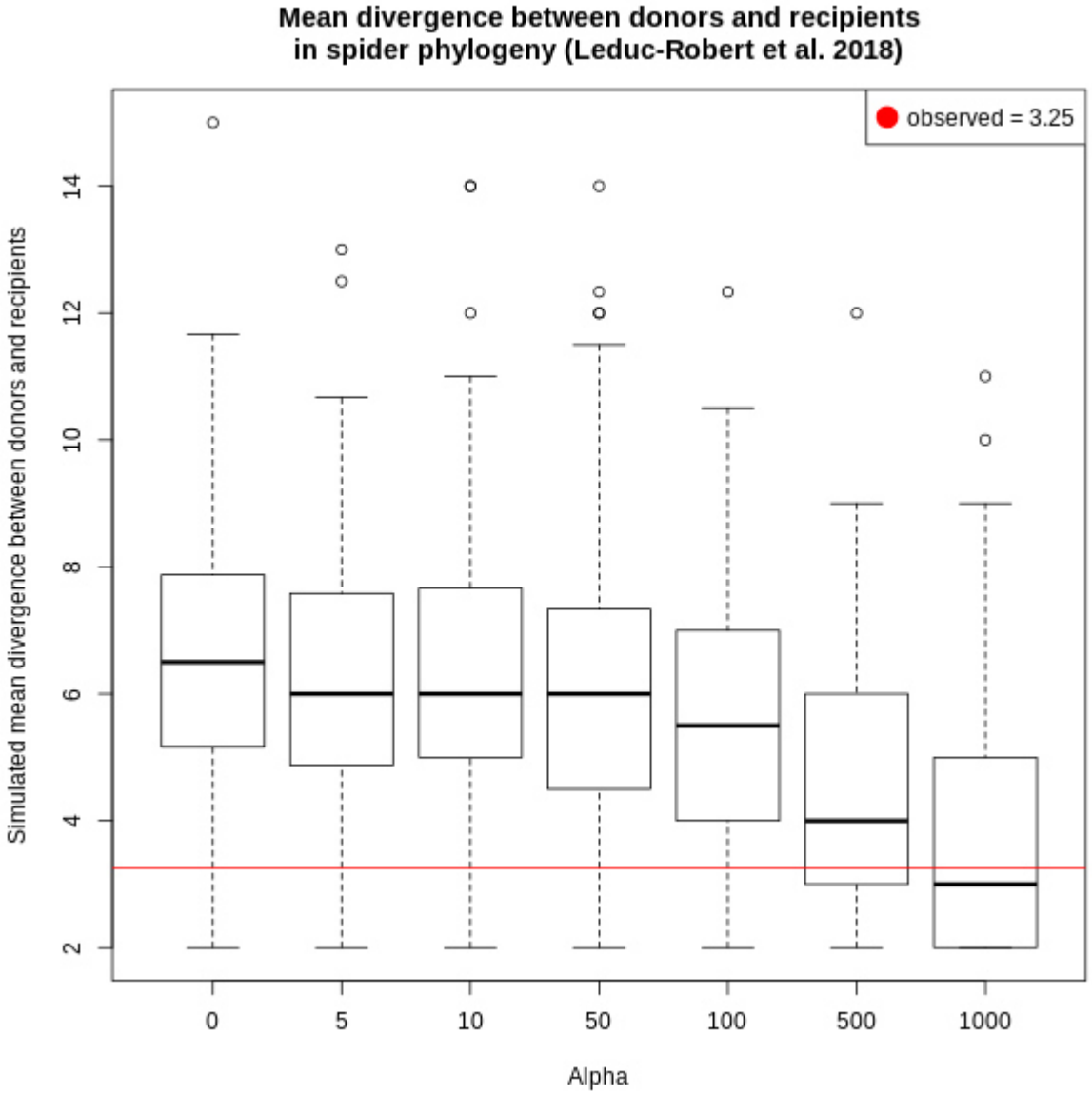
Supplementary Figures 3-7




Mean divergence between donors and recipients
in bos complexe phylogeny (Wu et al. 2018)
observed = 4.4
Simulated mean divergence between donors and recipients
2.0
2.5
3.0
3.5
4.0
4.5
5.0
0
5
10
50
100
500
1000
Alpha


Mean divergence between donors and recipients
in Gambiae complexe phylogeny (Fontaine et al. 2015)
observed = 3.66666666666667
Simulated mean divergence between donors and recipients
2.0
2.5
3.0
3.5
4.0
4.5
5.0
0
5
10
50
100
500
1000
Alpha


Mean divergence between donors and recipients
in woodcreepers phylogeny (Pulido-Santacruz et al. 2020)
observed = 2.6
Simulated mean divergence between donors and recipients
2
3
4
5
6
7
8
9
0
5
10
50
100
500
1000
Alpha


Mean divergence between donors and recipients
in spider phylogeny (Leduc-Robert et al. 2018)
observed = 3.25
Simulated mean divergence between donors and recipients
2
4
6
8
10
12
14
0
5
10
50
100
500
1000
Alpha

5. *Erroneous interpretations in the D_{FOIL} test*

In the main text we argue that just as the D-statistic, D_{FOIL} is subject to misinterpretations. We detail this statement here, and need to recall first how it is computed.

2.1. How the D_{FOIL} is computed

For a combination of 5 lineages with symetrical topology (((P1,P2),(P3,P4)),O), the 4 D-like statistics with the following equations are computed (Pease and Hahn 2015):

$$D_{FO}= \frac{(\Sigma BABAA + \Sigma BBBAA + \Sigma ABABA + \Sigma AAABA) - (\Sigma BAABA + \Sigma BBBAA + \Sigma ABBAA + \Sigma AABAA)}{(\Sigma BABAA + \Sigma BBBAA + \Sigma ABABA + \Sigma AAABA) + (\Sigma BAABA + \Sigma BBBAA + \Sigma ABBAA + \Sigma AABAA)}$$

$$D_{IL}= \frac{(\Sigma ABBAA + \Sigma BBAAA + \Sigma BAABA + \Sigma AAABA) - (\Sigma ABABA + \Sigma BBABA + \Sigma BABAA + \Sigma AABAA)}{(\Sigma ABBAA + \Sigma BBAAA + \Sigma BAABA + \Sigma AAABA) + (\Sigma ABABA + \Sigma BBABA + \Sigma BABAA + \Sigma AABAA)}$$

$$D_{FI}= \frac{(\Sigma BABAA + \Sigma BABBA + \Sigma ABABA + \Sigma ABAAA) - (\Sigma ABBAA + \Sigma ABBBA + \Sigma BAABA + \Sigma BAAAA)}{(\Sigma BABAA + \Sigma BABBA + \Sigma ABABA + \Sigma ABAAA) + (\Sigma ABBAA + \Sigma ABBBA + \Sigma BAABA + \Sigma BAAAA)} D_{OL}$$

$$= \frac{(\Sigma BAABA + \Sigma BABBA + \Sigma ABBAA + \Sigma ABAAA) - (\Sigma ABABA + \Sigma ABBBA + \Sigma BABAA + \Sigma BAAAA)}{(\Sigma BAABA + \Sigma BABBA + \Sigma ABBAA + \Sigma ABAAA) + (\Sigma ABABA + \Sigma ABBBA + \Sigma BABAA + \Sigma BAAAA)}$$

Binomial tests are performed to evaluate whether the difference between both elements framing the minus sign of each equation was significant, in order to assign a "+", "-" or "0" sign. In (Pease and Hahn 2015), 8 unique patterns of D_{FOIL} were linked to different polarized introgression events (with an explicit direction) and another 2 to pairs of non polarized (both directions) events (see Table 1. in Pease and Hahn 2015).

2.2. D_{FOIL}, a 5-taxon extension of D-statistics, rarely solves the issue raised by ghost introgressions.

If we examine the DFOIL statistic with the possibility of presence of ghost lineages, we can observe two additional D_{FOIL} patterns, "00++" and "00--" that can be interpreted as non polarized events. Furthermore, any non polarized event can be explained either by an introgression between an ancestor lineage of one clade and a species from the opposite clade or an introgression from a midgroup ghost lineage to the second species of the opposite clade. For example, the Dfoil pattern "++00" arises from the event P1P2<->P3 but could also be observed following the event Ghost->P4. For "--00", events are P1P2<->P4 and Ghost->P3. For the two new patterns, "00--" and "00++", events are P3P4<->P1 and Ghost->P2 and events are P3P4<->P2 and Ghost->P1 respectively. It should be noted that, similarly to the D-statistic, an introgression from the outgroup lineages or an external lineages to the quintet will produce the same pattern as a midgroup ghost interpretation. Given that the ancestor of P3 and P4 is always older than the ancestor of P1 and P2, this implies that a lineage with no

descendant available inside the ingroup is the donor for those two D_{FOIL} patterns, either a sister lineage to P3P4 or a midgroup ghost lineage. Conversely polarized events can not be explained by any events involving midgroup ghost lineages. This means that Dfoil can only be erroneously interpreted if the pattern is non polarized.

Fontaine, Michael C., James B. Pease, Aaron Steele, Robert M. Waterhouse, Daniel E. Neafsey, Igor V. Sharakhov, Xiaofang Jiang, et al. 2015. "Extensive Introgression in a Malaria Vector Species Complex Revealed by Phylogenomics." *Science* 347 (6217): 1258524. https://doi.org/10.1126/science.1258524.

Hailer, Frank, Verena E. Kutschera, Björn M. Hallström, Denise Klassert, Steven R. Fain, Jennifer A. Leonard, Ulfur Arnason, and Axel Janke. 2012. "Nuclear Genomic Sequences Reveal That Polar Bears Are an Old and Distinct Bear Lineage." *Science* 336 (6079): 344–47. https://doi.org/10.1126/science.1216424.

Hudson, R. R. 2002. "Generating Samples under a Wright-Fisher Neutral Model of Genetic Variation." *Bioinformatics* 18 (2): 337–38. https://doi.org/10.1093/bioinformatics/18.2.337.

Leduc-Robert, Geneviève, and Wayne P. Maddison. 2018. "Phylogeny with Introgression in Habronattus Jumping Spiders (Araneae: Salticidae)." *BMC Evolutionary Biology* 18 (1): 24. https://doi.org/10.1186/s12862-018-1137-x.

Pease, James B., and Matthew W. Hahn. 2015. "Detection and Polarization of Introgression in a Five-Taxon Phylogeny." *Systematic Biology* 64 (4): 651–62. https://doi.org/10.1093/sysbio/syv023.

Pulido-Santacruz, Paola, Alexandre Aleixo, and Jason T. Weir. 2020. "Genomic Data Reveal a Protracted Window of Introgression during the Diversification of a Neotropical Woodcreeper Radiation*." *Evolution* 74 (5): 842–58. https://doi.org/10.1111/evo.13902.

Staab, Paul R., and Dirk Metzler. 2016. "Coala: An R Framework for Coalescent Simulation." *Bioinformatics* 32 (12): 1903–4. https://doi.org/10.1093/bioinformatics/btw098.

Wu, Dong-Dong, Xiang-Dong Ding, Sheng Wang, Jan M. Wójcik, Yi Zhang, Małgorzata Tokarska, Yan Li, et al. 2018. "Pervasive Introgression Facilitated Domestication and Adaptation in the Bos Species Complex." *Nature Ecology & Evolution* 2 (7): 1139–45. https://doi.org/10.1038/s41559-018-0562-y.

4

Reconnaître l'existence de lignées fantômes inverse les résultats des méthodes de détections utilisant les longueurs de branches des arbres de gènes

Les événements de flux de gènes sont courants à toutes les échelles du monde vivant. Ces évènements peuvent être détectés avec des méthodes phylogénétiques notamment, en comparant l'histoire évolutive des fragments échangés à l'histoire évolutive des espèces ou des gènes n'ayant pas été transférés.

Dans ce manuscrit sont présentées trois réévaluations, avec des simulations, de méthodes basées sur l'utilisation des longueurs de branches pour détecter des flux de gènes et résoudre l'histoire évolutive des espèces considérées. On montre que détecter des flux de gènes avec ces méthodes en prenant en compte les espèces fantômes peut inverser les conclusions tirées dans ces différentes méthodes :

1. Le D3, une variante du test ABBA-BABA, utilise des arbres de trois taxa et les longueurs de branches pour détecter des événements d'introgression. Cependant, de la même manière que le test ABBA-BABA, celui-ci ne prend

pas en compte l'impact des espèces fantômes sur son interprétation.

2. Pour résoudre la phylogénie d'un complexe d'espèces gambiae, Fontaine et al (2015) on utilisé l'hypothèse suivante : si les flux de gènes diminuent la distance évolutive et les longueurs de branches dans les arbres de gènes, alors les gènes qui présentent les longueurs de branches les plus importantes n'ont probablement pas subi d'échanges et devraient donc représenter l'histoire évolutive des espèces. La phylogénie qu'ils trouvent implique que la majorité des gènes sont introgressés. Prendre en compte les espèces fantômes permettrait d'inférer une phylogénie alternative et plus conforme à l'histoire des gènes.
3. La mitochondrie et le chloroplaste sont deux organelles qui ont été acquises par endosymbiose. Les mitochondries, ubiquitaires chez les eucaryotes, dériveraient d'un endosymbionte Alphaproteobacterien. Les chloroplastes, présents chez certaines espèces de plantes et algues, proviendraient de l'endosymbiose d'une cyanobactérie par les cellules eucaryotes. Dans l'espoir de clarifier la chronologie d'acquisition de ces deux organelles, qui est encore source de débat, plusieurs méthodes utilisant les longueurs de branches ont été développées avec pour objectif d'ordonner les multiples événements d'acquisition des différents gènes les composant. L'hypothèse sous-jacente de ces méthodes est la suivante : si deux gènes ont été transférés horizontalement dans le même receveur, l'arbre de gènes présentant les longueurs de branches les plus courtes indiquerait que ce gène aurait été acquis plus récemment que le second gène présentant lui des longueurs de branches plus importantes. On montre que les longueurs de branches ne permettent pas de donner une chronologie aux évènements d'acquisition si les espèces fantômes sont considérées.

Recognizing the existence of ghost lineages reverses the results of evolutionary studies on genetic transfers

Théo Tricou[1], Eric Tannier[1,2] and Damien M. de Vienne[1,*]

[1] Univ Lyon, Université Lyon 1, CNRS, Laboratoire de Biométrie et Biologie Évolutive UMR5558, F-69622 Villeurbanne, France
[2] INRIA Grenoble Rhône-Alpes, F-38334 Montbonnot, France

***Corresponding authors: E-mail:** damien.de-vienne@univ-lyon1.fr

Abstract: A common expectation is that introgressions, transfers, gene flow, or any kind of horizontal transfer of genetic material across taxa or populations results in phylogenies with shorter branches for the lineages involved. Following this prediction, authors proposed to use phylogenetic branch lengths as a tool to identify introgressed genomic regions, to choose good markers to reconstruct correct species trees (*i.e.* gene trees with longer branch lengths) or to relatively order acquisitions of genes during major transitions (eukaryogenesis or origin of chloroplasts). However, the use of branch length in these contexts is misleading because it overlooks the important impact of ghost lineages on the theoretical expectations. We show that when considering that extinct, unknown and unsampled taxa are predominant, which is far from unrealistic, conclusions of studies using branch lengths can be the exact opposite.

Introduction

Genomic flow across taxa appears as an important evolutionary force, affecting all domains of Life, and occurring at all evolutionary time scales: from introgression between populations to trans-phylum endosymbiosis. Detection of these events, hereafter referred to as Horizontal Genomic Fluxes (HGF) can be performed with phylogenetic approaches. The principle is that phylogenies reconstructed from portions of the genomes that were transferred (from single sites to full chromosomes) may contradict topologically the phylogeny of the taxa analyzed (the species tree), and/or show shorter branch lengths (on average and in lineages involved in the

transfer). These simple expectations are at the basis of myriad studies and methods on HGF (Adato et al. 2015; Fontaine et al. 2015; Pittis and Gabaldón 2016; Rosenzweig et al. 2016; Dalquen et al. 2017; Hahn and Hibbins 2019; Hibbins and Matthew W Hahn 2019; Pfeifer and Kapan 2019; Forsythe et al. 2020a, 2020b; Vosseberg et al. 2020; Susko et al. 2021; Suvorov et al. 2021).

We focus here specifically on the link between HGF and branch lengths. We show that the way branch lengths are interpreted in studies dealing with genomic fluxes overlooks the impact that ghost lineages can have on these expectations. By ghost lineages, we mean any taxa that is absent from the analysis, *i.e.* the extant taxa that are simply excluded, the extant taxa that are still unknown, and all taxa that are extinct. While the first category is a choice and can be taken into consideration in the interpretation of the results obtained, the two others are not. More than 99.9% of all species that ever lived are now extinct (Raup 1991), and the number of extant species still uncatalogued is almost an order of magnitude higher than reported ones (around 1.3 over 8.7 Million estimated in 2011, Mora et al. 2011), and many orders of magnitudes higher if considering microbial species (Locey and Lennon 2016).

We reanalyzed three different studies that used the expected link between HGF and branch length in three different contexts: for the detection of introgressed loci (the D3 method Hahn and Hibbins 2019), for resolving the correct branching order of Anopheles species (Fontaine et al. 2015) and for timing the acquisitions of genes associated with the emergence of eukaryotic cells (Pittis and Gabaldón 2016; Vosseberg et al. 2020) and those at the origin of the chloroplast membrane (Sato 2020). With simple simulations, we show evidence that under the very weak assumption that many taxa are invisible, using the argument of branch length to draw conclusions involving HGF is misleading, and can even lead to conclusions that are the opposite of those drawn initially. We advocate for a better consideration of ghost lineages in evolutionary studies and propose an alternative hypothesis: any signal of horizontal genomic flux should be interpreted at first as coming from a ghost.

Results

Preamble

The three examples that we explore in this study (the three next sections) illustrate the same principle regarding the overlooked impact of ghost lineages on branch length after HGF: instead of decreasing branch lengths, transfers of genetic material may increase it in some cases when ghost lineages are the source of the transfers. This is illustrated in Figure 1. While HGF between species present in the analysis produce gene trees with shorter branch lengths (scenario A and B), HGF from the ghost lineage "X" produces a gene tree with longer branches (scenario C). As a consequence, considering phylogenies with smaller branches as the result of genomic fluxes can in fact be the opposite of what really occurred: phylogenies with shorter branches might stem in reality from the absence of introgression.

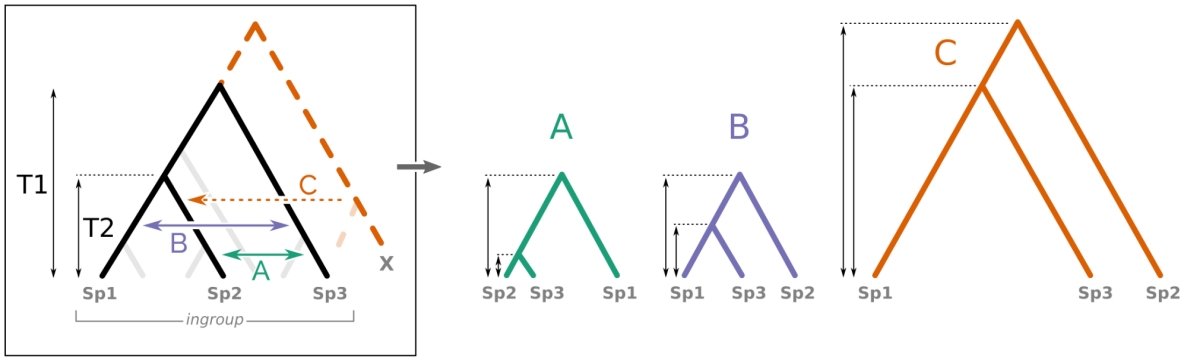


Figure 1. Effect of some horizontal genomic fluxes (HGF) on the branch lengths of phylogenetic trees of the horizontally transferred genomic regions. HGF from a ghost lineage (X) exterior to the ingroup containing the three lineages considered (scenario C) produces a phylogenetic tree with increased branch lengths as compared to the species tree, while HGF within the three lineages of interest (scenarios A and B) do the opposite.

