Identifying remote homology and gene remodelling using network-based approaches

Duncan Sussfeld

THÈSE DE DOCTORAT

NNT : 2024UPASL112

université PARIS-SACLAY

Identifying remote homology and gene remodelling using network-based approaches

Identification d'homologies distantes et de gènes remodelés par des approches de réseaux

Thèse de doctorat de l'Université Paris-Saclay

École doctorale n° 577, Structure et Dynamique des Systèmes Vivants (SDSV)

Spécialité de doctorat : Évolution

Graduate School : Sciences de la Vie et de la Santé, Référent : Université d'Évry Val d'Essonne

Thèse préparée dans l'unité de recherche **Génomique Métabolique (Université Paris-Saclay, Univ Evry, CNRS, CEA, Genoscope)**, sous la direction de **Eric PELLETIER**, Directeur de Recherche, la co-direction de **Philippe LOPEZ**, Professeur des Universités, et le co-encadrement de **Eric BAPTESTE**, Directeur de Recherche

Thèse soutenue à Paris, le 03 Décembre 2024, par

Duncan SUSSFELD

Composition du Jury

Membres du jury avec voix délibérative

Chris BOWLER, Directeur de Recherche CNRS, Ecole Normale Supérieure	Président & Examinateur
Marco FONDI, Chercheur, Université de Florence	Rapporteur & Examinateur
Ingrid LAFONTAINE, Professeure, Sorbonne Université	Rapporteure & Examinatrice
Catherine LAROSE, Chargée de Recherche CNRS, Université Grenoble Alpes	Examinatrice

université PARIS-SACLAY

ÉCOLE DOCTORALE
Structure et dynamique des systèmes vivants (SDSV)

Titre : Identification d'homologies distantes et de gènes remodelés par des approches de réseaux.

Mots clés : Evolution moléculaire ; Bioinformatique ; Algorithmes de graphes ; Homologie distante ; Gènes remodelés

Résumé : L'augmentation toujours plus importante de données génomiques et métagénomiques appelle de nouveaux développements méthodologiques et bio-informatiques, afin de caractériser avec davantage de précision les phénomènes évolutifs dans leur ensemble. En particulier, certaines des méthodes usuelles pour étudier l'évolution des (familles de) gènes s'avèrent inadaptées lorsque la parenté entre séquences n'est que partiellement supportée. Ainsi, la définition et la reconstruction de familles de gènes se heurtent à l'obstacle de l'homologie distante, qui passe sous le seuil de détection des alignements de séquences. De même, les mécanismes d'évolution combinatoire, tels que les fusions et fissions de gènes, remettent en cause les représentations purement arborescentes de l'évolution des familles de gènes. L'application de méthodes complémentaires basées sur les réseaux de similarité de séquences permet de contourner certaines de ces lacunes, en proposant une représentation holistique des similarités entre gènes. La détection et l'analyse d'homologues très divergents de familles de gènes fortement conservées dans des jeux de données environnementaux est notamment facilitée par la recherche itérative d'homologie fondée sur les réseaux. Cette fouille itérative de métagénomes révèle une immense diversité de variants environnementaux dans ces familles, qui divergent de la diversité connue tant par leur séquence que par la structure des protéines qu'ils encodent, et elle permet de suggérer des pistes pour guider de futures explorations de la matière noire microbienne. En outre, en prenant en compte des liens d'homologie partielle entre séquences génétiques, les réseaux de similarité de séquences permettent une identification systématique des évènements de fusion et de fission de gènes. Il devient ainsi possible d'évaluer l'impact de ces processus au cours de l'évolution de lignées biologiques d'intérêt, permettant de comparer le rôle qu'ils ont joué lors de l'émergence de phénotypes multicellulaires complexes dans plusieurs telles lignées. Plus généralement, ces approches basées sur les réseaux illustrent l'intérêt de prendre en compte une pluralité de modèles pour étudier une plus grande variété de processus évolutifs.

université PARIS-SACLAY

ÉCOLE DOCTORALE
Structure et dynamique des systèmes vivants (SDSV)

Title: Identifying remote homology and gene remodelling using network-based approaches.