Using branch lengths to detect introgression events (the D3 method) is often misleading

D_3 (Hahn and Hibbins 2019) is a recently published test that proposes to use branch lengths (*i.e.* pairwise distances between species) to detect introgressions in a three-taxa tree. Referring to the notations in Figure 1, the test is supposed to detect gene flows between sp2 and sp3 or sp1 and sp3 (cases A and B, respectively), by computing the statistics:

$$D_3 = \frac{d_{sp2-sp3} - d_{sp1-sp3}}{d_{sp2-sp3} + d_{sp1-sp3}} \quad (1)$$

According to the original description of the test, if no introgression occurs, D_3 is equal to 0 regardless of the presence of Incomplete Lineage Sorting (ILS), but in the case of gene flow, D_3

can be either significantly positive, revealing introgression between sp1 and sp3, or significantly negative, revealing introgression between sp2 and sp3.

So the D_3 method (and others Adato et al. 2015; Rosenzweig et al. 2016; Hahn and Hibbins 2019; Forsythe et al. 2020; Pfeifer et al. 2020; Suvorov et al. 2021, based on the same principle) rely on the assumption that short branch lengths will be observed following introgression events. But what if the introgression occurs between a taxa outside the tree (such as X in Figure 1) and the sp2 taxa for example (see case C in Figure 1)? Such gene flow increases the distance between sp2 and sp3 without affecting the distance between sp1 and sp3. This results in a significantly positive D_3 that is interpreted as gene flow between sp1 and sp3 while none of these taxa are in fact involved in the introgression. While it has been shown several times that the D-statistic and other variants can be deceived by ghost lineages (Durand et al. 2011; Hibbins and Hahn 2021), it is an issue for which the scale is still unknown (but see Tricou et al. 2021).

In practice, the chance that ghost introgressions produce erroneously interpreted D_3 statistics can be estimated. To do so we simulated random species trees and random introgressions thererin, with the software *ms* (see Material and Methods), and for all possible samples of three-taxa where D_3 was significantly different from 0, we evaluated whether this result was imputable to (i) an introgression within the group containing the three taxa of interest (the ingroup), even if the donor taxa was not the one directly involved (*i.e.* a sister donor Eaton and Ree 2013), or (ii) to a "ghost introgression" from outside the ingroup, resulting in erroneous interpretation of the test. Unsurprisingly, the probability of erroneous interpretation of the D_3 statistics was high when the size of the ingroup was small relative to the total size of the tree, almost reaching 100% (Figure 2). Biologically speaking, considering that there are more taxa outside a given clade than inside does not seem unrealistic, which allows us to propose that any D_3 statistics significantly different from 0 should be interpreted at first as the result of an introgression from outside the tree formed by the three taxa considered.

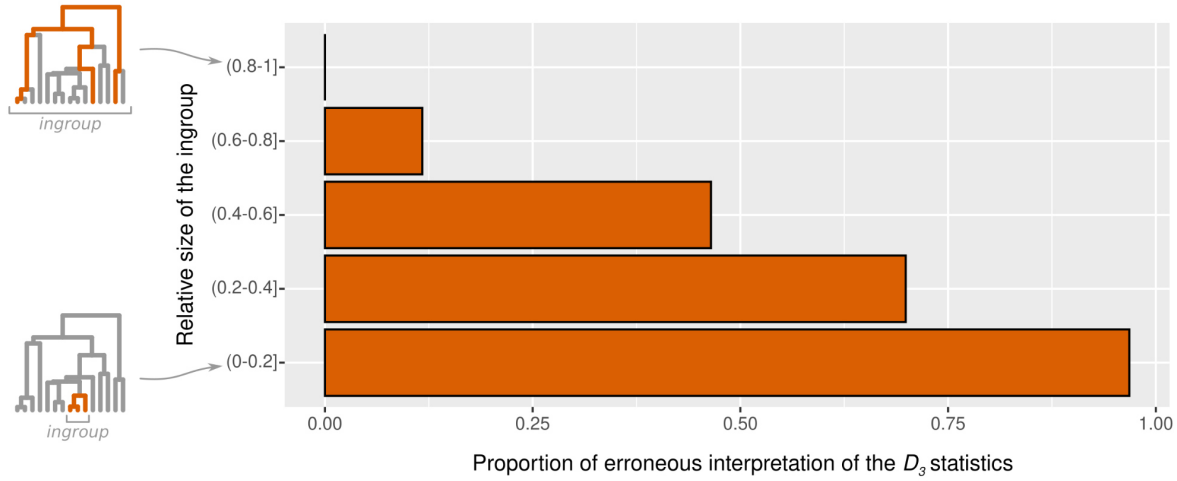


Figure 2. Proportion of cases (in simulations) where a D_3 test is significantly different from 0 because of an introgression from a ghost lineage situated outside the ingroup. These cases might be incorrectly interpreted (see text) and lead to the misidentification of both taxa involved in the introgression. The proportion of erroneous interpretation of the test is computed for different relative sizes of the ingroup, i.e. the size of the smallest subtree containing the three taxa studied divided by the total number of species in the tree (see sketches on the left side).

Using branch lengths to resolve the correct species branching order in mosquitoes is misleading in the presence of ghost lineages

It is a common practice in phylogenetics to use gene markers known to not be horizontally transferred in order to reconstruct the “true” species phylogeny. This is meaningful, because HGF can change the topology of the gene phylogenies and should thus be discarded. Following this, if one considers that HGF decreases branch lengths in gene phylogenies, it seems like a straightforward expectation that among multiple possibilities of species tree topologies, the one supported by the genes with the largest branch lengths will be the correct one. These genes are those that are expected to not have experienced genomic fluxes and to thus be concordant with the "true" species tree topology.

This approach was proposed and used to recover the "correct" branching order of three Anopheles species (*An. arabiensis, An. gambiae* and *An. coluzzii* in Fontaine et al. 2015), *i.e.* the one supported by the genes with the largest branches. As explained in the preamble, however, the hypotheses behind this approach (in terms of branch lengths) may well be violated because in the presence of ghost lineages, long branches may well be the result of introgression and not an indication of the absence of it.

To test the effect that ghost lineages may have on the strategy proposed in (Fontaine et al. 2015), we simulated a species tree from which three species were chosen to mimic the three Anopheles species used in the original study (Figure 3, top). Using *ms*, we simulated data (gene trees) under two simple evolutionary scenarios (see Material and Methods): one involving an introgression between ingroup species B and C (scenario 1), and one involving an introgression from a ghost lineage outside the ingroup to the species B (scenario 2, Figure 3). We separated the obtained gene trees in two categories, those with a topology similar to the species tree and those with a discordant topology. We then computed for each category the mean divergence times (T1 and T2) in the trees (see Figure 3).

We observed that under the scenario not involving a ghost lineage (scenario 1), divergence times in gene trees with discordant topologies (as compared to the species tree) were on average smaller than in gene trees with topologies similar to the species tree (Figure 3). This is in accordance with the expectations of the test as formulated in the original study: "Because introgression will reduce sequence divergence between the species exchanging genes, we expect that the correct species branching order revealed by gene trees constructed from non introgressed sequences will show deeper divergences than those constructed from introgressed sequence" (Fontaine et al. 2015).

However, when the introgression comes from a ghost lineage (such as in scenario 2), we observe opposite results, with gene trees with discordant topologies (as compared to the species tree) exhibiting deeper divergence than gene trees supporting the species tree topology. In this case, considering that the "true" species tree topology is the one supported by the genes with deeper divergence times will result in an erroneous conclusion.

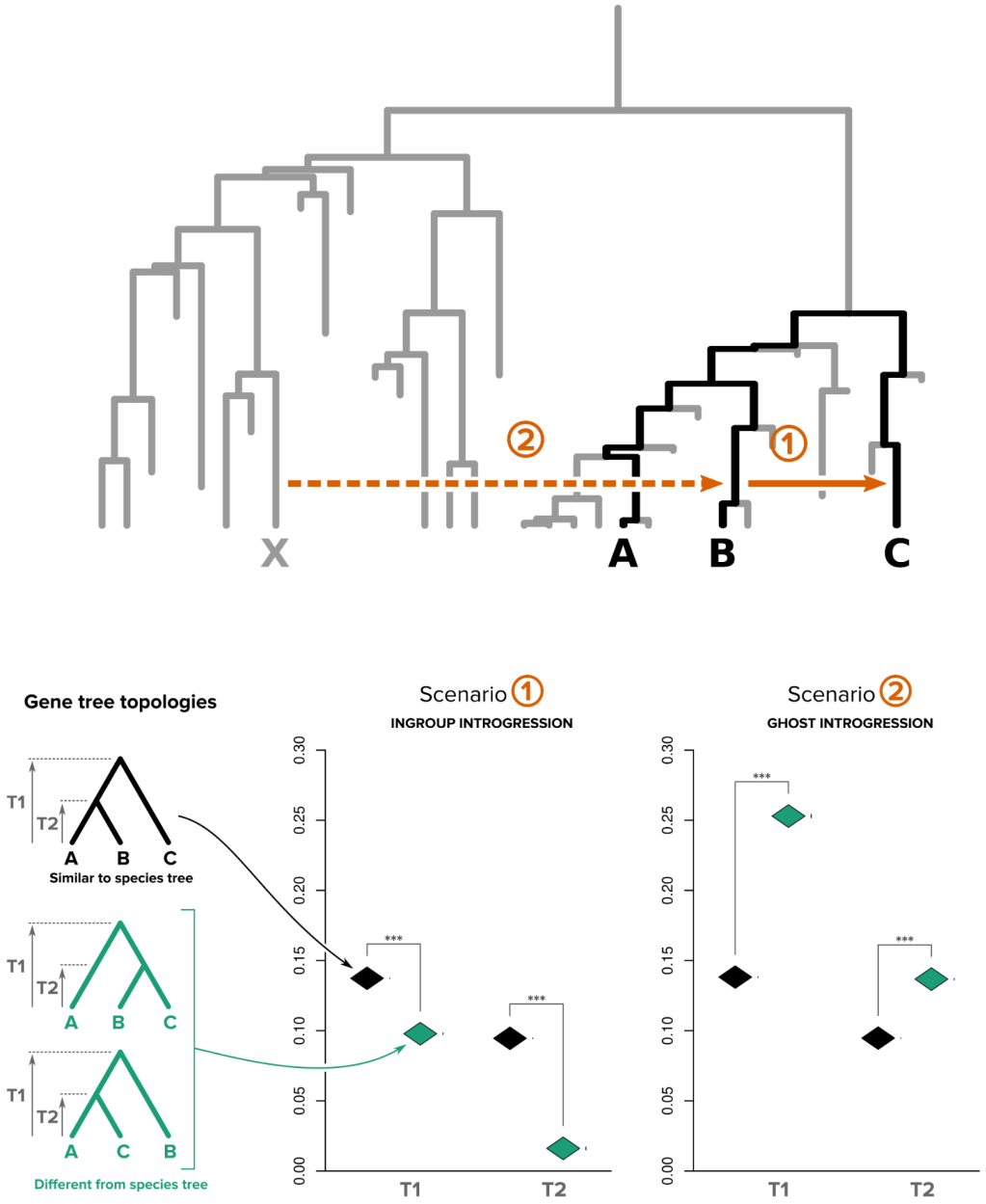


Figure 3. Impact of ghost lineages on the use of branch lengths to choose between alternative species topologies. Top panel: complete species tree, black branches represent the ABC tree while grey branches represent ghost lineages. The evolution of genomes (1000 genes) is simulated in the tree with two different introgressions: scenario 1, the ingroup introgression (solid orange arrow), between B and C and scenario 2, the ghost introgression (dotted orange arrow), between a ghost lineage X and B. Bottom panel. Left: three possible topologies for the 3 selected groups (A,B and C). Right: Mean divergence times T1 and T2 are computed for both scenarios and for all genes supporting either the species tree topologie "ABC" or the two discordantes topologies "BCA" and "ACB" (whiskers standard error of the mean, *** for t-test with P-value < 2.2e-16).

Using branch lengths to order acquisition events (the stem-length method) is misleading in the presence of ghost lineages

In recent papers, a method termed "stem-length" was proposed to help ordering different acquisitions of genes during eukaryogenesis (Pittis and Gabaldón 2016; Vosseberg et al. 2020) and was also applied to better characterize gene acquisitions and their relative timing at the origin of the chloroplast membrane (Sato 2020). This stem-length method, depicted in Figure 4, relies on the expectation that early acquisitions of genes should result in long branches (or stems) at the base of the receiving lineage in the phylogenetic trees of the transferred genes, while late acquisition should lead to shorter stems (Figure 4, top). This approach was used to address the long-standing question of the early or late acquisition of mitochondria during eukaryogenesis (Pittis and Gabaldón 2016), concluding that shorter stems in phylogenetic trees of eukaryotic genes with alphaproteobacterial origin was supporting the latter.

However, if ghost lineages are considered, the expectations of the stem-length method can be totally reversed. In the case where the donor lineage has no descendants, either because they all went extinct or because they have not been discovered yet, the stem lengths will not be determined by the time where the transfer occurred, but by the time of the divergence between the missing (ghost) clade from which the transfer originated and its closest non-missing relative (Figure 4, bottom). Under these circumstances, the correlation between the order of acquisitions and the stem-length can easily be lost.

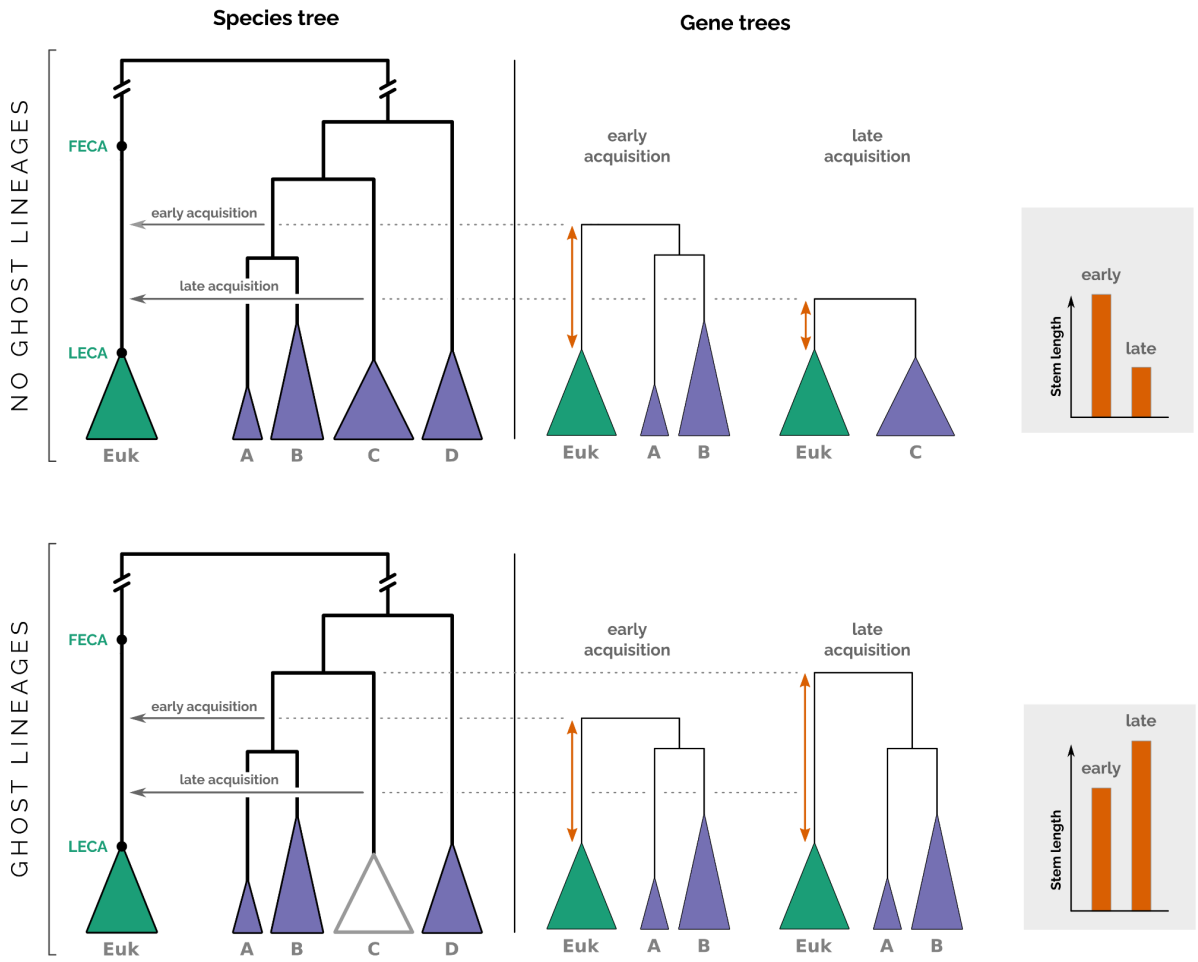


Figure 4. Illustration of the impact of ghost lineages on the use of branch length (the stem-length method) to time acquisitions. When all clades are available, i.e. there are no ghost lineages (top), early acquisitions produce gene trees with long stems (orange arrows), while late acquisitions produce gene trees with short stems, as reported on the grey box on the right. If some clade is missing, i.e. there are ghost lineages (bottom), the opposite observation can be made. Because the donor lineage (here the ancestor of C) left no descendant, and because it splitted with the rest of the clades before the time of the early acquisition, the stem lengths are inverted: early acquisition leads to shorter stem lengths than late acquisition (grey box).

To quantify this, we performed a simple simulation where species trees with 1000 leaves were generated following a birth-death process and pairs of acquisition events were repeatedly sampled in the tree to mimic early and late acquisitions. We then sampled a variable proportion of the extant species (between 10% and 1% of the total number) and looked at the effect of this sampling on the correct or erroneous prediction of the order of the events if using the stem-length method (Figure 5). We observed that when 10% of the species were considered, already ~33.5%

of the predictions were wrong (would predict that event A occured before event B while the opposite happened). This proportion reached almost 50% (the maximum possible, equivalent to random prediction) when 1% of the species were sampled (Figure 5A).

We also observed that the error of the stem-length method increased when the time interval between the two events is shorter. Indeed, from the ~33.5% of wrong predictions with 10% of species sampled, we reached nearly ~41% of wrong predictions with the same proportion of species sampled, if only looking at events whose interval was less than 10% of the tree height (Figure 5B).

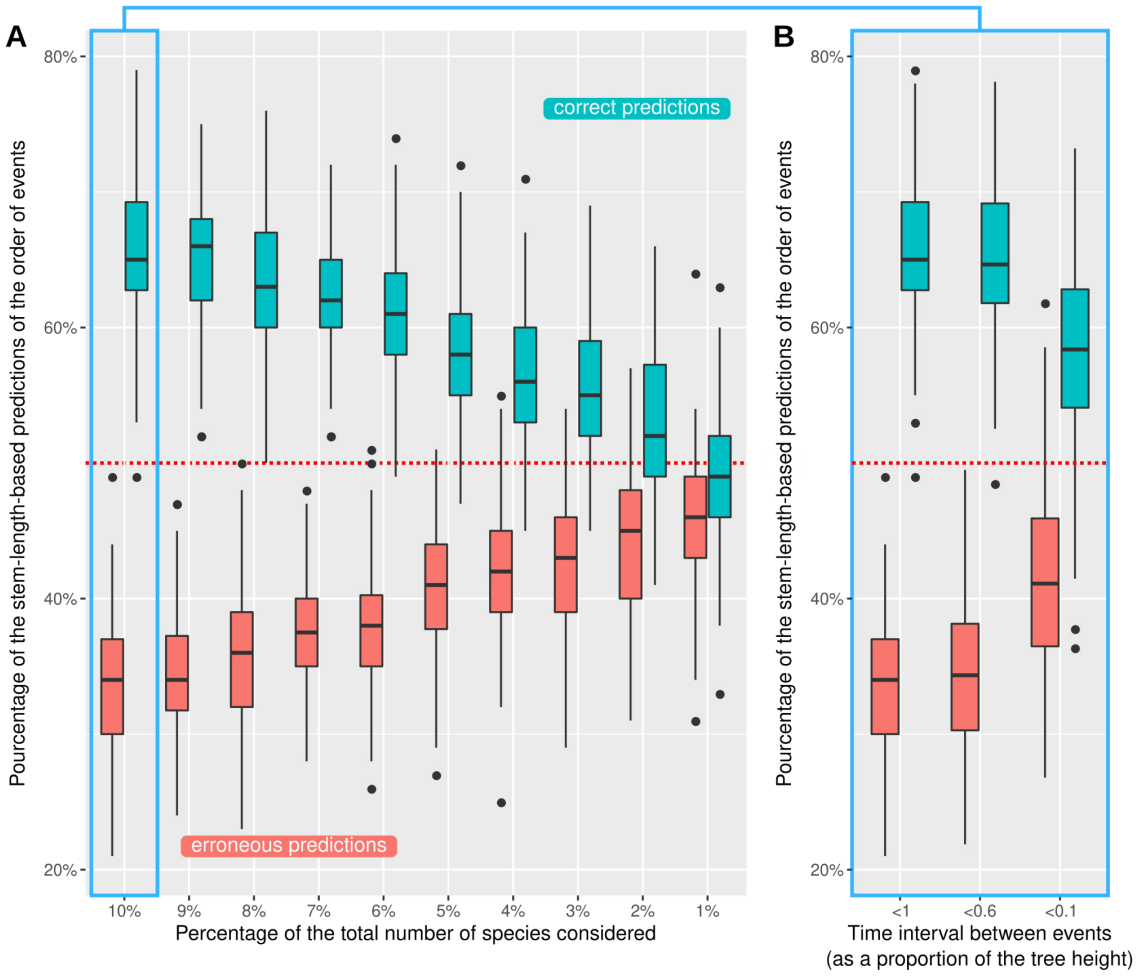


Figure 5. Effects of species sampling (**A**) and time interval between acquisition events (**B**) on the percentage of correct (blue bars) and erroneous (red bars) predictions of the order of events when using the stem-length method. The red dashed line represents what would be observed if predictions were at random. Cases where the stem-length method cannot order the events (because stems have the exact same length) are removed, which explains why blue and red bars do not always sum to 100%.

The simulations performed here are simple, but are illustrative of the impact that ghost lineages can have on approaches like the stem-length method for ordering events. Being more biologically realistic would require, for any clade on which the stem-length method is to be applied, to be able to answer the following questions: *how likely is it that none of the descendants of lineages from which transfers originated are observed today?* and *How far apart did the events considered occurred, relative to the total time span considered?*

If referring to the biological data for which the method was devised and on which it was applied originally (Pittis and Gabaldón 2016; Vosseberg et al. 2020), we can provide the following elements. The acquisitions analysed are genes of bacterial origin transferred to protoeukaryotes before the Last Eukaryotic Common Ancestor (LECA), that is between ~1.1-1.2 and ~2.3-2.7 Billion years ago (reviewed in Chernikova et al. 2011), and after the so-called FECA (First Eukaryotic Common Ancestor). It is hard to infer the macroevolutionary history of bacteria, especially at such deep evolutionary timescales, but it is clear that of the lineages living at this period, most went extinct (Louca et al. 2018), and some of them certainly during mass extinctions (Weinbauer and Rassoulzadegan 2007). It appears also clear that most extant bacterial lineages are still unknown (Locey and Lennon 2016; Louca et al. 2018), and that they may not be scattered uniformly across the known diversity. Indeed, unsampled lineages are likely to form major clades, as illustrated by the discovery of a complete new phyla (CPR) back in 2016 (Hug et al. 2016).

Considering all these elements, the possibility of erroneous ⁓or at least random⁓ prediction of the order of events on such data does not seem unlikely. This questions the use of such methods in cases where only a small fraction of all the diversity (extinct and extant) is known, which may well be the rule more than the exception in biological studies.

Discussion

Branch length approaches are versatile methods allowing to study different aspects of horizontal genomic fluxes at different scales, from intraspecies introgression (between populations) to trans-phylum gene acquisitions. These approaches all rely on the expectation that any Horizontal Genomic Flux (HGF) should result in a shorter (phylo)genetic distance between the donor and the recipient entities when looking at the transferred genomic sequence(s) than when looking at other (presumably vertically transferred) sequences for the same entities.

However, this only holds if all lineages, or at least sister lineages of all involved lineages, are present in the study. Indeed, and as demonstrated here, an apparent decrease in the (phylo)genetic distance between two entities can in fact result from an increase of the (phylo)genetic distance between other (non-considered) ones. It appears therefore important to take into account ghost lineages when interpreting the result of branch-length-based HGF inference methods.

Our simulations demonstrate that under the weak assumption that many lineages are ghost, not only does the possibility of an HGF coming from unknown species increases, but also the possibility that the identification of both the donor and the recipient of the transfer is erroneous, leading to the identification of two lineages that have nothing to do with the lineages truly involved. When interpreting the results of a D3 test, the possibility of introgression from ghost lineages should be systematically taken into consideration as an alternative possible scenario and should be considered (at least) as being as probable as the usual interpretation. Similarly, we show that HGF from a ghost lineage could in some cases increase branch length in gene trees, instead of decreasing it as is commonly expected. This absence of an unambiguous pattern should prevent the usage of branch length to identify appropriate markers for phylogenetic reconstruction. Finally, because HGF can increase and decrease branch length depending on the sampling effort and the amount of ghost lineages, using it to relatively time transfer events needs precaution.

Here, we highlight the issue that most HGF-detection results are examined without considering ghost lineages. Failure to consider the impact of ghost lineages on the interpretation of such tests could impair their usefulness and change conclusions that we draw from them. As such, we suggest that the interpretation of the results of gene-flow-related methods should be done with ghost lineages as the topmost hypothesis.

Material and Methods

Effect of ghost lineages on the D3 test for introgression

Simulating species trees and introgressions

We simulated 200 random species trees with a birth-death model using the tool *Zombi* (Davín et al. 2020). Speciation rate and extinction rate were respectively fixed to 1 and 0.9. Simulations stopped when 40 extant species were reached. With a custom python script, we converted species

tree topology in a suitable format to use in the coalescent simulator *ms* (Hudson 2002). Branch lengths were converted in units of generation and fixed the age of the root of the trees to 10^6 generations. To simulate the introgression, a single event of migration was imposed over 1 generation for a fraction **f**= 50% of the donor population to invade the recipient population. This migration rate was used to ensure that introgression detection using D3 would not be biased by false positives. Then, for each species tree we used *ms*, implemented in the R package *Coala* (Staab and Metzler 2016), to simulate 1000 gene trees, evolving in populations of fixed size (Ne) of 100,000 individuals.

Computing the D3

To compute the D3 for a trio of lineages with topology ((P1,P2),P3) in a species tree, we compute D_{13} and D_{23}, respectively the sum of the distance (or branch length) separating P1 and P3 and the sum of the distance separating P2 and P3 across all 1000 gene trees. We then computed the D3 following equation (1). This was done for each trio of lineages with topology ((P1,P2),P3) in each species tree simulated. The significance of D3 was tested by bootstrap resampling of 1000 gene trees with 1000 replicates. We then calculate the Z-score and consider a D3 significant if Z>3 or Z<3, following Green et al (2010). Finally, as we tracked for each simulation the true donor and recipient lineages, we assessed for each given D3 if the test was significantly due to an ingroup introgression or a ghost introgression.

Using branch lengths to determine the correct species tree topology

Species tree simulation with ingroup introgression or ghost introgression

We used *Zombi* to simulate a species tree. If not stated otherwise, parameters were the same as the one used in the previous section (see section 1.1). The simulation stopped when 16 extant species were reached. On this tree, three species with the topology ((A,B)C) were selected similarly to what is used in (Fontaine et al. 2015). From this point, all other lineages were considered as ghost lineages. We then converted the topology of the tree in a *ms* readable tree, to use for coalescent simulation. Using ms, two dataset were generated, in both the number of generation separating the tip from the root of the tree was fixed to $5*10^6$ generations and an event of migration, for a fraction ***f***= 20%. For the first dataset the migration took place from B to C, between two extant species (ingroup introgression). For the second the migration took place

from a random ghost lineage outside the triplet phylogeny and B (ghost introgression). Finally for both models, we simulated 1000 gene trees, evolving in populations of fixed size (Ne) of 100,000 individuals.

Branch length in gene trees.

For all gene trees, we determined in which topology the three species A, B and C were, either ((AB)C) (also representing the species tree branching order) or one of the two discordant topologies that arise from ILS or introgression, ((BC)A) or ((AC)B). Subsequently, for both models and for each gene tree, the value of species divergence times T1 and T2 were computed (but see Figure 3) following the equation from Fontaine *et al.* 2015 (see supplementary material S3.2 from their paper).

Timing the acquisition of genes in presence of ghost lineages: the stem-length method

To compare simulated *versus* predicted order of events using the stem-length method (Pittis and Gabaldón 2016; Vosseberg et al. 2020) in the presence of ghost lineages, we performed the following simulations:

- Generate 100 trees under a birth-death process (speciation rate = 1, extinction rate = 0.5) using the *rphylo* function in the R package *ape* (Paradis and Schliep 2019), stopping the simulations when tree reached 1000 leaves.
- Randomly sample two points in each tree representing two origins of transfers (or acquisitions) and record their timing and the time interval ***dt*** between them (as a fraction of the total tree height).
- Sample a proportion ***p*** of the leaves in each tree.
- Evaluate the new order of the events if using the stem-length method on the pruned tree.
- Record whether the order of events before and after sampling agree (1), disagree (0) or are indistinguishable (NA). The latter case occurs if the stem-lengths after sampling are equal for the two events.
- Compute the proportion of each case out of the 100 replicates.

Values of ***p*** were chosen between 10% (100 leaves) and 1% (10 leaves) of the total number of leaves, every 1%, and the whole process was repeated 100 times to obtain a variance around the observed proportions of correct and erroneous predictions on the order of the events.

To explore the effect of the time interval (***dt***) between the events on the proportion of erroneous predictions, we subsampled the pairs of events with ***dt***<1, ***dt***<0.6 and ***dt***<0.1 and recomputed the proportion of correct and erroneous predictions each time. This analysis was restricted to the case where ***p***=10%.

Acknowledgments

We thank Adrián A. Davín for sharing his insight on simulations and Gergely Szollosi for useful discussions. This work was supported by the French National Research Agency (Grants ANR-18-CE02-0007-01 and ANR-19-CE45-0010). Simulations were performed using the computing facilities of the CC LBBE/PRABI..

Software Availability

All codes used to generate and analyze simulations performed in this study are available at: XXXXXXX.

References

Adato O., Ninyo N., Gophna U., Snir S. 2015. Detecting Horizontal Gene Transfer between Closely Related Taxa. PLOS Comput. Biol. 11:e1004408.

Chernikova D., Motamedi S., Csürös M., Koonin E.V., Rogozin I.B. 2011. A late origin of the extant eukaryotic diversity: divergence time estimates using rare genomic changes. Biol. Direct. 6:26.

Dalquen D.A., Zhu T., Yang Z. 2017. Maximum Likelihood Implementation of an Isolation-with-Migration Model for Three Species. Syst. Biol. 66:379–398.

Davín A.A., Tricou T., Tannier E., de Vienne D.M., Szöllősi G.J. 2020. Zombi: a phylogenetic simulator of trees, genomes and sequences that accounts for dead linages. Bioinformatics. 36:1286–1288.

Durand E.Y., Patterson N., Reich D., Slatkin M. 2011. Testing for Ancient Admixture between Closely Related Populations. Mol. Biol. Evol. 28:2239–2252.

Eaton D.A.R., Ree R.H. 2013. Inferring Phylogeny and Introgression using RADseq Data: An Example from Flowering Plants (Pedicularis: Orobanchaceae). Syst. Biol. 62:689–706.

Fontaine M.C., Pease J.B., Steele A., Waterhouse R.M., Neafsey D.E., Sharakhov I.V., Jiang X., Hall A.B., Catteruccia F., Kakani E., Mitchell S.N., Wu Y.-C., Smith H.A., Love R.R., Lawniczak M.K., Slotman M.A., Emrich S.J., Hahn M.W., Besansky N.J. 2015. Extensive introgression in a malaria vector species complex revealed by phylogenomics. Science. 347.

Forsythe E.S., Nelson A.D.L., Beilstein M.A. 2020a. Biased gene retention in the face of introgression obscures species relationships. Genome Biol. Evol.

Forsythe E.S., Sloan D.B., Beilstein M.A. 2020b. Divergence-Based Introgression Polarization. Genome Biol. Evol. 12:463–478.

Hahn M.W., Hibbins M.S. 2019. A Three-Sample Test for Introgression. Mol. Biol. Evol.

36:2878–2882.
Hibbins M., Hahn M. 2021. Phylogenomic approaches to detecting and characterizing introgression. .
Hibbins M.S., Matthew W Hahn. 2019. The Timing and Direction of Introgression Under the Multispecies Network Coalescent. :15.
Hudson R.R. 2002. Generating samples under a Wright-Fisher neutral model of genetic variation. Bioinformatics. 18:337–338.
Hug L.A., Baker B.J., Anantharaman K., Brown C.T., Probst A.J., Castelle C.J., Butterfield C.N., Hernsdorf A.W., Amano Y., Ise K., Suzuki Y., Dudek N., Relman D.A., Finstad K.M., Amundson R., Thomas B.C., Banfield J.F. 2016. A new view of the tree of life. Nat. Microbiol. 1:1–6.
Locey K.J., Lennon J.T. 2016. Scaling laws predict global microbial diversity. Proc. Natl. Acad. Sci. 113:5970–5975.
Louca S., Shih P.M., Pennell M.W., Fischer W.W., Parfrey L.W., Doebeli M. 2018. Bacterial diversification through geological time. Nat. Ecol. Evol. 2:1458–1467.
Mora C., Tittensor D.P., Adl S., Simpson A.G.B., Worm B. 2011. How Many Species Are There on Earth and in the Ocean? PLoS Biol. 9:e1001127.
Paradis E., Schliep K. 2019. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. Bioinformatics. 35:526–528.
Pfeifer B., Alachiotis N., Pavlidis P., Schimek M.G. 2020. Genome scans for selection and introgression based on k-nearest neighbour techniques. Mol. Ecol. Resour. 20:1597–1609.
Pfeifer B., Kapan D.D. 2019. Estimates of introgression as a function of pairwise distances. BMC Bioinformatics. 20:207.
Pittis A.A., Gabaldón T. 2016. Late acquisition of mitochondria by a host with chimaeric prokaryotic ancestry. Nature. 531:101–104.
Raup D.M. 1991. Extinction: bad genes or bad luck? New York: W.W. Norton.
Rosenzweig B.K., Pease J.B., Besansky N.J., Hahn M.W. 2016. Powerful methods for detecting introgressed regions from population genomic data. Mol. Ecol. 25:2387–2397.
Sato N. 2020. Complex origins of chloroplast membranes with photosynthetic machineries: multiple transfers of genes from divergent organisms at different times or a single endosymbiotic event? J. Plant Res. 133:15–33.
Staab P.R., Metzler D. 2016. Coala: an R framework for coalescent simulation. Bioinformatics. 32:1903–1904.
Susko E., Steel M., Roger A.J. 2021. Conditions under which distributions of edge length ratios on phylogenetic trees can be used to order evolutionary events. bioRxiv.:2021.01.16.426961.
Suvorov A., Kim B.Y., Wang J., Armstrong E.E., Peede D., D'Agostino E.R.R., Price D.K., Wadell P., Lang M., Courtier-Orgogozo V., David J.R., Petrov D., Matute D.R., Schrider D.R., Comeault A.A. 2021. Widespread introgression across a phylogeny of 155 Drosophila genomes. bioRxiv.:2020.12.14.422758.
Tricou T., Tannier E., Vienne D.M. de. 2021. Ghost lineages deceive introgression tests and call for a new null hypothesis. bioRxiv.:2021.03.30.437672.
Vosseberg J., van Hooff J.J.E., Marcet-Houben M., van Vlimmeren A., van Wijk L.M., Gabaldón T., Snel B. 2020. Timing the origin of eukaryotic cellular complexity with ancient duplications. Nat. Ecol. Evol.:1–9.
Weinbauer M., Rassoulzadegan F. 2007. REVIEW: Extinction of microbes: evidence and potential consequences. Endanger. Species Res. 3:205–215.

5

Détecter les groupes d'espèces fantômes grâce aux gènes transférés horizontalement

Ce manuscrit présente un premier concept de méthode capable de prédire la diversité éteinte et de détecter des clades non échantillonnés/inconnus dans une phylogénie avec les transferts horizontaux de gènes.

La majorité de la biodiversité qui a existé et qui existe aujourd'hui sur Terre nous est inconnue et les méthodes d'étude et d'identification de cette diversité fantôme sont limitées. Une des plus grosse source d'information sur la biodiversité éteinte est dans l'analyse des fossiles (Donoghue *et al.*, 1989). Cependant, la quantité de fossiles est à la fois limitée, hétérogène dans le temps, mais aussi entre les clades. Par ailleurs, bien que la majorité de l'évolution et de la transmission de matériel génétique se fasse verticalement (*i.e.* d'ascendant à descendant), les flux de gènes sont très fréquents à travers tout l'arbre du vivant.

Cela implique qu'une partie des gènes qui sont présents dans les espèces observables aujourd'hui, peut être apparue et/ou avoir évolué pendant un certain temps au sein d'espèces qui sont maintenant éteintes (*i.e.* qui n'ont pas donné de descendants) ou tout simplement inconnues (Maddison, 1997b; Galtier and Daubin, 2008). Cette affirmation implique que les gènes transférés

horizontalement contiennent des informations sur la nature et l'existence de la biodiversité éteinte et inconnue. Détecter ces transferts de gènes pourrait donc être une façon de prédire la diversité que l'on ignore encore.

Pour développer cette méthode, j'ai utilisé des données simulées avec Zombi (Davín *et al.*, 2020). Dans un premier temps, il était nécessaire de pouvoir comparer les transferts détectés par réconciliation avec ceux simulés. Cette comparaison est impossible avec des données empiriques puisque ici ce sont les espèces fantômes et les gènes qu'elles transfèrent qui nous intéressent. Ces espèces sont forcément absentes de nos données puisqu'elles nous sont inconnues.

Dans un premier temps des arbres d'espèces avec de l'extinction et de l'échantillonnage sont simulés avec *Zombi*. Toujours avec cet outil, des génomes on évolué le long des branches des arbres d'espèces, avec des événements de transferts horizontaux. Enfin, on utilise l'outil de réconciliation ALE pour détecter les événements de duplication, de perte et de transfert en utilisant l'arbre des espèces vivantes et observables (*i.e.* sans les espèces éteintes et non échantillonnées).

On montre sur une simulation simple qu'il est possible d'utiliser les transferts de gènes inférés par un outil de réconciliation pour localiser la position d'un groupe fantôme dans un arbre d'espèces.

Gene flow can reveal ghost lineages

Ghostbusters

Theo Tricou[1,*], Éric Tannier[1,2], and Damien M. de Vienne[1,*]

[1] *Université de Lyon, Université Lyon 1, UMR CNRS 5558 Laboratoire de Biométrie et Biologie Évolutive, 69622 Villeurbanne, France*
[2] *INRIA Grenoble Rhône-Alpes, Montbonnot-Saint-Martin F-38334, France*
[*] *Corresponding author: Theo Tricou, theo.tricou@univ-lyon1.fr; Damien de Vienne, damien.de-vienne@univ-lyon1.fr*

1 Introduction

Horizontal Gene Transfer (HGT), the transmission of genetic material across species boundaries, is a major driver of evolution in bacteria and archaea but also in eukaryotes (Daubin and Szöllősi, 2016). HGTs have long been considered as a source of noise for species tree reconstruction, but in the last decade they started to be seen as real opportunities for answering fundamental questions in Evolutionary Biology. For instance, it was demonstrated that HGTs could be used for finding the root of the Tree of Life (Abby *et al.*, 2012), could provide insight in the relative timing of speciation events is species trees (Szöllősi *et al.*, 2012), or could even help dating the Tree of Life (Davín *et al.*, 2018). HGTs, by definition, involve a donor and a recipient species. Identifying the recipient species is considered a simple task (but see (Tricou *et al.*, 2021) and Chapter 4) as it is the one whose genome hosts the sequence that has been identified as being horizontally transferred. Identifying the donor, however, is much more complicated because it may be extinct or simply unknown.

This so-called ghost biodiversity (extinct and unknown) may well constitute most of everything that has ever existed on Earth. Indeed, only a few million species have been described so far, which is small compared to those suspected to exist (Mora *et al.*, 2011), and most species that ever lived on earth are now extinct. The difference between the diversity that we know and the one we don't know is even larger in bacteria and archaea than in eukaryotes, because extinct species in these groups hardly fossilize (Waggoner, 1996) and because their estimated diversity is orders of magnitude larger than their known one (Locey and Lennon, 2016). Recent results revealing that a large amount of previously unsuspected diversity can be detected with metagenomic approaches (Hug *et al.*, 2016) reinforces this view.

This enormous proportion of extinct and unknown species (hereafter referred as *ghost lineages*), combined with the ubiquity of horizontal gene transfers in Evolution, implies that some genes that are present in species that are known today may have originated and/or evolved for a period of time in ghost lineages (Maddison, 1997; Galtier and Daubin, 2008; Szöllősi *et al.*, 2013). This also means that horizontally acquired genes may carry some information on the nature and existence of this ghost diversity.