Keywords: Molecular evolution ; Bioinformatics ; Graph algorithmics ; Remote homology ; Remodelled genes

Abstract: The ever-increasing accumulation of genomic and metagenomic data calls for new methodological developments in bioinformatics, in order to characterise evolutionary phenomena as a whole with better accuracy. In particular, some of the canonical methods to study the evolution of genes and gene families may be ill-suited when the relatedness of sequences is only partially supported. For instance, the definition and reconstruction of gene families face the hurdle of remote homology, which falls beneath the detection thresholds of sequence alignments. Likewise, combinatorial mechanisms of evolution, such as gene fusion and gene fission, challenge the purely tree-based representations of gene family evolution. The use of complementary methods based on sequence similarity networks allows us to circumvent some of these shortcomings, by offering a more holistic representation of similarities between genes. The detection and analysis of highly divergent homologues of strongly conserved families in environmental sequence datasets, in particular, is facilitated by iterative homology search protocols based on networks. This iterative mining of metagenomes reveals an immense diversity of environmental variants in these families, diverging from the known diversity in primary sequence as well as in the tertiary structure of the proteins they encode. It is thus able to suggest possible directions of future explorations into microbial dark matter. Furthermore, by factoring in relationships of partial homology between gene sequences, sequence similarity networks allow for a systematic identification of gene fusion and fission events. It thus becomes possible to assess the effects of these processes on the evolution of biological lineages of interest, enabling us for instance to compare the role that they played in the emergence of complex multicellular phenotypes between several such lineages. More generally, these network-based approaches illustrate the benefits of taking a plurality of models into account, in order to study a broader range of evolutionary processes.

Acknowledgements – Remerciements

Mes tout premiers remerciements vont naturellement à l'encontre de mes trois encadrants de thèse. Eric, Eric, Philippe, je tiens à témoigner à vous trois ma plus profonde et sincère gratitude pour votre accompagnement et votre soutien au cours de ces trois années. Je pense que l'on se rend rarement compte du pétrin dans lequel on s'embarque en commençant une thèse, et votre présence constante et bienveillante aura été un appui irremplaçable pour mener à bien ce projet. Eric P, avant tout, merci mille fois d'avoir accepté d'être mon directeur de thèse. La fréquence de nos échanges aura été moindre qu'avec mes autres encadrants, mais ils auront été tout aussi instructifs et appréciés. Merci pour tous tes conseils dans le cadre de nos recherches, et merci d'avoir répondu à mes appels anxieux de doctorant en fin de thèse. Ton soutien m'aura été précieux. Merci également à toi Philippe, pour ton infinie gentillesse et ta bienveillance, toi qui étais également toujours là pour répondre à la moindre de mes questions, merci pour les rires et les encouragements que tu m'as procurés. Enfin, Eric B, merci d'avoir partagé avec moi ton savoir et ta curiosité, tous deux inépuisables, merci de m'avoir accompagné à travers les galères et de m'avoir soutenu tout au long de ma thèse. Merci infiniment à vous trois de l'avoir rendue possible, par votre investissement et votre confiance, par votre bienveillante sympathie qui a fait du laboratoire un excellent lieu de travail. Pour ces trois ans, vous aurez toujours ma gratitude. Je suis fier de pouvoir faire figurer mon nom sur ce document aux côtés des vôtres.

Je tiens également à remercier les autres membres du laboratoire, collègues devenus amis, qui ont rendu l'ambiance du labo si mémorable. Merci à Charles, à Yuping, à Rafael, à Lucas, pour les discussions animées du midi, les occasionnels verres du soir, et les diverses activités culinaires par lesquelles on embaumait le couloir d'arômes discutables. Merci à Danielle pour ta bonne humeur, ton soutien administratif, et encore ta bonne humeur. Merci Jérôme pour toute ta sympathie, et pour les innombrables paquets de McVitie's qui furent si souvent éphémères une fois la porte du labo passée. Merci à Eduardo, ta gentillesse n'a d'égale que l'admiration que je porte à l'étendue de ta culture, merci pour les anecdotes, les discussions saugrenues, et les cassages de tête sur divers problèmes matheux. Enfin, tous mes remerciements à Hugo, mon compagnon de ~~galère~~ thèse depuis plus de trois ans. Merci d'avoir répondu à toutes mes questions biologiques, peu importe leur bassesse, j'espère avoir pu te rendre la pareille en informatique. Merci pour les plaisanteries comme les discussions sérieuses, ce fut mon privilège d'être en thèse en même temps que toi.