In this brief paper, we explore the idea that HGTs could be used to detect or predict ghost diversity. We first present the theoretical concept behind this idea, and we then use simple simulations to show that, thanks to HGTs, the presence of a clade and its phylogenetic position can be correctly retrieved even though not a single genome is available for this clade.

The proof of concept presented here opens new lines of research for the future: the scarcity or absence of fossils, especially in Archaea and Bacteria, may not be synonymous of an absence of data to explore the past anymore.

Figure 1: Schematic justification of the use of HGT for detecting ghost diversity. Transfers from ghost groups to groups that are sampled (left) produces gene trees (center) whose reconciliation with the species tree can give a signal for the presence of a ghost group branching at the position where the transfer originates (right tree).

2 Theoretical concept behind the use of HGT detection to identify ghost diversity

Reconciliation methods are popular means of detecting evolutionary events, and among them horizontal gene transfers. By comparing (reconciling) gene and species trees, these methods infer the most likely evolutionary scenario that can explain the observed discordance (if any) between the trees, invoking evolutionary events such as gene duplications, transfers and losses.

An under-explored property of reconciliation approaches is that they also carry some potential for exploring unknown diversity. Indeed, and as illustrated in Figure 1, reconciliation of gene and species trees allows detecting genes present in extant species that have been acquired horizontally from species that are now extinct or still unknown. In a scenario in which a ghost clade is present in the species tree considered (red clade in Figure 1) and has one of his genes transferred to other extant lineages (gene x in Figure 1), one way to explain the observed discordance between the gene and the species tree could be a horizontal transfer leaving from the branch that would normally support the ghost clade (red branch in the rightmost tree of Figure 1) and going to the same recipient species in the green clade.

If the ghost clade is big and transfers are frequent, then many genes may give the same signal, thus revealing the presence and localisation of this ghost clade. When reconciling many gene trees with a species tree, an excess of transfers originating from the branch supporting the ghost clade is thus expected. Under the assumption that the number of transfers originating from a branch is correlated with the length of this branch, one way of detecting the presence of the ghost clade would be to identify branches in the species tree whose number of transfers emerging from it is high compared to their length.

Using simple simulations (see material and methods), we test this possibility of detecting ghost clades with reconciliation-based HGT detection.

3 Materials and Methods

To test our capacity to detect a ghost clade using HGT, we devised a 5-steps approach: (i) simulate a large species tree with one large ghost clade, (ii) let genes evolve along the branches of this tree, with duplications, losses and transfers, (iii) remove all species that are unknown or extinct (ghosts) in the species tree and the gene trees, (iv) reconcile gene and species trees and (v) search for a signal left by the ghosts, if any. These steps are detailed hereafter.

(i) The species was simulated using the birth-death simulator available in Zombi (Davín *et al.*, 2020). Speciation and extinction rates were set to 1 and 0.5, respectively, and simulation stopped when the tree reached 100 extant lineages. The tree obtained comprised 218 species in total (100 extant and 118 extinct, Figure 2A). In this tree a monophyletic group of 72 species (comprising 30 extant species) was manually chosen and had its branch length reduced by a factor of 0.8. This formed what we called the ghost clade (red clade in Figure 2A).

(ii) Zombi (Davín *et al.*, 2020) was used to evolve a genome of 500 genes along the branches of the complete species tree, using the following parameters: duplication rate = 1, loss rate = 3 and transfer rate = 5.

(iii) The 500 gene trees obtained and the species tree were pruned to remove non-extant species, using a dedicated python script.

(iv) Reconciliation analyses were conducted using ALE (Szöllősi *et al.*, 2015) on each one of the 500 pruned gene trees, using default parameters. ALE is a probabilistic approach that estimates duplications, transfers and losses (DLT) rates. Using those rates, the program then outputs 100 sampled reconciled amalgamated gene trees from which the frequency of transfer events is computed. This produced a set of 500 reconciliation scenarios, involving duplications, transfers and losses.

(v) For each branch in the species tree, the sum of all inferred transfers leaving from it (over all 500 reconciliation scenarios) was computed, and compared to its length. As explained before, we expect an excess of transfers to be observed on the branch supporting the ghost clade. To test for this effect, taking into account the fact that longer branches are also expected to show more transfers, we performed a linear regression of the number of transfers inferred with ALE by the branch length and looked for outliers.

For each branch we computed is "outlierness" by computing the probability to belong to the expected distribution of the number of transfers given their length. First we standardized the observed number of transfers leaving the branch by subtracting the corresponding fitted value inferred from the linear model and dividing by the model standard deviation. Second, we compared this value to the normal distribution computed from all standardized values to compute a p-value. Finally, p-values were adjusted using Benjamini-Hochberg procedure (control the false discovery rate, the expected proportion of false discoveries amongst the rejected hypotheses) (Klipper-Aurbach *et al.*, 1995) with a false discovery rate cutoff (FDR) of 0.05. Any branch below this threshold revealed a number of transfers originating from it exceeding the expected one considering its length, an outlier.

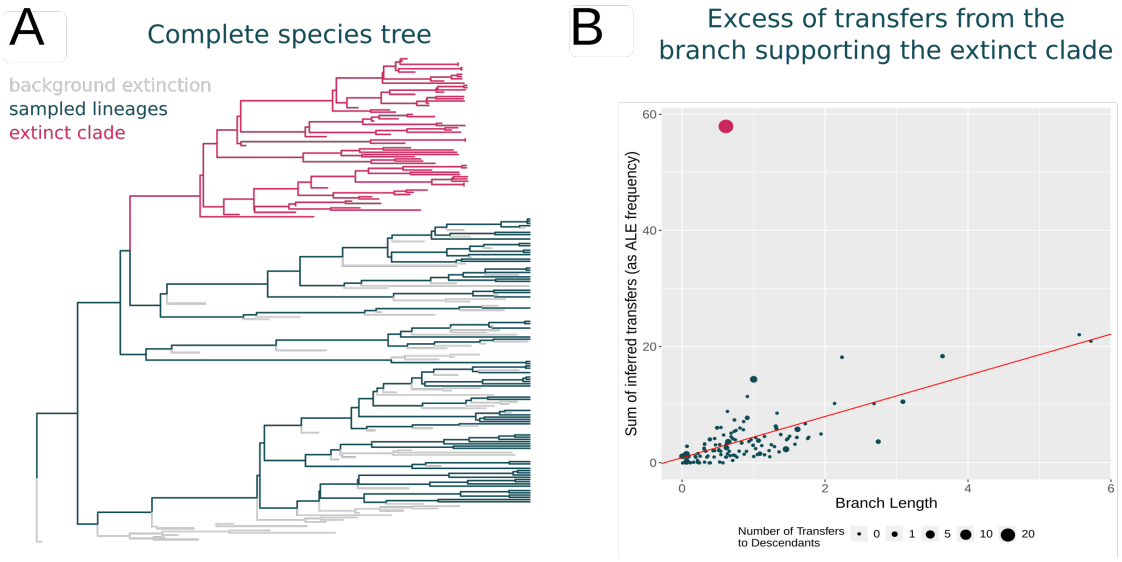


Figure 2: (A) The complete species tree used for the simulation presented here. Extant species are represented in dark green, extinct lineages in grey and the ghost clade in red (see text). (B) Distribution of the number of transfers leaving a branch (y-axis) as a function of the branch length (x-axis). The red line is a linear model ($R^2 = 0.26$, $P = 1.043 \times 10^{-16}$). The red dot correspond to the branch supporting the ghost clade in the extant species tree.

4 Results

Across all gene trees obtained after evolving a genome of 500 genes along the branches of a large species tree (Figure 2A), 1101 transfers, 16 duplications and 1776 losses were simulated with Zombi (Davín *et al.*, 2020). After reconciliation, 514.59 transfers, 1.01 duplications and 798.51 losses were detected by ALE.

Figure 2B depicts the sum of transfers inferred from each possible donor branch (as a sum of all HGTs frequency) as a function of their length. We observed that the longer the branch, the higher the number of transfers leaving from this donor ($R^2 = 0.26$, $P = 1.043 \times 10^{-16}$).

We detected one outlier from this distribution (FDR adjusted $P = 1 \times 10^{-16}$). This donor branch showed a much higher number of transfers than expected given its length. With a length of 0.6 (in unit of time step), if the donor followed the same linear model than others, we would expect around 2 transfers. Here, we observed nearly 30 times that value, with 58.16 transfers inferred by ALE as originating from this specific branch. This donor corresponds to the branch that supported the ghost clade (red clade in Figure 2A). This excess of transfers is due to transfers that occurred between ghost lineages within the ghost clade and extant lineages outside. This result suggests that a ghost clade in a species tree phylogeny could theoretically be located via the signal of HGT, and this corroborates our suggestion that HGT could be used to explore the unknown diversity.

5 Discussion

In the vein of others who exploited the signal from HGTs (Abby *et al.*, 2012; Davín *et al.*, 2018; Morel *et al.*, 2020), we show here that HGTs could be used as a surrogate to identify ghost lineages, *i.e.* extinct, unknown or unsampled lineages in species trees. These are very preliminary results but are also very encouraging. Of course, this example is simplistic, because the true gene and species trees are known, including branch

lengths, because transfers are not too rare or too frequent, because they are homogeneous along the branches, and because only one clade is extinct. However, it gives an exciting first proof of concept which, according to us, justifies to go further in the direction proposed in this project.

This method may be one solution to have access to extinct biodiversity in groups of species that do not fossilize. The proof of concept presented here is an important validation of the feasibility of the method. We are confident that interesting results will come out of it. For example, large groups of species are still unknown, notably because cultivation-independent surveys using metagenomics approaches will miss entire phyla that show divergent 16S ribosomal RNA sequences (as recently demonstrated for the CPR bacterial group, Brown *et al.* (2015)). Such groups could be identified with the method presented here if transfers occurred between these ghost groups and lineages already described. The observed amount of gene sharing in prokaryotes (Beiko *et al.*, 2005) is big enough for being confident that such transfers occur, especially if species live in similar environments.

References

Abby, S. S., Tannier, E., Gouy, M., and Daubin, V. 2012. Lateral gene transfer as a support for the tree of life. *Proceedings of the National Academy of Sciences*, 109(13): 4962–4967.

Beiko, R. G., Harlow, T. J., and Ragan, M. A. 2005. Highways of gene sharing in prokaryotes. *Proceedings of the National Academy of Sciences*, 102(40): 14332–14337.

Brown, C. T., Hug, L. A., Thomas, B. C., Sharon, I., Castelle, C. J., Singh, A., Wilkins, M. J., Wrighton, K. C., Williams, K. H., and Banfield, J. F. 2015. Unusual biology across a group comprising more than 15% of domain bacteria. *Nature*, 523(7559): 208–211.

Daubin, V. and Szöllősi, G. J. 2016. Horizontal Gene Transfer and the

History of Life. *Cold Spring Harbor Perspectives in Biology*, 8(4): a018036.

Davín, A. A., Tannier, E., Williams, T. A., Boussau, B., Daubin, V., and Szöllősi, G. J. 2018. Gene transfers can date the tree of life. *Nature Ecology & Evolution*, 2(5): 904–909.

Davín, A. A., Tricou, T., Tannier, E., de Vienne, D. M., and Szöllősi, G. J. 2020. Zombi: a phylogenetic simulator of trees, genomes and sequences that accounts for dead linages. *Bioinformatics*, 36(4): 1286–1288.

Galtier, N. and Daubin, V. 2008. Dealing with incongruence in phylogenomic analyses. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 363(1512): 4023–4029.

Hug, L. A., Baker, B. J., Anantharaman, K., Brown, C. T., Probst, A. J., Castelle, C. J., Butterfield, C. N., Hernsdorf, A. W., Amano, Y., Ise, K., Suzuki, Y., Dudek, N., Relman, D. A., Finstad, K. M., Amundson, R., Thomas, B. C., and Banfield, J. F. 2016. A new view of the tree of life. *Nature Microbiology*, 1(5): 1–6.

Klipper-Aurbach, Y., Wasserman, M., Braunspiegel-Weintrob, N., Borstein, D., Peleg, S., Assa, S., Karp, M., Benjamini, Y., Hochberg, Y., and Laron, Z. 1995. Mathematical formulae for the prediction of the residual beta cell function during the first two years of disease in children and adolescents with insulin-dependent diabetes mellitus. *Medical hypotheses*, 45(5): 486–490.

Locey, K. J. and Lennon, J. T. 2016. Scaling laws predict global microbial diversity. *Proceedings of the National Academy of Sciences*, 113(21): 5970–5975.

Maddison, W. P. 1997. Gene Trees in Species Trees. *Systematic Biology*, 46(3): 523–536.

Mora, C., Tittensor, D. P., Adl, S., Simpson, A. G. B., and Worm, B. 2011. How Many Species Are There on Earth and in the Ocean? *PLoS Biology*, 9(8): e1001127.

Morel, B., Kozlov, A. M., Stamatakis, A., and Szöllősi, G. J. 2020. GeneRax: A Tool for Species-Tree-Aware Maximum Likelihood-Based Gene Family Tree Inference under Gene Duplication, Transfer, and Loss. *Molecular Biology and Evolution*, 37(9): 2763–2774.

Szöllősi, G. J., Boussau, B., Abby, S. S., Tannier, E., and Daubin, V. 2012. Phylogenetic modeling of lateral gene transfer reconstructs the pattern and relative timing of speciations. *Proceedings of the National Academy of Sciences*, 109(43): 17513–17518.

Szöllősi, G. J., Tannier, E., Lartillot, N., and Daubin, V. 2013. Lateral Gene Transfer from the Dead. *Systematic Biology*, 62(3): 386–397.

Szöllősi, G. J., Davín, A. A., Tannier, E., Daubin, V., and Boussau, B. 2015. Genome-scale phylogenetic analysis finds extensive gene transfer among fungi. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 370(1678): 20140335.

Tricou, T., Tannier, E., and Vienne, D. M. d. 2021. Ghost lineages deceive introgression tests and call for a new null hypothesis. *bioRxiv*, page 2021.03.30.437672.

Waggoner, B. 1996. Bacteria and protists from Middle Cretaceous amber of Ellsworth County, Kansas. *PaleoBios*, 17(1): 20–26.

6

Discussion et Perspectives

Contents

6.1 Résumé des principaux résultats de la thèse

Mon travail a consisté à utiliser des simulations pour générer des arbres phylogénétiques d'espèces contenant des espèces fantômes, puis à faire évoluer des génomes le long de leurs branches. Enfin j'ai évalué les capacités de différents outils bioinformatiques à correctement détecter les évènements de flux de gènes et

regardé s'il était possible d'utiliser leur signal comme outil de détection de la biodiversité.

Premièrement, j'ai présenté dans le chapitre 2 *Zombi*, un outil pour simuler l'évolution d'espèces, de génomes et de séquences qui prend en compte la participation intégrante des lignées éteintes. Le développement de cet outil m'a permis d'explorer plus en détail l'impact des lignées fantômes sur notre capacité à détecter et identifier les acteurs des flux de gènes. Deuxièmement, j'ai illustré dans le chapitre 3 comment l'interprétation du test ABBA-BABA, une méthode populaire de détection de l'introgression, était influencée par l'absence de considération des espèces fantômes et comment son interprétation pouvait être inversée si l'introgression identifiée venait d'une espèce inconnue ou insoupçonnée. Troisièmement, j'ai ré-analysé trois résultats de la littérature dans le chapitre 4 et j'ai montré que la présence de lignées fantômes remettait en question les conclusions des études utilisant les longueurs des branches des arbres phylogénétiques pour détecter les flux de gènes, les ordonner dans le temps, ou pour identifier de "bons" gènes pour reconstruire l'histoire évolutive des organismes. Enfin, pour montrer que les flux de gènes fantômes ne sont pas qu'une source de biais et d'erreurs d'interprétations des méthodes de détection, j'ai présenté dans le chapitre 5 une preuve de concept de la possibilité de détection et de localisation des clades fantômes dans une phylogénie grâce à la détection de gènes transférés.

6.2 Vers une détection des lignées fantômes

Initialement, le titre de ma thèse était "Détecter la diversité éteinte et inconnue avec les transferts horizontaux de gènes". Le projet était centré sur le développement d'une méthode de détection, et sur l'application de cette méthode sur des données simulées et empiriques. Les résultats avec les simulations (dont certains sont présentés dans le chapitre 5) laissaient penser que la détection des gros groupes fantômes pouvait être accomplie en utilisant le signal des transferts horizontaux. De façon plus ambitieuse, la fréquence des transferts semblait même pouvoir être utilisée pour détecter des évènements macro-évolutifs comme des extinction de masse ou des radiations. Mais l'application sur les données empiriques a révélé certaines limites de l'approche. Dans les sections suivantes je reviens sur tout le travail effectué sur la détection des espèces fantômes et

l'application aux données empiriques, et pourquoi ces résultats nous ont amené à réorienter le sujet de la thèse vers la question de l'impact des espèces fantômes sur notre capacité à détecter les flux de gènes.

6.2.1 Délimiter une zone de détection avec les simulations

Une des étapes du développement de la méthode de détection des lignées fantômes grâce aux transferts à été de définir les conditions dans lesquelles la méthode pouvait être appliquée. Dans les simulations que nous avons réalisées, on arrivait à montrer que la méthode était robuste dans une large gamme de paramètres, biologiquement réalistes. Je résume ici les principales conclusions auxquelles nous sommes arrivées pour les paramètres testés :

- Taille du groupe : la taille du groupe a peu d'influence sur notre capacité à le détecter. Tant que le nombre de lignées composant le groupe est important comparé au nombre d'espèces fantômes réparties dans l'arbre, alors il est possible de le retrouver avec les transferts. Avec nos simulations, la corrélation entre le nombre de transferts et la longueur des branches était tellement forte que la moindre déviation était identifiable. Mais ces résultats questionnent aussi la robustesse de notre méthodes dans un contexte où plusieurs grands clades sont inconnus et où la quantité d'espèces fantômes surpasse en nombre celle des espèces de la phylogénie, un scénario qui semble finalement très probable.

- Taux de transfert : il est important d'évaluer si la quantité de transferts nécessaire à l'identification des groupes éteints est en accord avec les quantités observées dans la nature. Ce que nos résultats suggèrent c'est qu'il est très facile de détecter un groupe inconnu même avec très peu de transferts (quelques dizaines en tout). Par contre, on perd la capacité de détecter un groupe si le nombre de transferts est trop élevé (>15 transferts par gène pour un arbre de 50 feuilles). Les taux de transferts compatibles avec la détection d'un groupe fantôme d'après nos simulations sont en accord avec le nombre de transferts par famille de gène inférés dans deux phylogénies de 28 fungis (11 387 familles de gènes) et 40 cyanobactéries (7 415 familles de gènes) : respectivement 0.07 et 0.16 transferts par gènes.

- Taux de duplication et perte : Ces deux paramètres sont traités ensemble car ils ont un effet antagoniste. La perte de gène diminue la taille des arbres de gènes et peut impacter les espèces qui reçoivent des transferts en provenance des

espèces fantômes. Plus il y a de perte de gènes, plus il y a de chance de perdre les gènes dans les espèces fantômes et les espèces receveuses, et donc de perdre le signal nécessaire pour détecter un groupe inconnu. À l'inverse, les duplications de gènes augmentent la taille des arbres de gènes, et donc augmentent le nombre de transferts possibles entre des espèces fantômes et le reste de l'arbre. Plus il y a de duplications, plus l'excès de transfert partant de la branche qui soutient le groupe inconnu est marqué, ce qui peut faciliter son identification.

L'ensemble de ces résultat suggère que les taux évolutifs rencontrés dans le vivant ne devrait pas ou peu influencer et entraver notre capacité à détecter la biodiversité inconnu.

6.2.2 Détecter des traces d'extinctions de masse chez les bactéries et archées

Une extension possible du test de la détection des espèces fantômes est de l'utiliser pour détecter des évènements macro-évolutifs majeurs comme des extinctions de masse ou des radiations. Par exemple, il n'y a aucune information indiquant si les six extinctions de masse identifiées au cours des 600 derniers millions d'années (Hallam and Wignall, 1997; Raup and Sepkoski, 1982) ont également affecté la diversité microbienne, et si oui dans quelle mesure. Si des transferts peuvent être détectés dans le passé, une diminution drastique de la biodiversité devrait être associée à une diminution du nombre de transferts horizontaux détectés. De plus, après de tels événements, une grande quantité d'habitats se retrouvent disponibles. Par conséquent, des phénomènes de radiations adaptatives sont attendus et de nombreuse espèces vont évoluer en peu de temps pour remplir ces niches nouvellement disponibles. Ces radiations sont souvent liées à des augmentations significatives du nombre d'échanges horizontaux (Ford *et al.*, 2015). La détection d'une variation dans le paysage des transferts de gènes d'un arbre phylogénétique, comme une chute suivi par une augmentation du nombre de transferts, pourrait nous fournir un début de réponse.