Next I wish to thank the different researchers I had the chance to collaborate with during my doctoral studies. Thank you in particular to Mark Cock, and to Peter Mulhair, Mary O'Connell and

James McInerney for letting me join their respective research projects. I am grateful for the experience you have provided me, and I hope to work with you again in the future.

Merci également à mes collègues d'enseignement à l'Université d'Evry. L'enseignement aura été pour moi une découverte fantastique, et c'est notamment grâce à votre accompagnement et votre confiance. Merci donc à Abdelghani Sghir, à Violette Da Cunha, à Valérie Chaudru, et un merci tout particulier à Carène Rizzon et Yolande Diaz pour leur profonde gentillesse.

Je remercie aussi sincèrement Philippe Gambette et Frédéric Cazals, pour leur accompagnement et leurs conseils lors de mes deux réunions de comité de thèse.

I am also very grateful to the members of my PhD jury for agreeing to participate in this final step of my doctoral adventures. All my thanks in particular to Ingrid Lafontaine and Marco Fondi for agreeing to review this manuscript, and to Catherine Larose and Chris Bowler for participating in my defence jury. Thank you for your interest and your insights, thank you for the lovely and lively scientific discussion we shared during the viva.

A l'aboutissement de mes études supérieures, démarrées il y a maintenant neuf ans à Marseille, je tiens particulièrement à exprimer ma gratitude envers les enseignant·es qui ont rendu possible ce parcours. Un merci tout particulier aux responsables (du moins de mon temps) de la Licence MPCI : Guillemette Chapuisat, Philippe Marsal, Marc Georgelin, Julia Charrier et Denis Lugiez. Avec cette licence vous avez créé un environnement exceptionnel pour des scientifiques en herbe, et j'espère que vous êtes conscient·es de son impact pour ceux qui y sont passés. Merci infiniment de m'avoir soutenu dans ces trois ans de licence et de m'avoir poussé vers les ENS. Je peux dire avec certitude que je ne serais pas arrivé ici sans vous. Merci également à Matthias Függer, qui m'a pour la première fois mis le pied à l'étrier de la bioinformatique à l'ENS Cachan.

Je tiens aussi à remercier tou·tes mes ami·es, que je ne me risquerai pas à énumérer par peur d'en omettre. Merci pour tous les rires et tout l'amour que l'on se porte, les indénombrables pintes, les raclettes et les bourguignons au tofu, les Père-Noël secrets, les sorties grimpe, les soirées jeux et les weekends par monts et par vaux. Toutes mes pensées en particulier à celles et ceux qui sont aussi thésard·es, à Michelle, Emma, Robin, Paul : ensemble, croyons en la vie après la thèse.

Je remercie également toute ma famille, les grands-pères et grand-mères, les oncles et tantes, les cousins et cousines : votre soutien est précieux et ma reconnaissance est immense. Merci pour les visites, pour les coups de fil, merci pour les invitations à venir vous voir, même lorsque je n'ai pas pu les honorer.

Many thanks as well to my friends and my family-in-law on the other side of the Channel, for cheering me on and, soon, for welcoming me on their bright, sunny, tropical British archipelago. I very much look forward to seeing more of you in the upcoming years.

C'est avec une émotion toute particulière que j'adresse mes remerciements les plus chaleureux à ma sœur et mes parents. Lilie, tes premières semaines parisiennes n'ont sûrement pas été les plus fun avec un frère/coloc aussi occupé, et j'ai hâte que l'on rattrape maintenant le temps perdu. Papa, je te remercie pour tes encouragements sans faille tout au long de cette thèse, même quand les formulations sont parfois hasardeuses. Enfin, Maman, toi qui as ouvert la voie du doctorat quelques années avant moi, comment résumer en quelques mots l'importance de ton soutien. Merci d'avoir partagé avec moi tes expériences et tes conseils (même si j'en ai ignoré la plupart, si ce n'est par principe car je reste quand même ton fils). Merci infiniment à vous trois.

Enfin, à toi Cameron, merci pour absolument, absolument tout. En sciences les certitudes sont rares, et je te remercie de m'en offrir au moins une. Rien de tout cela n'aurait été imaginable sans ton indéfectible soutien. Thank you.

List of Figures

Table of Contents

Chapter I. Introduction

Over the course of my doctoral studies, I have developed and applied several network-based methods that aim to reconstruct the evolutionary history of gene families. In particular, my research focused on those families that present one of two types of exacerbated divergence, typically exceeding the levels of variation that can be retrieved by sequence alignments. The first focus of this thesis consists in the retrieval of remote homologues of ancient, core gene families, which have diverged from the known diversity beyond detectability by canonical methods despite the marked evolutionary conservation of their known counterparts. In particular, we mined environmental metagenomic data for gene variants that could suggest the existence of uncharacterised lineages branching deep in the tree of life. These results are detailed in Chapter II of this thesis. The second research focus, which we develop in Chapter III, is the identification of gene remodelling events, in particular gene fusion and fission. We applied a systematic detection approach that aims to go beyond the scope of domain-centred analyses. By quantifying gene remodelling in two distinct lineages that

acquired complex multicellular phenotypes, we can draw comparative insights into the effect of combinatorial gene and genome dynamics on emergent multicellularity. The common thread of this thesis is thus the use of sequence similarity networks to investigate 'non-canonical' relationships of homology. In this introductory Chapter, we discuss the notion of homology in a broader sense, as well as the way in which it relates to sequence similarity, and therefore the relevance of methods based on sequence similarity networks in this context.

1. The ever-growing abundance of biological sequence data

Charles Darwin's *On the Origin of Species*, first published in 1859, is inarguably the seminal text of evolutionary biology as we understand it today. In it, Darwin introduced the central concept of natural selection: living organisms must compete for access to limited resources necessary to their survival, leading the fittest individuals to prevail over their less well-adapted counterparts, thus producing more offspring in the next generation. Offspring inherit many characteristics of their ancestors, but not always with exactly identical fidelity: slightly shorter legs, or lighter wings, or wider leaves. Hereditary traits are passed down across generations with some variation, a notion Darwin called descent with modification. In turn, these variations may turn out beneficial or deleterious for survival, and therefore may be passed down, or lost, in further generations.

One weakness in Darwin's theory at the time was the absence of a known physical support for heredity, as he observed himself in the first chapter of *Origin*. The first half of the 20th century saw more and more mechanistic advancements to the understanding of vertical heredity, most prominently the re-discovery of Gregor Mendel's work on inheritance and the development of mathematical population genetics, which were unified with Darwinian theory in the 1940s under the name of Modern Synthesis. However, it was only in the early 1950s that the DNA molecule, which had been discovered nearly a century earlier, was confirmed to be the physical template onto which hereditary genetic information is encoded.

The DNA molecule consists of two polynucleotide strands, coiled together to give DNA its classic double-helix structure. Each strand is a succession of nucleotides, each containing one of four nucleobases: adenine (A), cytosine (C), guanine (G) or thymine (T). In addition, the two strands of a DNA molecule are complementary: adenine and thymine always face each other, as do cytosine and guanine. The genetic information is therefore encoded redundantly by each strand, as the contents of

one strand dictates the contents of the other[1]. What this linear polynucleotidic structure of DNA means, for the bioinformatician, is that any DNA molecule can be abstractly represented textually, by a simple string of A, C, G and Ts mirroring the sequence of nucleotides along a strand. This representation encapsulates all the genetic information encoded by the molecule, in a data format that can be easily read, stored and processed by humans or computers. Similarly, RNA and proteins (the functional products of biological processes encoded by DNA) are also linear polymeric molecules. RNA is a single-strand molecule that has the same sequence as its coding gene, with the exception of thymines that are substituted by uracils (U). Proteins, on the other hand, are chains of residues from a canonical set of 20 amino-acids (Figure 1A). Each block of three nucleotides (called a codon) along a messenger RNA dictates one amino-acid in the resulting peptide chain, following a correspondence known as genetic code (Figure 1B). In other words, not only can the support of heredity and template for biological function (DNA) be represented and studied from its sequence, but its functional products (mRNAs, proteins, non-coding RNAs) can too. Accessing the genome of organisms thus grants a unique window into their evolution and the ways in which they function.