La figure 6.1 est une illustration de cette variation dans le paysage des transferts. L'arbre à gauche a été simulé avec Zombi (Davín *et al.*, 2020). Un événement d'extinction de masse a été simulé entre les deux droites verticales bleus réduisant

le nombre d'espèces vivante par un facteur 20. Un génome de 500 de gènes a été simulé le long des branches de l'arbre avec des évènements de transferts de gènes, puis des réconciliations (avec ALE Szöllősi *et al.*, 2015) ont été faites entre les arbres de gènes et l'arbre des espèces vivantes. Dans cet arbre, à droite, est représenté par un cercle le nombre de transferts inféré par branche. Plus un cercle rouge est opaque plus il y a de transfert partant de cette branche. Comme prédit au-dessus, une zone dans laquelle les cercles sont nettement plus opaque que dans le reste de l'arbre peut être observée. Cette zone correspond à l'époque de l'extinction de masse dans l'arbre d'espèces complet, à gauche. Comme la méthode de détection des groupes fantômes présentée dans le chapitre 5, ce concept n'est encore qu'une simple illustration avec une simulation très simpliste. Néanmoins c'est une piste prometteuse pour explorer des évènements macro-évolutifs du passé.

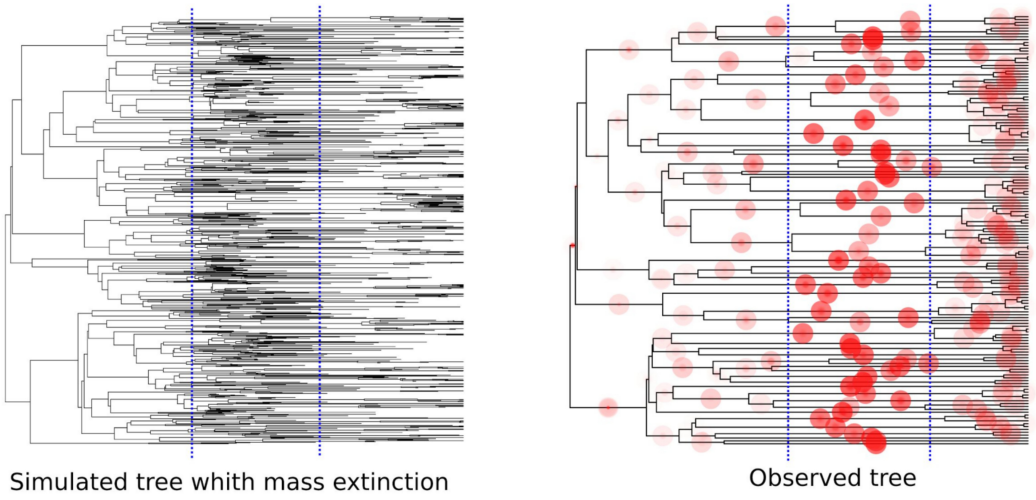


Figure *6.1* – *Identifier des extinctions de masse avec les transferts de gènes. À gauche, l'arbre d'espèces. Entre les deux droite est représenté l'évènement d'extinction de masse. À droite, l'arbre des espèces vivantes et les cercle correspondant avec la quantité de transfert.*

6.2.3 L'application sur des données empiriques

Nous avons tenté de valider la méthode de détection de groupes fantômes (présentée chapitre 5) sur des données biologiques. Une façon de valider notre approche est de simuler la présence d'un groupe fantôme en le retirant. Ainsi, pour une phylogénie donnée, un groupe monophylétique d'une taille conséquente est retiré à la main, puis des réconciliations entre les arbres de gènes et l'arbre d'espèces sont réalisées pour détecter des transferts et chercher le signal de de ce

groupe. J'ai appliquée cette méthode à plusieurs ensembles de données : une phylogénie de 160 espèces de *Caulobacteraceae* (111 Rhizobiales, 11 Caulabacterales et 38 Rhodobacterales) reconstruite pendant ma première année de thèse, une phylogénie de 104 *Agaricomycotina* (Varga *et al.*, 2019), trois phylogénies de trois phylums d'archées et enfin la phylogénie de 36 cyanobactéries utilisée abondamment dans le contexte de l'étude des transfert horizontaux de gènes (Szöllősi *et al.*, 2012; Szöllősi *et al.*, 2013; Szöllősi *et al.*, 2013; Szöllősi *et al.*, 2015; Davín *et al.*, 2018; Morel *et al.*, 2019; Szöllősi *et al.*, 2021).

L'application de la méthode à des données empiriques a révélé plusieurs limitations et problèmes.

La première est que le nombre des transferts partant des branches n'était pas ou peu corrélée à la longueur des branches de l'arbre phylogénétique, alors même que cette corrélation était à la base de la méthode envisagée de détection des groupes fantômes (voir chapitre 5). Cette absence de corrélation est bien illustrée dans les données des Rhizobiales (Figure 6.2). Le nombre de transferts partant des branches courtes (à gauche dans le graphique) est très variable, et c'est même une des branches les plus courtes qui présente la plus grande quantité de transferts. Cette absence de corrélation n'est pas surprenante *a posteriori*. Rien ne laisse supposer que les évènement de transferts soient répartis de façon homogène, on sait par exemple qu'il existe des autoroutes de transfert préférentielles entre certaines organismes ou groupes d'organisme (Beiko *et al.*, 2005). De plus, on utilise ici des arbres non datés, donc les longueurs des branches ne représentent pas la durée d'existence d'une lignée mais seulement un taux de mutation.

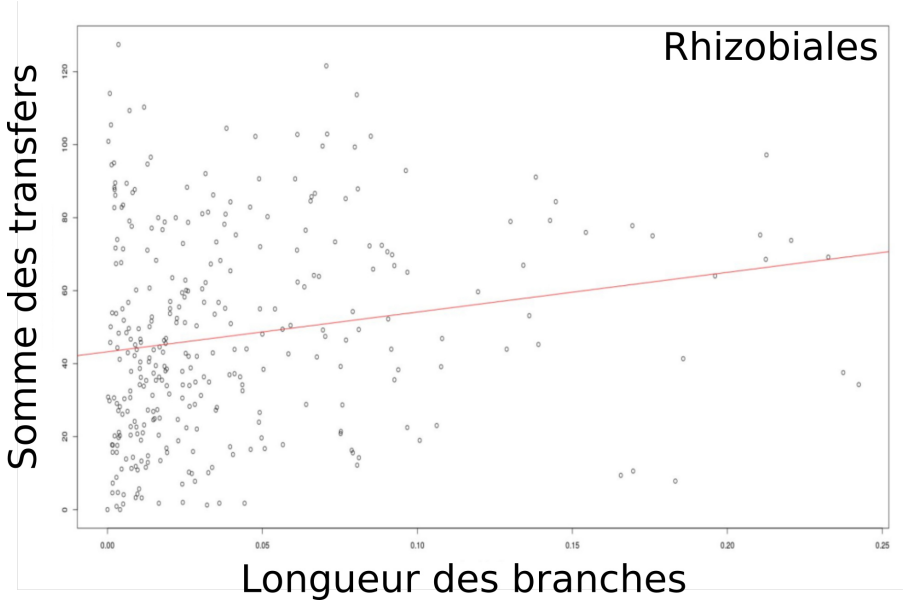


Figure *6.2* – *Distribution des transferts chez les Rhizobiales. Le jeu de données des Rhizobiales est composé de 426 familles de gènes universels. Comme dans le chapitre 5, le graphique représente les nombre de transferts inféré par ALE partant d'une branche en fonction de la longueur de cette branche. La droite rouge représente une régression linéaire.*

La seconde est que notre méthode repose sur un excès attendu du nombre de transferts partant des branches supportant un groupe fantômes comparé aux autres. Or, en comparant le nombre de transferts avant et après échantillonnage, on a systématiquement observé que le nombre de transferts partant de la branche supportant le groupe inconnu n'augmentait pas ou pas beaucoup après échantillonnage. Dans certains cas on observait même l'inverse, une diminution du nombre de transferts partant de la branche. Ces résultats inattendus suggèrent que les transferts partant d'un groupe inconnu ne sont pas inférés comme partant de la branche le soutenant après l'avoir retiré. Soit un scénario de réconciliation alternatif existe qui explique la discordance entre l'arbre d'espèces et les arbres de gènes. Soit ces transferts disparaissent tout simplement, ils ne sont plus inférés par la réconciliation.

Cette dernière observation, non conformes à nos attentes d'après les simulations, nous a poussé à nous questionner sur la capacité des méthodes de réconciliation — et des méthodes de détection des flux de gènes en générale — à correctement détecter les transferts en présence de lignées fantômes. C'est suite à ces résultats que la direction de la thèse à été changée, pour se concentrer non plus sur la détection de la biodiversité inconnue mais sur l'effet de celle-ci sur nos capacité

à correctement détecter les flux de gènes.

6.3 Tester l'impact des lignées fantômes sur les outils de réconciliation

Les résultats obtenus sur données empiriques (voir ci-dessus) nous poussent à nous questionner sur la robustesse des outils de réconciliation face à la présence d'espèces fantômes. Une suite évidente et immédiate de mon travail serait donc d'évaluer l'impact des espèces fantômes sur la capacité des outils de réconciliation à détecter précisément les transferts horizontaux de gènes et à identifier correctement les couples d'organismes donneurs et receveurs. Plusieurs résultats suggèrent qu'une part non négligeable des transferts inférés par ces méthodes serait incorrecte. Abby *et al.* (2010) ont montré, par le biais de simulations (sans espèces fantômes), que les outils de réconciliation détectent en général le nombre correct d'évènements de transferts (aussi observé avec le modèle ODT implémenté dans ALE Szöllősi *et al.*, 2012) mais que l'identification des donneurs et receveurs est, elle, bien moins précise. Par la suite, Chauve *et al.* (2017) ont développé une méthode permettant de filtrer les résultats de réconciliation calculées par ALE pour conserver le scénario de transferts avec la cohérence temporelle maximale. Cette méthode est utilisée pour ordonner de façon relative les nœuds des arbres phylogénétiques d'espèces. Leurs résultats suggèrent là encore que pour maximiser les contraintes temporelles induites pas les transferts, entre 5% et 20% des transferts doivent être supprimés.

Les outils de réconciliation sont encore assez nouveaux et leur utilisation assez peu répandue. Peu d'études ont été conduites pour évaluer leur capacité à réellement identifier des évènements correctes et, à ma connaissance, seulement deux intègrent des espèces fantômes : Chauve *et al.* (2017), qui simulent de l'échantillonnage, et Weiner and Bansal (2021) qui utilisent Zombi (Davín *et al.*, 2020) pour simuler de l'extinction. Dans le cadre ma thèse, j'ai eu l'opportunité de co-encadrer, avec mes encadrants, Syrine Benali une stagiaire de Master 2. Pendant son stage, elle a exploré l'effet de l'échantillonnage sur les capacités d'identification correcte des donneurs et des receveurs d'un transfert avec ALE.

Avec Zombi un génome de 100 gènes a été simulé le long des branches d'un arbre de 160 feuilles. Une première réconciliation a été calculée sur chacun des 100 arbres de gènes. Ensuite, 45 espèces ont été sélectionnées aléatoirement et

supprimées de l'arbre d'espèces et des arbres de gènes, pour simuler des espèces inconnues. Une seconde réconciliation a alors été calculée entre l'arbre d'espèces réduit et les arbres de gènes réduits. Pour chaque famille de gènes, on a ensuite comparé les prédiction des donneurs et receveurs de transferts avant et après échantillonnage et regardé si ces prédictions étaient en accord. Dans la figure 6.3, chaque point représente un transfert inféré par ALE avant et après avoir supprimé (*échantillonné*) des espèces (160 feuilles en abscisse contre 115 feuilles en ordonnée). Tous les points qui dévient de la bissectrice sont des transferts dont le score varie avant et après l'échantillonnage. On observe qu'un grand nombre de transferts se retrouvent soit sur l'axe des abscisses soit sur l'axe de ordonnées, ce qui correspond soit à des transferts qui ont disparu après échantillonnage soit qui sont apparus après échantillonnage. Ainsi, l'échantillonnage fait beaucoup varier les transferts inférés par la réconciliation avec ALE. Ces résultats préliminaires suggèrent que le nombre d'espèces fantômes (ici échantillonnées) influe beaucoup sur la capacité à correctement identifier les donneurs et les receveurs.

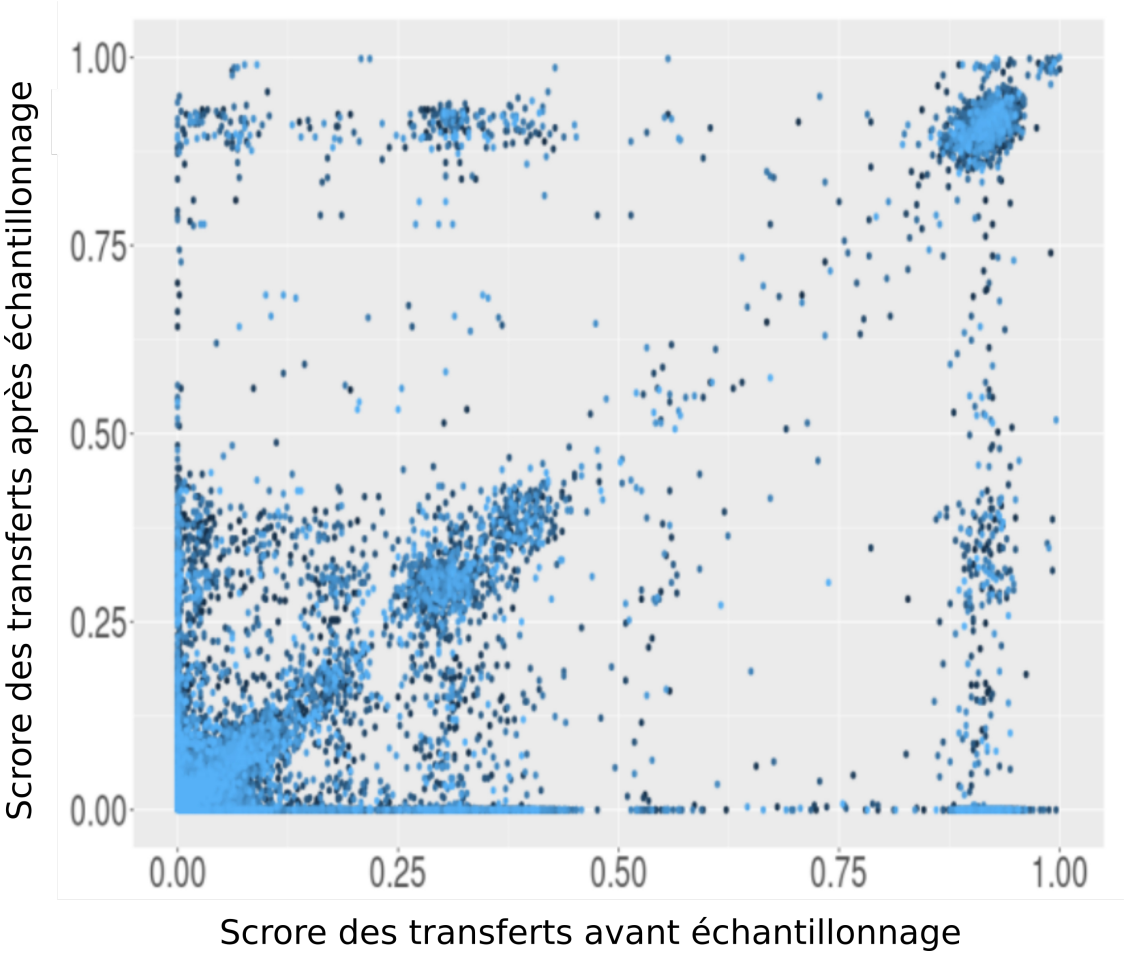


Figure *6.3* – *Représentation des scores de transfert inféré par ALE avant échantillonnage dans un arbre de 160 feuilles (abscisse) et après échantillonnage dans un arbre de 115 feuilles (ordonnée). Chaque point du graphique représente un événement de transfert. Si un point est sur la bissectrice du graphique alors les scores du transfert inféré avant l'échantillonnage (abscisse) et après échantillonnage (ordonnée) sont les mêmes. Les points sur l'axe des abscisses représentent les transferts inférés avant l'échantillonnage et qui ont disparu après l'échantillonnage. Inversement les points sur l'axe des ordonnées représentent les transferts absents avant l'échantillonnage mais qui ont été inférés après l'échantillonnage.*

Ce sont ici des résultats préliminaires qui nécessitent plus de contrôle et d'analyse pour être validés et comprendre comment les réconciliations sont affectées par les espèces fantômes. Il serait donc intéressant de poursuivre ces travaux et les étendre à un plus large panel d'outils de réconciliations, et pas seulement ALE. Et de tester l'impact de la taille des arbres d'espèces, de l'échantillonnage et de la quantité d'espèces éteintes sur les capacité des outils de réconciliations à correctement détecter les évènements de duplications, de transferts et de pertes de gènes. De tels travaux apporteraient des précisions sur les limites de ces outils ainsi que sur des pistes possibles d'améliorations.

6.4 Le casse tête de la détection des flux de gènes

Mes travaux pendant cette thèse ont très majoritairement consisté à simuler des scénarios de flux de gènes. Ainsi j'ai toujours eu une connaissance des vrais scénarios de flux de gènes, leur provenance véritable et les vrais receveurs. Cependant la détection des flux de gènes est une tâche très complexe. Les événement de flux inférés par les différentes méthodes varient considérablement.

En règle générale, les méthodes paramétriques sont relativement peu précises et sont sujettes à un grand nombre de faux positifs (Friedman and Ely, 2012). Prenons par exemple les méthodes qui utilisent le taux de GC (voir la section 1.2.1). Une des limitates de ces méthodes est que non seulement le taux de GC est variable entre individus, mais il l'est aussi tout le long de l'ADN (Bohlin *et al.*, 2010), ce qui limite la sensibilité de la détection de flux de gènes par ces outils. Une autre limite est qu'on ne peut détecter que des transferts depuis des organismes ayant des taux de GC assez éloignés pour avoir un signal, ce qui est d'autant moins probable que les transferts ont lieu vers des espèces qui seraient phylogénétiquement proches. De plus, une limitation importante des méthodes paramétriques est qu'elles ne réussissent pas à détecter les flux de gènes trop anciens car les séquences transférées de longue date vont être sujettes aux mêmes forces évolutives que le reste du génome (Lawrence and Ochman, 1997), et vont voir leurs composition tendre vers celles du génome hôte.

Les méthodes phylogénétiques ne sont pas sans défaut non plus. Ces méthodes nécessitent de connaître les vraies topologies des arbres d'espèces et de gènes pour réussir à détecter précisément des évènements de flux de gènes, or dans bien des cas, il existe de fortes incertitudes. Par exemple, dans le cas du complexe des *Anopheles gambiae* (Fontaine *et al.*, 2015), la quantité d'introgression entre les espèces composants ce clade brouille le signal phylogénétique et entrave la reconstruction d'un arbre d'espèces précis. Une phylogénie d'espèce erronée est donc une mauvaise référence pour évaluer si des flux de gènes ont impacté l'évolution des lignées. De la même façon, les erreurs dans la reconstruction de l'histoire des gènes vont entraîner la détection de transferts simplement parce que la topologie est erronée et pas parce que des discordances entre les histoires évolutives des espèces et des gènes existent. ALE tente de résoudre ce problème en calculant les réconciliations sur des échantillons d'arbre d'espèces généré par des méthodes bayésiennes et va ainsi prendre en compte cette incertitude pour

identifier les flux de gènes.

Une dernière limitation des méthodes phylogénétiques, qui n'est pas un biais méthodologique, est que les discordances entre les arbres de gènes et d'espèces peuvent provenir d'autres évènements biologiques que des transferts. On peut citer les paralogies cachées qui peuvent produire des topologies discordantes par l'enchaînement d'évènements de duplication et de perte de gènes. La comparaison de ces topologies à celle de l'arbre d'espèces peut aboutir à l'inférence érronée de flux de gènes (Devos *et al.*, 2012). De la même façon, un excès d'ILS peut aussi brouiller l'interprétation de tests, comme le test ABBA-BABA. Ces excès peuvent survenir suite à des histoires démographiques complexes (comme des phénomènes de goulet d'étranglement) (Martin *et al.*, 2015; Lawson *et al.*, 2018).

La validation de ces méthodes de détection se fait souvent par le biais d'analyses comparatives, reposant sur des simulations. Cependant ces simulations sont souvent simplistes, à l'instar de celles que nous avons réalisées dans cette thèse. Simuler des scénarios de flux de gènes réalistes n'est pas trivial. La véritable fréquence et la distribution des flux de gènes dans le vivant est encore grandement méconnue ce qui rend difficile l'estimation des taux de transfert a utiliser pour des simulations. Bien que ce soit une méthode imparfaite, la simulation de donnée est malgré tout la meilleure approche à notre disposition pour tester la robustesse des méthodes de détection de flux.

6.5 Simuler l'inconnu

L'utilisation quasi exclusive de simulations dans mon travail soulève aussi des questions quant à la pertinence de leur utilisation. Si on prend l'exemple de la détection des flux de gènes, on va vouloir tester la capacité d'une méthode a correctement détecter les flux sur des simulations *ad hoc*, c'est-à-dire des simulations faites dans le seul but de la tester. C'est par exemple le cas du logiciel *Zombi* qui a été conçu dans l'optique de prendre en compte les transferts qui proviennent de lignées fantômes pour tester notre capacité à détecter de tels évènements. Dans cette situation certains éléments de la méthode sont inévitablement intégrés dans le simulateur, qui est alors susceptible de ne générer que des instances faciles pour cette méthode et n'a aucune chance d'atteindre la complexité des données réelles. Encore une fois, Zombi a été développé par des

utilisateurs de l'outil de réconciliation ALE. Mais même lorsque les simulations sont basées sur un logiciel général qui n'a pas été conçu pour une étude spécifique (Dalquen *et al.*, 2012; Sjöstrand *et al.*, 2013; Arenas and Posada, 2014; Mallo *et al.*, 2016), certains principes sous-jacents importants demeurent, partagés entre les méthodes de simulation et d'inférence simplement parce qu'ils sont largement acceptés (souvent implicitement) dans la communauté bioinformatique. Par exemple, ne simuler que des espèces vivantes décrites, celles qu'on observe, alors que les espèces fantômes sont ignorées et non simulées. Dans une telle situation, les méthodes ne sont testées que dans un monde conçu pour elles, ce qui n'évalue pas leur efficacité dans le monde réel.

6.6 Conclusion

Pour conclure, j'ai souligné dans cette thèse l'importance de prendre en compte les espèces fantômes pour l'étude des flux de gènes. J'ai montré l'impact de la biodiversité fantôme sur nos capacités à détecter ces flux entre organismes, mais aussi proposé d'exploiter ce signal pour l'exploration de la biodiversité inconnue. Ces travaux représentent une des premières contributions (au côté de Szöllősi *et al.* (2015); Davín *et al.* (2018); Martin Kuhlwilm (2020); Ottenburghs (2020)) au changement de mentalité de la communauté scientifique sur l'importance de la biodiversité fantôme.

7

Annexes

Contents

Durant les trois années de ma thèse, j'ai été amené à collaborer avec plusieurs collègues et à mettre mes compétences en phylogénie et en analyse de données au profit d'autres projets. Je présente ici brièvement ces collaborations et les articles ou résultats qui en ont été issus.

7.1 Étude de l'évolution de la taille des génomes chez les drosophiles

Ce projet est porté par Annabelle Haudry et en collaboration avec deux anciens doctorants du LBBE, Vincent Merel et Thibault Latrille (maintenant tous les deux en postDoc à Lausanne en Suisse). Il porte sur l'évolution de la taille des génomes des drosophiles, qui semble être fortement influencée par le contenu en éléments transposables, et la taille efficace des populations. Pour ce projet j'ai reconstruit la phylogénie de 82 lignées de drosophiles (Figure 7.1). Cette phylogénie a été utilisé pour comparer la taille des génomes et leurs composition en éléments transposables. Un article est actuellement en préparation par Vincent et Annabelle.

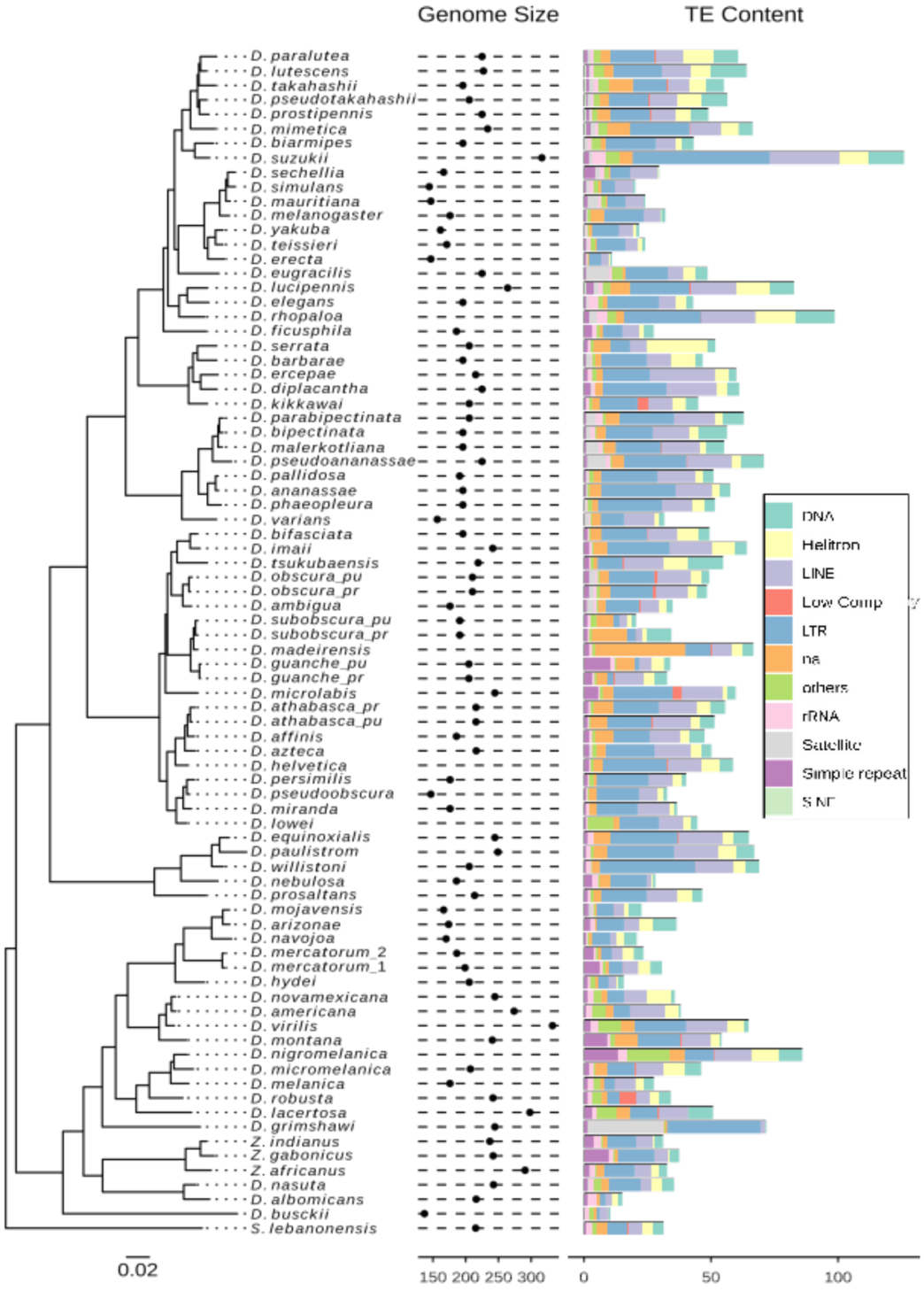


Figure ***7.1*** – *Phylogénie de 82 Drosophiles. La tailles des différents génomes est représenté au centre. Le contenu en éléments transposables est résumé à droite pour chaque génomes.*

7.2 Étude de l'évolution des Cannabaceae

Ce projet cannabis/houblon s'inscrit dans la thèse de Djivan Prentout un doctorant du LBBE. Djivan travaille sur l'évolution des chromosomes sexuels chez les plantes. Il a aussi le sujet de thèse le plus sexy depuis l'invention de la science puisqu'il a travaillé sur le sexe chez *Cannabis sativa*, *Humulus lupulus* ou encore *Vitis vinifera* (Cannabis, Bière et Vin). Cette collaboration a été rendue possible grâce (à cause ?) du Coronavirus qui a grandement vidé le LBBE pendant l'été 2020. Le projet en question portait sur la description d'une vieille paire de chromosomes sexuels homologues entre deux espèces qui ont divergé il y a plus de 20 millions d'années, à savoir, *Cannabis sativa* et *Humulus lupulus*.

J'ai eu pour tache de reconstruire les arbres phylogénétiques des gènes, précédemment identifiés par Djivan comme étant liés au sexe chez le cannabis et le houblon (Figure 7.2). Ce travail s'est conclu par la rédaction d'un article par Djivan "Plant genera Cannabis and Humulus share the same pair of well differentiated sex chromosomes" publié dans la revue *New Phytologist* (Prentout *et al.*, 2021). L'article complet est accessible à : https ://nph.onlinelibrary.wiley.com/doi/10.1111/nph.17456.

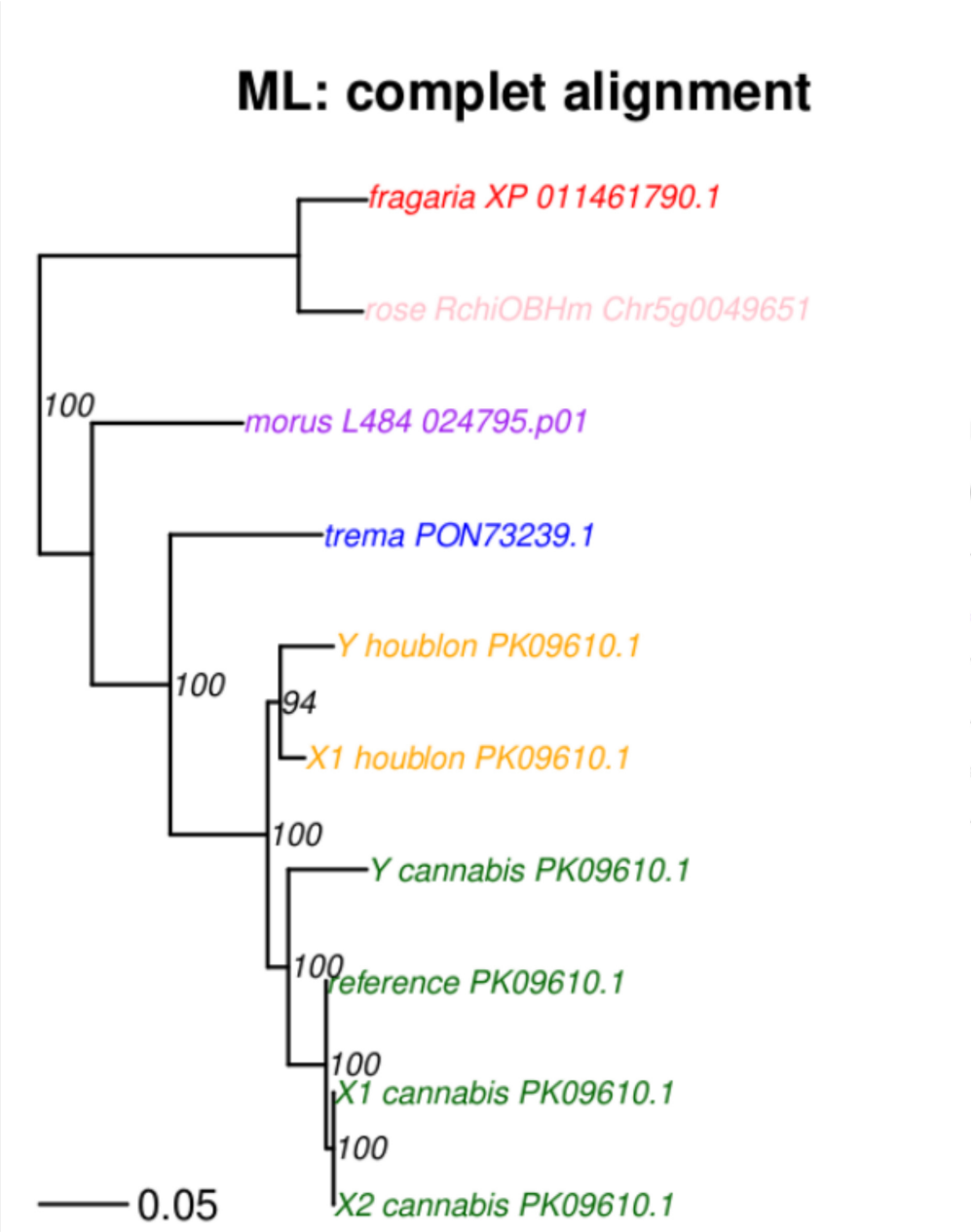


Figure 7.2 – *Phylogénie en maximum de vraisemblance d'une famille de gène de Cannabaceae.*

7.3 Une méthode de détection de flux de gènes pour étudier l'histoire évolutive des langues

J'ai mené un autre projet pendant la crise du Coronavirus. En février 2020, Dan Dediu (laboratoire Dynamique Du Langage, Université Lyon 2) à donné un séminaire au LBBE intitulé "Language and speech are evolutionary systems : the

influence of our own biology on language evolution and diversity". Pour comprendre et construire l'histoire des langues, de nombreux auteurs se sont inspirés des méthodes de la biologie évolutive. En particulier la reconstruction sous la forme d'arbre, comme l'arbre du vivant. En 2016, Gerhard Jäger et Søren Wichmann ont proposé un arbre global des langues et trouvent des preuves que les familles linguistiques du monde entier sont liées les unes aux autres de manière géographiquement cohérente, se regroupant en continents et reflétant même des événements assez spécifiques de l'histoire humaine. Plusieurs autres travaux ont été menés pour reconstruire des arbres des différentes langues et dialectes (Serva and Petroni, 2008; Kolipakam *et al.*, 2018). De plus, l'emprunt lexical (lexical borrowings) est un aspect primordial à l'évolution des langues (Pasquini and Serva, 2019). Ce phénomène s'apparente à un transfert horizontal de traits linguistiques d'une langue à une autre. L'idée est d'utiliser les méthodes de détection des flux de gènes et de les appliquer aux données linguistiques pour tenter d'identifier des évènements de transferts horizontaux entre les langues. Il est à noter que l'utilisation des arbres pour représenter l'évolution des langues est un sujet très sensible dans la communauté des linguistes. Beaucoup sont contre l'idée de représenter l'évolution des langues sous la forme d'un arbre et la notion qu'elle implique d'une origine unique de toutes les langues. Nous avons néanmoins travaillé avec Dan Dediu et Marc Tang (à l'époque en post-Doc, désormais CR CNRS au Muséum National d'Histoire Naturelle) pour tenter d'appliquer le test ABBA-BABA sur des données de langues.

Les traits linguistiques sont classé dans 3 catégories : lexique, phonologie, morphosyntaxe. Chacun des ces traits peuvent être résumés sous la forme de "cognate" binaires représentant la présence ou l'absence du trait dans les langue considérées. Prenons un trait lexical par exemple, c'est-à-dire que pour un "mot", on liste les origines pour les résumer en une variable qualitative. Pour les mots avec le sens de "TOUT" (la notion d'ensemble), le français et l'italien ont la même origine latine et l'anglais et l'allemand une origine Proto-Germanic 7.4. Du coup français et anglais on le même 'cognate', et cette variable est ensuite changée en variable binaire. Cette variable binaire peut être lu comme un motif ABBA ou BABA si il y a eu échange de trait.

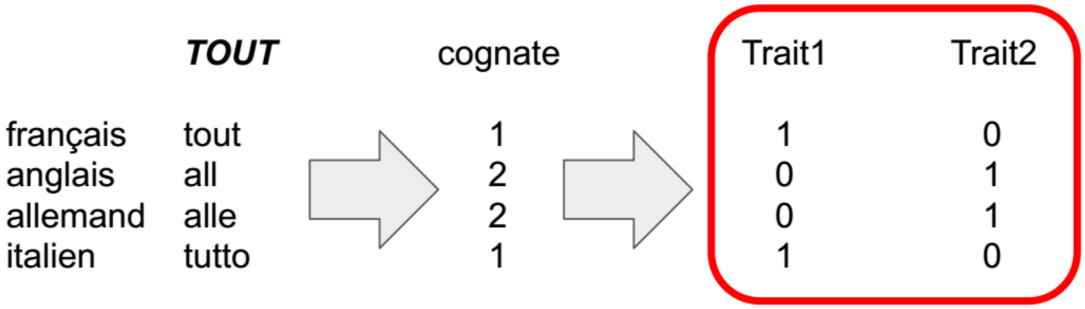
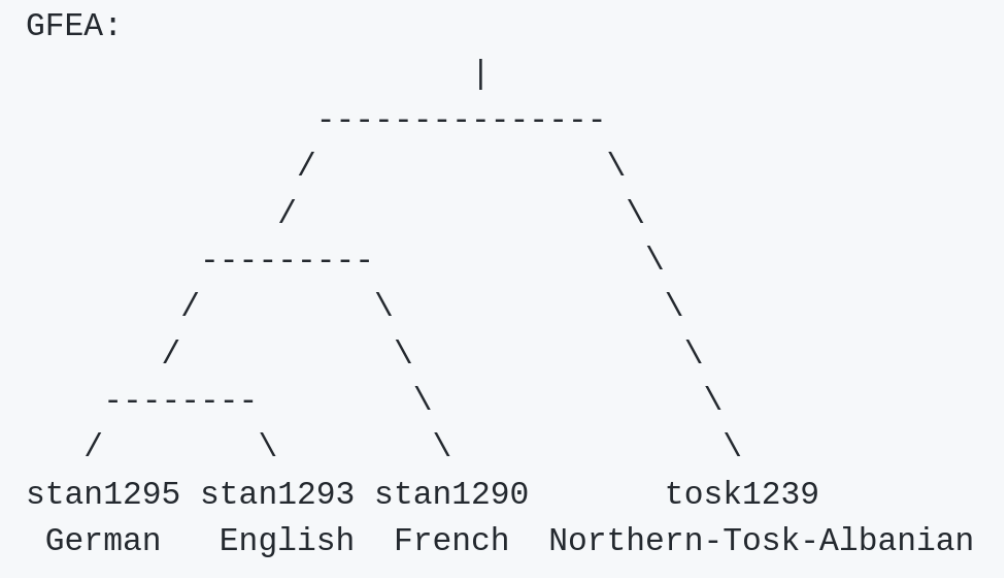


Figure 7.3 – *Comment passer d'un mot à un motif ABBA-BABA. La notion de "Tout" est transformé en cognate. Le français et l'italien partage la même origine, "tout" et "tutto" du latin "tottus", cognate 1. Alors que le "all" anglais est "alle" allemand ont une origine proto-germanique, cognate 2. Ces deux cognates peuvent être résumés en traits binaires "1001" et "0110", très similaire au motif ABBA utilisé par la statistique-D.*

Pour valider la méthode, j'ai tenté de retrouver le contact, ou échange, lexical entre le français et l'anglais. L'arbre de quatre langue utilisé est composé de l'allemand, du français, de l'anglais et en groupe extérieur l'albanais avec la topologie représentée dans la figure 7.4. Une statistique-D a été calculée pour chaque catégorie de trait ainsi que pour l'ensemble de 783 traits disponibles. Un test binomial est utilisé pour tester si la statistique-D est significative. Le seul D significativement différent de zéro est celui calculé sur les traits Lexicon (Figure 7.4), ce qui suggère bien un échange de traits lexicaux entre le français et l'anglais. Il ne semble pas y avoir de trace d'échange dans les autres catégories de traits. De façon plus surprenante lorsque tous les traits sont utilisé pour calculer le D alors on ne retrouve plus de flux entre les langues. Il est difficile de dire si ce qu'on détecte ici (ou ne détecte pas) est véritablement représentatif de l'histoire. La très faible quantité des traits, seulement 783, ne permet pas un calcul très robuste, surtout que seuls 50 traits sont sous la forme ABBA ou BABA. Néanmoins pour nos collaborateurs linguistes, ces résultats étaient très encourageants. Notre collaboration n'a pas continué après le premier confinement dû au COVID et par manque de temps, mais il serait intéressant de pouvoir continuer ce projet assez hors du commun et peut être détecter des flux de traits linguistiques encore inconnus. Ce projet est en pause actuellement. Marc Tang ayant changé de laboratoire, il est possible que le projet ne reprenne jamais.