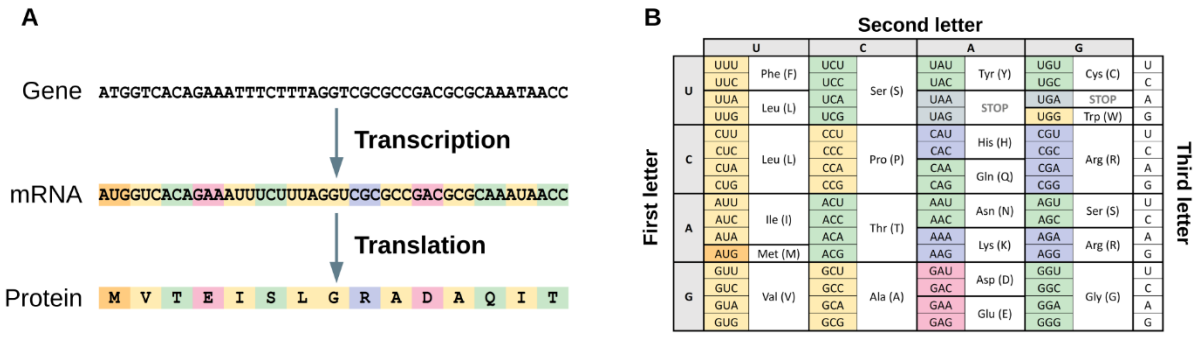


Figure 1: The central dogma of molecular biology.
(**A**) The biological instructions encoded by genes are transcribed into messenger RNAs by RNA polymerase, and mRNAs are then translated into proteins by ribosomes.
(**B**) Translation follows the codon-to-amino acid correspondence of the genetic code.

The first major breakthrough in DNA sequencing came with Frederick Sanger's chain-termination method in 1977, and the publication of the first full genome sequence, that of bacteriophage ϕX174. At first fully manual, the Sanger method was progressively refined and automated throughout the end of the 20th century, with improvements to time and cost requirements as well as reading accuracy. This allowed the first draft of the full human genome to be published in 2001 after over a decade of work by the Human Genome Project and an estimated cost of $2.7 billion.

[1] This is precisely the crux of the DNA duplication process, which allows singular cells to duplicate into multiple copies carrying the same genetic baggage: the DNA strands are separated, and each strand functions as a template for the creation of its new complementary strand, resulting in two copies of the initial molecule.

Sequencing costs remained high in the early 2000s, until the irruption of Next Generation Sequencing (NGS) on the DNA sequencing market, which made genome sequencing dramatically more affordable (Figure 2A). In mid-2007, where NGS was just starting to replace Sanger sequencing in laboratories, raw sequencing costs were of roughly $500 per million base pairs (Mbp); a year later, NGS had dropped costs to $8/Mbp, and $0.35/Mbp by mid-2010. Recent estimates place the cost of sequencing one Mbp at $0.006 in 2022, a drop by five orders of magnitudes in just 15 years.