DATA	Feat Number	ABBA	BABA	D statistic	Pvalue (binom.test())
Lexicon	579	17	5	0.545454545454545	0.0169005393981934
Morphosyntax	116	7	13	-0.3	0.263175964355469
Phono	88	1	7	-0.75	0.0703125
all	783	25	25	0	1

Figure 7.4 – *Topologies d'un quartet de langues composé de l'allemand, du français, de l'anglais et de l'albanais. La table résume la quantité de données disponible pour chaque catégorie de trait, ainsi que le nombre de cognates qui ont pu être résumé en motif de type ABBA et BABA. Pour chaque catégorie, une statistique-D a été calculée et un test binomial utilisé pour tester si le D est significativement différent de zéro (seuil de 0.05). Enfin une dernière statistique-D a été calculé sur l'ensemble des donnés.*

7.4 Impact de la distribution du taux de recombinaison à fine échelle sur la dynamique du paysage régulateur chez les carnivores.

Ce projet canidés s'inscrit dans la thèse de Julien Joseph, doctorant de deuxième année au LBBE, sous la direction de Laurent Duret. L'ensemble du projet a été pensé par Julien, avec pour but premier de travailler avec un grand groupe de doctorant du LBBE et comme but secondaire t'étudier les paysages de recombinaison chez les carnivores.

La perte de PRDM9 chez les canidés a entraîné une stabilisation des points

chauds de recombinaison, et donc une accumulation de substitution vers GC dans ces points chauds. Ces points chauds se sont enrichis en dinucléotides CpG, et sont reconnus par les algorithmes cherchant à détecter des îlots CpGs. Il a été montré que les dinucléotides CpG (methylés ou pas) avaient des rôles complexes dans la régulation de l'expression des gènes. Nous cherchons donc à quantifier les gains/pertes d'îlots CpGs le long de la phylogénie des carnivores en voyant si l'on obtient des différences entre les canidés d'une part, et les félidés et les phocidés qui ont toujours un PRDM9 fonctionnel et donc des points chauds instables, de l'autre. L'attendu étant que chez les canidés les gains soient plus importants, et les pertes plus faibles que chez les Félidés et les Phocidés. Une deuxième étape serait d'explorer si il existe une corrélation entre l'évolution de ces séquences et l'évolution de l'expression des gènes qui se situent au voisinage.

Dans ce projet, mon rôle a été de reconstruire la phylogénie de 19 génomes de carnivores. Il m'a fallu pour cela annoter ou ré-annoter 5 génomes de carnivores. Pour la reconstruction phylogénétique j'ai utilisé les alignements de 1655 familles de gènes universel et unicopies. La topologie de l'arbre est globalement concordante avec celles qu'on peut trouver dans la littérature. Le branchement de *Otocyon megalotis* (Renard à oreilles de chauve-souris) comme espèces frère des *Canini* (tribu de carnivores caniformes composée des chiens et des loups) ne correspond pas à ce qui peut être observé dans d'autre étude. Il est généralement retrouvé entre les *Vulpes* et *Nyctereutes procyonoides*. Ce projet est toujours en cours et devrait mener à la publication d'un article par Julien.

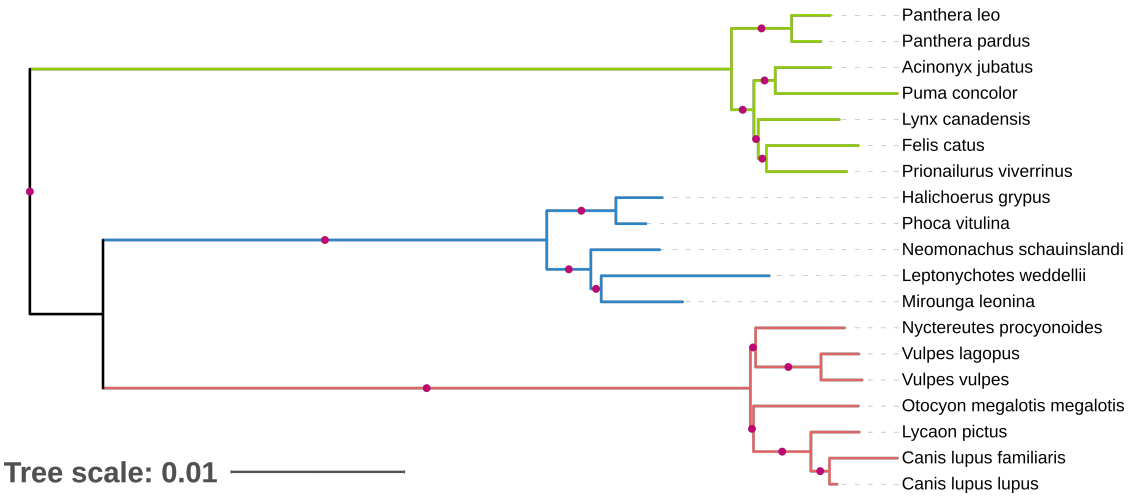


Figure 7.5 – *Phylogenie des carnivores reconstruite avec 1655 marqueur universel et unicopies. Les ronds pourpres représentent un support bootstraps de 100. Les branches rouges représentent le groupe des canidés, en bleu les phocidés et en vert les félidés.*

Bibliographie

Abby, S. S., Tannier, E., Gouy, M., and Daubin, V. 2010. Detecting lateral gene transfers by statistical reconciliation of phylogenetic forests. *BMC bioinformatics*, 11(1) : 1–13. *Cited at page 116*

Abby, S. S., Tannier, E., Gouy, M., and Daubin, V. 2012. Lateral gene transfer as a support for the tree of life. *Proceedings of the National Academy of Sciences*, 109(13) : 4962–4967. *Cited at page 33*

Adato, O., Ninyo, N., Gophna, U., and Snir, S. 2015. Detecting Horizontal Gene Transfer between Closely Related Taxa. *PLOS Computational Biology*, 11(10) : e1004408. *Cited at page 30*

Allen, R., Rittmann, B. E., and Curtiss III, R. 2019. Axenic biofilm formation and aggregation by synechocystis sp. strain pcc 6803 are induced by changes in nutrient concentration and require cell surface structures. *Applied and environmental microbiology*, 85(7) : e02192–18. *Cited at page 16*

Alroy, J. 2002. How many named species are valid ? *Proceedings of the National Academy of Sciences*, 99(6) : 3706–3711. *Cited at page 16*

Andersson, G., Karlberg, O., Canbäck, B., and Kurland, C. G. 2003. On the origin of mitochondria : a genomics perspective. *Philosophical Transactions of the Royal Society of London. Series B : Biological Sciences*, 358(1429) : 165–179. *Cited at page 23*

Arenas, M. and Posada, D. 2014. Simulation of genome-wide evolution under heterogeneous substitution models and complex multispecies coalescent histories. *Molecular biology and evolution*, 31(5) : 1295–1301. *Cited at page 121*

Avery, O. T., MacLeod, C. M., and McCarty, M. 1944. Studies on the chemical nature of the substance inducing transformation of pneumococcal types induction of transformation by a desoxyribonucleic acid fraction isolated from pneumococcus type iii. *Journal of experimental medicine*, 79(2) : 137–158. *Cited at page 21*

Bansal, M. S., Kellis, M., Kordi, M., and Kundu, S. 2018. Ranger-dtl 2.0 : rigorous reconstruction of gene-family evolution by duplication, transfer and loss. *Bioinformatics*, 34(18) : 3214–3216. *Cited at page 28*

Becq, J., Churlaud, C., and Deschavanne, P. 2010. A benchmark of parametric methods for horizontal transfers detection. *PLoS One*, 5(4) : e9989. *Cited at page 24*

Beiko, R. G. and Charlebois, R. L. 2007. A simulation test bed for hypotheses of genome evolution. *Bioinformatics*, 23(7) : 825–831. *Cited at page 37*

Beiko, R. G., Harlow, T. J., and Ragan, M. A. 2005. Highways of gene sharing in prokaryotes. *Proceedings of the National Academy of Sciences*, 102(40) : 14332–14337. *Cited at page 114*

Bhattacharya, D. and Medlin, L. 1995. The phylogeny of plastids : a review based on comparisons of small-subunit ribosomal rna coding regions. *Journal of Phycology*, 31(4) : 489–498. *Cited at page 23*

Biller, P., Knibbe, C., Beslon, G., and Tannier, E. 2016. Comparative genomics on artificial life. In *Conference on Computability in Europe*, pages 35–44. Springer. *Cited at page 34*

Bohlin, J., Snipen, L., Hardy, S. P., Kristoffersen, A. B., Lagesen, K., Dønsvik, T., Skjerve, E., and Ussery, D. W. 2010. Analysis of intra-genomic gc content homogeneity within prokaryotes. *BMC genomics*, 11(1) : 1–8. *Cited at page 119*

Boto, L. 2010. Horizontal gene transfer in evolution : facts and challenges. *Proceedings of the Royal Society B : Biological Sciences*, 277(1683) : 819–827. *Cited at pages 21, 33*

Brown, C. T., Hug, L. A., Thomas, B. C., Sharon, I., Castelle, C. J., Singh, A., Wilkins, M. J., Wrighton, K. C., Williams, K. H., and Banfield, J. F. 2015. Unusual biology across a group comprising more than 15% of domain bacteria. *Nature*, 523(7559) : 208–211. *Cited at page 15*

Burgarella, C., Barnaud, A., Kane, N. A., Jankowski, F., Scarcelli, N., Billot, C., Vigouroux, Y., and Berthouly-Salazar, C. 2019. Adaptive Introgression : An Untapped Evolutionary Mechanism for Crop Adaptation. *Frontiers in Plant Science*, 10 : 4. *Cited at page 41*

Ceballos, G., Ehrlich, P. R., and Dirzo, R. 2017. Biological annihilation via the ongoing sixth mass extinction signaled by vertebrate population losses and

declines. *Proceedings of the national academy of sciences*, 114(30) : E6089–E6096. *Cited at page 11*

Chapman, A. D. 2009. *Numbers of living species in Australia and the world*. Department of the Environment Water Heritage and the Arts, Canberra A.C.T. *Cited at pages 12, 13*

Chauve, C., Rafiey, A., Davín, A. A., Scornavacca, C., Veber, P., Boussau, B., Szöllősi, G. J., Daubin, V., and Tannier, E. 2017. MaxTiC : Fast ranking of a phylogenetic tree by Maximum Time Consistency with lateral gene transfers. *bioRxiv*, page 127548. *Cited at page 116*

Chen, K., Durand, D., and Farach-Colton, M. 2000. Notung : a program for dating gene duplications and optimizing gene family trees. *Journal of Computational Biology*, 7(3-4) : 429–447. *Cited at page 28*

Christenhusz, M. J. and Byng, J. W. 2016. The number of known plants species in the world and its annual increase. *Phytotaxa*, 261(3) : 201–217. *Cited at page 11*

Costello, M. J., May, R. M., and Stork, N. E. 2013. Can we name earth's species before they go extinct ? *science*, 339(6118) : 413–416. *Cited at page 16*

Dalquen, D. A., Anisimova, M., Gonnet, G. H., and Dessimoz, C. 2012. ALF—A Simulation Framework for Genome Evolution. *Molecular Biology and Evolution*, 29(4) : 1115–1123. *Cited at pages 37, 121*

Dannemann, M. and Racimo, F. 2018. Something old, something borrowed : admixture and adaptation in human evolution. *Current Opinion in Genetics & Development*, 53 : 1–8. *Cited at pages 23, 41*

Daubin, V., Lerat, E., and Perrière, G. 2003. The source of laterally transferred genes in bacterial genomes. *Genome biology*, 4(9) : 1–12. *Cited at page 24*

David, L. A. and Alm, E. J. 2011. Rapid evolutionary innovation during an archaean genetic expansion. *Nature*, 469(7328) : 93–96. *Cited at page 28*

Davín, A. A., Tannier, E., Williams, T. A., Boussau, B., Daubin, V., and Szöllősi, G. J. 2018. Gene transfers can date the tree of life. *Nature Ecology & Evolution*, 2(5) : 904–909. *Cited at pages 33, 114, 121*

Davín, A. A., Tricou, T., Tannier, E., de Vienne, D. M., and Szöllősi, G. J. 2020. Zombi : a phylogenetic simulator of trees, genomes and sequences that accounts for dead linages. *Bioinformatics*, 36(4) : 1286–1288. *Cited at pages 35, 37, 98, 112, 116*

Devos, D. P. *et al.* 2012. Evaluating the evolutionary origins of unexpected character distributions within the bacterial planctomycetes-verrucomicrobia-chlamydiae superphylum. *Frontiers in microbiology*, 3 : 401. *Cited at page 120*

Dodd, M. S., Papineau, D., Grenne, T., Slack, J. F., Rittner, M., Pirajno, F., O'Neil, J., and Little, C. T. 2017. Evidence for early life in earth's oldest hydrothermal vent precipitates. *Nature*, 543(7643) : 60–64. *Cited at page 16*

Donoghue, M. J., Doyle, J. A., Gauthier, J., Kluge, A. G., and Rowe, T. 1989. The importance of fossils in phylogeny reconstruction. *Annual review of Ecology and Systematics*, 20(1) : 431–460. *Cited at page 97*

Durand, E. Y., Patterson, N., Reich, D., and Slatkin, M. 2011. Testing for Ancient Admixture between Closely Related Populations. *Molecular Biology and Evolution*, 28(8) : 2239–2252. *Cited at page 26*

Eaton, D. A. R. and Ree, R. H. 2013. Inferring Phylogeny and Introgression using RADseq Data : An Example from Flowering Plants (Pedicularis : Orobanchaceae). *Systematic Biology*, 62(5) : 689–706. *Cited at page 41*

Edelman, N. B., Frandsen, P. B., Miyagi, M., Clavijo, B., Davey, J., Dikow, R. B., García-Accinelli, G., Belleghem, S. M. V., Patterson, N., Neafsey, D. E., Challis, R., Kumar, S., Moreira, G. R. P., Salazar, C., Chouteau, M., Counterman, B. A., Papa, R., Blaxter, M., Reed, R. D., Dasmahapatra, K. K., Kronforst, M., Joron, M., Jiggins, C. D., McMillan, W. O., Palma, F. D., Blumberg, A. J., Wakeley, J., Jaffe, D., and Mallet, J. 2019. Genomic architecture and introgression shape a butterfly radiation. *Science*, 366(6465) : 594–599. *Cited at pages 23, 41*

Eyres, I., Boschetti, C., Crisp, A., Smith, T. P., Fontaneto, D., Tunnacliffe, A., and Barraclough, T. G. 2015. Horizontal gene transfer in bdelloid rotifers is ancient, ongoing and more frequent in species from desiccating habitats. *BMC biology*, 13(1) : 1–17. *Cited at page 22*

Fontaine, M. C., Pease, J. B., Steele, A., Waterhouse, R. M., Neafsey, D. E., Sharakhov, I. V., Jiang, X., Hall, A. B., Catteruccia, F., Kakani, E., Mitchell, S. N., Wu, Y.-C., Smith, H. A., Love, R. R., Lawniczak, M. K., Slotman, M. A., Emrich, S. J., Hahn, M. W., and Besansky, N. J. 2015. Extensive introgression in a malaria vector species complex revealed by phylogenomics. *Science*, 347(6217) : 1258524. *Cited at pages 30, 119*

Ford, A. G. P., Dasmahapatra, K. K., Rüber, L., Gharbi, K., Cezard, T., and Day, J. J. 2015. High levels of interspecific gene flow in an endemic cichlid

fish adaptive radiation from an extreme lake environment. *Molecular Ecology*, 24(13) : 3421–3440. *Cited at page 112*

Forsythe, E. S., Nelson, A. D. L., and Beilstein, M. A. 2020. Biased gene retention in the face of introgression obscures species relationships. *Genome Biology and Evolution*. *Cited at page 29*

Friedman, R. and Ely, B. 2012. Codon usage methods for horizontal gene transfer detection generate an abundance of false positive and false negative results. *Current microbiology*, 65(5) : 639–642. *Cited at page 119*

Froese, R., Pauly, D., Roskov, Y., Ower, G., Orrell, T., Nicolson, D., Bailly, N., Kirk, P., and Penev, L. 2019. Species 2000 & itis catalogue of life. 2019 annual checklist. *Cited at page 11*

Galtier, N. and Daubin, V. 2008. Dealing with incongruence in phylogenomic analyses. *Philosophical Transactions of the Royal Society B : Biological Sciences*, 363(1512) : 4023–4029. *Cited at pages 34, 97*

Gogarten, J. P., Fournier, G., and Zhaxybayeva, O. 2008. Gene Transfer and the Reconstruction of Life's Early History from Genomic Data. *Space Science Reviews*, 135(1-4) : 115–131. *Cited at page 32*

Green, R. E., Krause, J., Briggs, A. W., Maricic, T., Stenzel, U., Kircher, M., Patterson, N., Li, H., Zhai, W., Fritz, M. H.-Y., Hansen, N. F., Durand, E. Y., Malaspinas, A.-S., Jensen, J. D., Marques-Bonet, T., Alkan, C., Prüfer, K., Meyer, M., Burbano, H. A., Good, J. M., Schultz, R., Aximu-Petri, A., Butthof, A., Höber, B., Höffner, B., Siegemund, M., Weihmann, A., Nusbaum, C., Lander, E. S., Russ, C., Novod, N., Affourtit, J., Egholm, M., Verna, C., Rudan, P., Brajkovic, D., Kucan, Z., Gusic, I., Doronichev, V. B., Golovanova, L. V., Lalueza-Fox, C., Rasilla, M. d. l., Fortea, J., Rosas, A., Schmitz, R. W., Johnson, P. L. F., Eichler, E. E., Falush, D., Birney, E., Mullikin, J. C., Slatkin, M., Nielsen, R., Kelso, J., Lachmann, M., Reich, D., and Pääbo, S. 2010. A Draft Sequence of the Neandertal Genome. *Science*, 328(5979) : 710–722. *Cited at pages 23, 26*

Griffith, F. 1928. The significance of pneumococcal types. *Epidemiology & Infection*, 27(2) : 113–159. *Cited at page 21*

Guindon, S. and Perriere, G. 2001. Intragenomic base content variation is a potential source of biases when searching for horizontally transferred genes. *Molecular biology and evolution*, 18(9) : 1838–1840. *Cited at page 24*

Guy, L. and Ettema, T. J. 2011. The archaeal 'tack'superphylum and the origin of eukaryotes. *Trends in microbiology*, 19(12) : 580–587. *Cited at page 11*

Hahn, M. W. and Hibbins, M. S. 2019. A Three-Sample Test for Introgression. *Molecular Biology and Evolution*, 36(12) : 2878–2882. *Cited at pages 29, 31*

Hall, J. P., Brockhurst, M. A., and Harrison, E. 2017. Sampling the mobile gene pool : innovation via horizontal gene transfer in bacteria. *Philosophical Transactions of the Royal Society B : Biological Sciences*, 372(1735) : 20160424. *Cited at pages 21, 33*

Hallam, A. and Wignall, P. B. 1997. *Mass extinctions and their aftermath*. Oxford University Press, UK. *Cited at page 112*

Harrison, R. G. and Larson, E. L. 2014. Hybridization, introgression, and the nature of species boundaries. *Journal of Heredity*, 105(S1) : 795–809. *Cited at page 22*

Hinchliff, C. E., Smith, S. A., Allman, J. F., Burleigh, J. G., Chaudhary, R., Coghill, L. M., Crandall, K. A., Deng, J., Drew, B. T., Gazis, R., *et al.* 2015. Synthesis of phylogeny and taxonomy into a comprehensive tree of life. *Proceedings of the National Academy of Sciences*, 112(41) : 12764–12769. *Cited at page 19*

Huang, D. I., Hefer, C. A., Kolosova, N., Douglas, C. J., and Cronk, Q. C. B. 2014. Whole plastome sequencing reveals deep plastid divergence and cytonuclear discordance between closely related balsam poplars, *Populus balsamifera* and *P. trichocarpa* (Salicaceae). *New Phytologist*, 204(3) : 693–703. *Cited at page 30*

Huang, J., Mullapudi, N., Sicheritz-Ponten, T., and Kissinger, J. C. 2004. A first glimpse into the pattern and scale of gene transfer in the apicomplexa. *International journal for parasitology*, 34(3) : 265–274. *Cited at page 22*

Huerta-Sánchez, E., Jin, X., Asan, Bianba, Z., Peter, B. M., Vinckenbosch, N., Liang, Y., Yi, X., He, M., Somel, M., Ni, P., Wang, B., Ou, X., Huasang, Luosang, J., Cuo, Z. X. P., Li, K., Gao, G., Yin, Y., Wang, W., Zhang, X., Xu, X., Yang, H., Li, Y., Wang, J., Wang, J., and Nielsen, R. 2014. Altitude adaptation in Tibetans caused by introgression of Denisovan-like DNA. *Nature*, 512(7513) : 194–197. *Cited at page 23*

Hug, L. A., Baker, B. J., Anantharaman, K., Brown, C. T., Probst, A. J., Castelle, C. J., Butterfield, C. N., Hernsdorf, A. W., Amano, Y., Ise, K., *et al.* 2016. A new view of the tree of life. *Nature microbiology*, 1(5) : 1–6. *Cited at pages 11, 15*

Husnik, F. and McCutcheon, J. P. 2018. Functional horizontal gene transfer from bacteria to eukaryotes. *Nature Reviews Microbiology*, 16(2) : 67–79. *Cited at page 22*

Jacox, E., Chauve, C., Szöllősi, G. J., Ponty, Y., and Scornavacca, C. 2016. ecceTERA : comprehensive gene tree-species tree reconciliation using parsimony. *Bioinformatics*, 32(13) : 2056–2058. *Cited at page 28*

Keeling, P. J. and Palmer, J. D. 2008. Horizontal gene transfer in eukaryotic evolution. *Nature Reviews Genetics*, 9(8) : 605–618. *Cited at pages 21, 22, 33*

Kolipakam, V., Jordan, F. M., Dunn, M., Greenhill, S. J., Bouckaert, R., Gray, R. D., and Verkerk, A. 2018. A bayesian phylogenetic study of the dravidian language family. *Royal Society open science*, 5(3) : 171504. *Cited at page 127*

Koonin, E. V., Makarova, K. S., and Aravind, L. 2001. Horizontal gene transfer in prokaryotes : quantification and classification. *Annual Reviews in Microbiology*, 55(1) : 709–742. *Cited at pages 21, 33*

Kundu, S. and Bansal, M. S. 2019. SaGePhy : an improved phylogenetic simulation framework for gene and subgene evolution. *Bioinformatics*, 35(18) : 3496–3498. *Cited at page 37*

Kunin, W. E. and Gaston, K. J. 2012. *The Biology of Rarity : Causes and consequences of rare—common differences*, volume 17. Springer Science & Business Media. *Cited at page 16*

Lawrence, J. G. and Ochman, H. 1997. Amelioration of bacterial genomes : rates of change and exchange. *Journal of molecular evolution*, 44(4) : 383–397. *Cited at page 119*

Lawson, D. J., van Dorp, L., and Falush, D. 2018. A tutorial on how not to over-interpret STRUCTURE and ADMIXTURE bar plots. *Nature Communications*, 9(1) : 3258. *Cited at page 120*

Lewis, W. H., Tahon, G., Geesink, P., Sousa, D. Z., and Ettema, T. J. 2021. Innovations to culturing the uncultured microbial majority. *Nature Reviews Microbiology*, 19(4) : 225–240. *Cited at page 13*

Locey, K. J. and Lennon, J. T. 2016. Scaling laws predict global microbial diversity. *Proceedings of the National Academy of Sciences*, 113(21) : 5970–5975. *Cited at page 13*

Maddison, W. P. 1997a. Gene trees in species trees. *Systematic biology*, 46(3) : 523–536. *Cited at page 34*

Maddison, W. P. 1997b. Gene Trees in Species Trees. *Systematic Biology*, 46(3) : 523–536. *Cited at page 97*

Mahé, K., Ernande, B., and Herbin, M. 2021. New scale analyses reveal centenarian African coelacanths. *Current Biology*, 0(0). *Cited at page 18*

Mallo, D., De Oliveira Martins, L., and Posada, D. 2016. SimPhy : Phylogenomic Simulation of Gene, Locus, and Species Trees. *Systematic Biology*, 65(2) : 334–344. *Cited at pages 36, 37, 121*

Martin, S. H., Dasmahapatra, K. K., Nadeau, N. J., Salazar, C., Walters, J. R., Simpson, F., Blaxter, M., Manica, A., Mallet, J., and Jiggins, C. D. 2013. Genome-wide evidence for speciation with gene flow in Heliconius butterflies. *Genome Research*, 23(11) : 1817–1828. *Cited at page 41*

Martin, S. H., Davey, J. W., and Jiggins, C. D. 2015. Evaluating the use of abba–baba statistics to locate introgressed loci. *Molecular biology and evolution*, 32(1) : 244–257. *Cited at page 120*

Martin Kuhlwilm 2020. Resurrection of the Ghosts. *BioEssays*, 42(6) : 2000057. *Cited at page 121*

Martin Kuhlwilm, Han, S., Sousa, V. C., Excoffier, L., and Marques-Bonet, T. 2019. Ancient admixture from an extinct ape lineage into bonobos. *Nature Ecology & Evolution*, 3(6) : 957–965. *Cited at page 34*

May, R. M. 2010. Tropical arthropod species, more or less ? *Science*, 329(5987) : 41–42. *Cited at page 13*

McGill, B. J., Dornelas, M., Gotelli, N. J., and Magurran, A. E. 2015. Fifteen forms of biodiversity trend in the anthropocene. *Trends in ecology & evolution*, 30(2) : 104–113. *Cited at page 11*

Menet, H., Daubin, V., and Tannier, E. 2021. Phylogenetic reconciliation. *HAL*. *Cited at page 27*

Meyer, M., Kircher, M., Gansauge, M.-T., Li, H., Racimo, F., Mallick, S., Schraiber, J. G., Jay, F., Prüfer, K., Filippo, C. d., Sudmant, P. H., Alkan, C., Fu, Q., Do, R., Rohland, N., Tandon, A., Siebauer, M., Green, R. E., Bryc, K., Briggs, A. W., Stenzel, U., Dabney, J., Shendure, J., Kitzman, J., Hammer, M. F.,

Shunkov, M. V., Derevianko, A. P., Patterson, N., Andrés, A. M., Eichler, E. E., Slatkin, M., Reich, D., Kelso, J., and Pääbo, S. 2012. A High-Coverage Genome Sequence from an Archaic Denisovan Individual. *Science*, 338(6104) : 222–226. *Cited at page 41*

Mora, C., Tittensor, D. P., Adl, S., Simpson, A. G. B., and Worm, B. 2011. How Many Species Are There on Earth and in the Ocean ? *PLoS Biology*, 9(8) : e1001127. *Cited at pages 11, 13, 14*

Morel, B., Kozlov, A. M., Stamatakis, A., and Szöllősi, G. J. 2019. GeneRax : A tool for species tree-aware maximum likelihood based gene family tree inference under gene duplication, transfer, and loss. preprint, Bioinformatics. *Cited at pages 33, 114*

Narasingarao, P., Podell, S., Ugalde, J. A., Brochier-Armanet, C., Emerson, J. B., Brocks, J. J., Heidelberg, K. B., Banfield, J. F., and Allen, E. E. 2012. De novo metagenomic assembly reveals abundant novel major lineage of archaea in hypersaline microbial communities. *The ISME journal*, 6(1) : 81–93. *Cited at page 15*

Norell MA 1992. Taxic Origin and Temporal Diversity : The Effect of Phylogeny, in Extinction and Phylogeny. *Cited at page 18*

Novichkov, P. S., Omelchenko, M. V., Gelfand, M. S., Mironov, A. A., Wolf, Y. I., and Koonin, E. V. 2004. Genome-wide molecular clock and horizontal gene transfer in bacterial evolution. *Journal of bacteriology*, 186(19) : 6575–6585. *Cited at page 30*

Nunoura, T., Takaki, Y., Kakuta, J., Nishi, S., Sugahara, J., Kazama, H., Chee, G.-J., Hattori, M., Kanai, A., Atomi, H., *et al.* 2011. Insights into the evolution of archaea and eukaryotic protein modifier systems revealed by the genome of a novel archaeal group. *Nucleic acids research*, 39(8) : 3204–3223. *Cited at page 15*

Ochman, H., Lawrence, J. G., and Groisman, E. A. 2000. Lateral gene transfer and the nature of bacterial innovation. *nature*, 405(6784) : 299–304. *Cited at page 22*

Ottenburghs, J. 2020. Ghost Introgression : Spooky Gene Flow in the Distant Past. *BioEssays*, 42(6) : 2000012. *Cited at pages 34, 41, 121*

Parr, C. S., Wilson, M. N., Leary, M. P., Schulz, K. S., Lans, M. K., Walley, M. L., Hammock, J. A., Goddard, M. A., Rice, M. J., Studer, M. M., *et al.* 2014. The encyclopedia of life v2 : providing global access to knowledge about life on earth. *Biodiversity data journal*, 2. *Cited at page 11*

Pasquini, M. and Serva, M. 2019. Horizontal transfers are a primary aspect of languages evolution. *EPL (Europhysics Letters)*, 125(3) : 38002. *Cited at page 127*

Patterson, N., Moorjani, P., Luo, Y., Mallick, S., Rohland, N., Zhan, Y., Genschoreck, T., Webster, T., and Reich, D. 2012. Ancient Admixture in Human History. *Genetics*, 192(3) : 1065–1093. *Cited at page 26*

Pittis, A. A. and Gabaldón, T. 2016. Late acquisition of mitochondria by a host with chimaeric prokaryotic ancestry. *Nature*, 531(7592) : 101–104. *Cited at pages 23, 30*

Poole, A. M. and Gribaldo, S. 2014. Eukaryotic origins : How and when was the mitochondrion acquired ? *Cold Spring Harbor perspectives in biology*, 6(12) : a015990. *Cited at page 23*

Prentout, D., Stajner, N., Cerenak, A., Tricou, T., Brochier-Armanet, C., Jakse, J., Käfer, J., and Marais, G. A. B. 2021. Plant genera Cannabis and Humulus share the same pair of well-differentiated sex chromosomes. *New Phytologist*, n/a(n/a). *Cited at page 125*

Prothero, D. R. 2013. *Bringing fossils to life : an introduction to paleobiology*. Columbia University Press, New York, third edition edition. *Cited at page 16*

Raup, D. M. 1986. Biological extinction in earth history. *Science*, 231(4745) : 1528–1533. *Cited at page 16*

Raup, D. M. 1991. *Extinction : bad genes or bad luck ?* W.W. Norton, New York. *Cited at page 16*

Raup, D. M. and Sepkoski, J. J. 1982. Mass Extinctions in the Marine Fossil Record. *Science*, 215(4539) : 1501–1503. *Cited at page 112*

Ravenhall, M., Škunca, N., Lassalle, F., and Dessimoz, C. 2015. Inferring horizontal gene transfer. *PLoS computational biology*, 11(5) : e1004095. *Cited at pages 24, 25*

Roskov, Y., Abucay, L., Orrell, T., Nicolson, D., Bailly, N., Kirk, P., Bourgoin, T., DeWalt, R., Decock, W., De Wever, A., *et al.* 2000. Species 2000 & itis catalogue of life. *2019 Annual Checklist. Digital resource at www. catalogueoflife. org/annual-checklist/2019. Species*, pages 2405–884. *Cited at page 11*

Rouard, M., Droc, G., Martin, G., Sardos, J., Hueber, Y., Guignon, V., Cenci, A., Geigle, B., Hibbins, M. S., Yahiaoui, N., Baurens, F.-C., Berry, V., Hahn, M. W., D'Hont, A., and Roux, N. 2018. Three New Genome Assemblies Support a Rapid

Radiation in Musa acuminata (Wild Banana). *Genome Biology and Evolution*, 10(12) : 3129–3140. *Cited at page 41*

Schopf, J. W. 2012. The fossil record of cyanobacteria. In *Ecology of cyanobacteria II*, pages 15–36. Springer. *Cited at page 16*

Schumer, M., Cui, R., Powell, D. L., Rosenthal, G. G., and Andolfatto, P. 2016. Ancient hybridization and genomic stabilization in a swordtail fish. *Molecular Ecology*, 25(11) : 2661–2679. *Cited at page 41*

Serva, M. and Petroni, F. 2008. Indo-european languages tree by levenshtein distance. *EPL (Europhysics Letters)*, 81(6) : 68005. *Cited at page 127*

Sjöstrand, J., Arvestad, L., Lagergren, J., and Sennblad, B. 2013. GenPhyloData : realistic simulation of gene family evolution. *BMC Bioinformatics*, 14(1) : 209. *Cited at pages 37, 121*

Spang, A., Saw, J. H., Jørgensen, S. L., Zaremba-Niedzwiedzka, K., Martijn, J., Lind, A. E., Van Eijk, R., Schleper, C., Guy, L., and Ettema, T. J. 2015. Complex archaea that bridge the gap between prokaryotes and eukaryotes. *Nature*, 521(7551) : 173–179. *Cited at page 15*

Strassert, J. F., Irisarri, I., Williams, T. A., and Burki, F. 2021. A molecular timescale for eukaryote evolution with implications for the origin of red algal-derived plastids. *Nature communications*, 12(1) : 1–13. *Cited at page 23*

Susko, E., Steel, M., and Roger, A. J. 2021. Conditions under which distributions of edge length ratios on phylogenetic trees can be used to order evolutionary events. *bioRxiv*, page 2021.01.16.426961. *Cited at page 30*

Suvorov, A., Kim, B. Y., Wang, J., Armstrong, E. E., Peede, D., D'Agostino, E. R. R., Price, D. K., Wadell, P., Lang, M., Courtier-Orgogozo, V., David, J. R., Petrov, D., Matute, D. R., Schrider, D. R., and Comeault, A. A. 2020. Widespread introgression across a phylogeny of 155 Drosophila genomes. *bioRxiv*, page 2020.12.14.422758. *Cited at page 30*

Syvanen, M. 1985. Cross-species gene transfer ; implications for a new theory of evolution. *Journal of theoretical Biology*, 112(2) : 333–343. *Cited at pages 21, 33*

Szöllősi, G. J., Boussau, B., Abby, S. S., Tannier, E., and Daubin, V. 2012. Phylogenetic modeling of lateral gene transfer reconstructs the pattern and relative timing of speciations. *Proceedings of the national academy of sciences*, 109(43) : 17513–17518. *Cited at pages 114, 116*

Szöllősi, G. J., Rosikiewicz, W., Boussau, B., Tannier, E., and Daubin, V. 2013. Efficient exploration of the space of reconciled gene trees. *Systematic biology*, 62(6) : 901–912. *Cited at page 114*

Szöllősi, G. J., Tannier, E., Lartillot, N., and Daubin, V. 2013. Lateral Gene Transfer from the Dead. *Systematic Biology*, 62(3) : 386–397. *Cited at page 114*

Szöllősi, G. J., Davín, A. A., Tannier, E., Daubin, V., and Boussau, B. 2015. Genome-scale phylogenetic analysis finds extensive gene transfer among fungi. *Philosophical Transactions of the Royal Society B : Biological Sciences*, 370(1678) : 20140335. *Cited at pages 21, 26, 27, 33, 113, 114, 121*

Szöllősi, G. J., Höhna, S., Williams, T. A., Schrempf, D., Daubin, V., and Boussau, B. 2021. Relative time constraints improve molecular dating. *bioRxiv*, page 2020.10.17.343889. *Cited at page 114*

Tannier, E., Boussau, B., and Daubin, V. 2019. Quand les branches de l'arbre du vivant s' entremêlent. *Pour la science*. *Cited at page 29*

Thomas, C. M. and Nielsen, K. M. 2005. Mechanisms of, and barriers to, horizontal gene transfer between bacteria. *Nature reviews microbiology*, 3(9) : 711–721. *Cited at pages 11, 21*

Tricou, T., Tannier, E., and de Vienne, Damien M. 2021. Ghost lineages deceive introgression tests and call for a new null hypothesis. *bioRxiv*, page 2021.03.30.437672. *Cited at page 27*

Varga, T., Krizsán, K., Földi, C., Dima, B., Sánchez-García, M., Sánchez-Ramírez, S., Szöllősi, G. J., Szarkándi, J. G., Papp, V., Albert, L., Andreopoulos, W., Angelini, C., Antonín, V., Barry, K. W., Bougher, N. L., Buchanan, P., Buyck, B., Bense, V., Catcheside, P., Chovatia, M., Cooper, J., Dämon, W., Desjardin, D., Finy, P., Geml, J., Haridas, S., Hughes, K., Justo, A., Karasiński, D., Kautmanova, I., Kiss, B., Kocsubé, S., Kotiranta, H., LaButti, K. M., Lechner, B. E., Liimatainen, K., Lipzen, A., Lukács, Z., Mihaltcheva, S., Morgado, L. N., Niskanen, T., Noordeloos, M. E., Ohm, R. A., Ortiz-Santana, B., Ovrebo, C., Rácz, N., Riley, R., Savchenko, A., Shiryaev, A., Soop, K., Spirin, V., Szebenyi, C., Tomšovský, M., Tulloss, R. E., Uehling, J., Grigoriev, I. V., Vágvölgyi, C., Papp, T., Martin, F. M., Miettinen, O., Hibbett, D. S., and Nagy, L. G. 2019. Megaphylogeny resolves global patterns of mushroom evolution. *Nature Ecology & Evolution*, 3(4) : 668–678. *Cited at page 114*

Vernot, B. and Akey, J. M. 2014. Resurrecting surviving neandertal lineages from modern human genomes. *Science*, 343(6174) : 1017–1021. *Cited at page 23*

de Vienne, D. M. 2016. Lifemap : Exploring the Entire Tree of Life. *PLOS Biology*, 14(12) : e2001624. *Cited at pages 19, 20*

Waggoner, B. 1996. Bacteria and protists from Middle Cretaceous amber of Ellsworth County, Kansas. *PaleoBios*, 17(1) : 20–26. *Cited at page 16*

Weiner, S. and Bansal, M. S. 2021. Improved duplication-transfer-loss reconciliation with extinct and unsampled lineages. *Algorithms*, 14(8) : 231. *Cited at page 116*

Williams, T. A., Foster, P. G., Nye, T. M., Cox, C. J., and Embley, T. M. 2012. A congruent phylogenomic signal places eukaryotes within the archaea. *Proceedings of the Royal Society B : Biological Sciences*, 279(1749) : 4870–4879. *Cited at page 11*

Williams, T. A., Cox, C. J., Foster, P. G., Szöllősi, G. J., and Embley, T. M. 2020. Phylogenomics provides robust support for a two-domains tree of life. *Nature Ecology & Evolution*, 4(1) : 138–147. *Cited at page 11*

Wills, M. 2007. Fossil ghost ranges are most common in some of the oldest and some of the youngest strata. *Proceedings of the Royal Society B : Biological Sciences*, 274(1624) : 2421–2427. *Cited at page 18*

Wilson, E. and Peter, F. M. 1988. Screening plants for new medicines. In *Biodiversity*. National Academies Press (US). *Cited at page 10*

Wilson, E. O. 1989. Threats to biodiversity. *Scientific American*, 261(3) : 108–117. *Cited at page 10*

Woese, C. 1998. The universal ancestor. *Proceedings of the national academy of Sciences*, 95(12) : 6854–6859. *Cited at page 32*

Woese, C. R. 2004. A new biology for a new century. *Microbiology and molecular biology reviews*, 68(2) : 173–186. *Cited at page 22*

Wu, D.-D., Ding, X.-D., Wang, S., Wójcik, J. M., Zhang, Y., Tokarska, M., Li, Y., Wang, M.-S., Faruque, O., Nielsen, R., Zhang, Q., and Zhang, Y.-P. 2018. Pervasive introgression facilitated domestication and adaptation in the Bos species complex. *Nature Ecology & Evolution*, 2(7) : 1139–1145. *Cited at page 41*

Zhang, W., Zhang, X., Li, K., Wang, C., Cai, L., Zhuang, W., Xiang, M., and Liu, X. 2018. Introgression and gene family contraction drive the evolution of lifestyle and host shifts of hypocrealean fungi. *Mycology*, 9(3) : 176–188. *Cited at page 41*