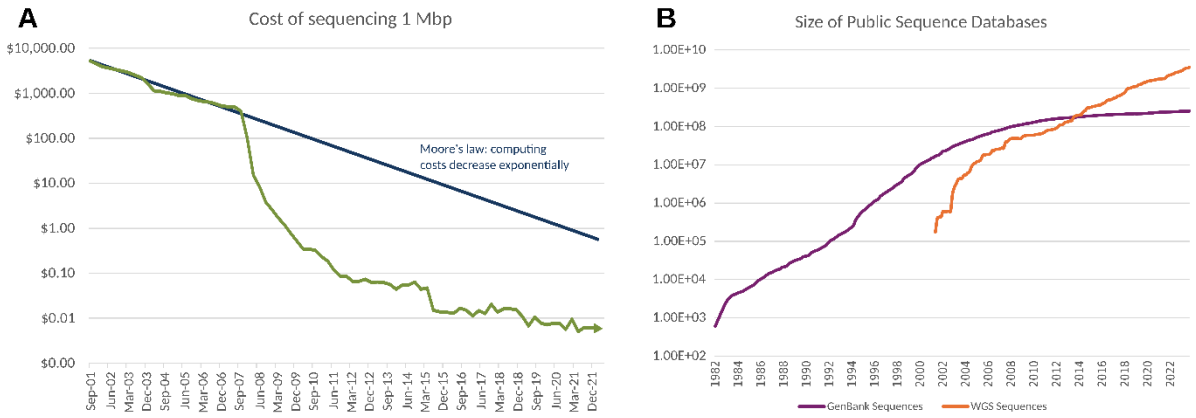


Figure 2: NGS methods provide access to unprecedented amounts of sequence data.
(**A**) In the early 2000s, Sanger sequencing became exponentially more affordable, in line with the predictions of Moore's law on the exponential increase of computing power over time. In the 15 years since NGS technologies entered the market, sequencing costs have decreased even more rapidly.
Data from: Wetterstrand KA. DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP) Available at: www.genome.gov/sequencingcostsdata. Accessed in October 2024.
(**B**) Public sequence databases such as GenBank contain more and more records each year, making necessary the development of new bioinformatic methods to address this ongoing informational torrent.
Data from: https://www.ncbi.nlm.nih.gov/genbank/statistics.

The democratisation of mass DNA sequencing has resulted in a genuine torrent of new genomic data (Figure 2B). Furthermore, beyond whole genome sequencing, other 'omics' developments have allowed access to many other kinds of biological data, such as RNA transcripts (transcriptomics), proteins (proteomics), or metabolites (metabolomics) [Hollywood, Brison, and Goodacre 2006, Aslam et al. 2017, Lowe et al. 2017, Stricker, Köferle, and Beck 2017]. New methods in metagenomics also make it possible for the genomic contents of whole environments to be sequenced at once, thus bypassing the constraining requirements of cultivation and isolation, and provide new insights into the inter-species interactions that sustain ecological systems. This massive influx of biological data is an unparalleled trove of information for the scientific community, and allows us to investigate evolutionary processes from many new angles. The magnitude of this genomic torrent also raises practical and computational challenges. Automated methods are now vital to produce, process and

analyse this biological information in an efficient and reliable way, and continued progress in bioinformatics is required to keep up with sequence datasets of ever-increasing sizes.

2. Comparing sequences to infer evolutionary relationships

2.1 – Understanding the evolution and function of biological sequences

Rendering the genomic information from nanoscopic DNA molecules in a 'language' convenient for humans and computers allows for further characterisation of this sequence dataset. From this point, and as is the case in any other scientific field, the new data must be interrogated in the light of already established knowledge. A biologist may glean a certain amount of information from a gene sequence itself, by looking into specific features such as CG-content and codon usage bias. Establishing the genomic signature of this sequence with such metrics can provide insights into its taxonomy, function and ecology [Coutinho, Franco, and Lobo 2015]. However, to understand the evolutionary history and the biological function of that gene with better certainty, its analysis will usually include comparing its sequence to other genes, with features that have already been characterised and validated. This is not so different from a historian who happens upon an unknown ancient text for the first time: if the contents of the text may already provide some information about its nature or its purpose, it is only within its greater historical context that the origin, importance and significance of the document can be truly evaluated. Over the past decades, dozens of generalist and specialised sequence databases, hosting billions of public DNA sequences, have been assembled to facilitate sequence comparisons for such purposes [Quast et al. 2013, Benson et al. 2013, Jolley, Bray, and Maiden 2018, The UniProt Consortium 2023].

The most common way to recontextualise a novel gene (or protein) sequence is to ascribe it to a known gene family, i.e. a set of homologous genes sharing a common evolutionary ancestor. Gene families are one of the main organisational principles of the global genetic space and are meant to represent coherent units of gene evolution. Comparing gene sequences to reconstruct homologous gene families can therefore help us understand the evolution and divergence of this family, following the expectation that more distantly related genes will have less similarity between their sequences. Moreover, for certain genes in particular that are known to be remarkably conserved (meaning that mutations in their sequences are particularly rare), evolutionary information at the gene level can be used to infer evolutionary relationships between their hosts. The prime example of such a marker gene is the one coding for 16S ribosomal RNA (Figure 3), present in all prokaryotic life forms. The

sequence of this gene consists of a succession of highly conserved regions, interspaced by nine more variable regions (numbered V1-V9). This allows the representation of evolutionary relationships with a large range of granularity, from strain identification to reconstructions of the overall tree of life[2] [Yang, Wang, and Qian 2016].

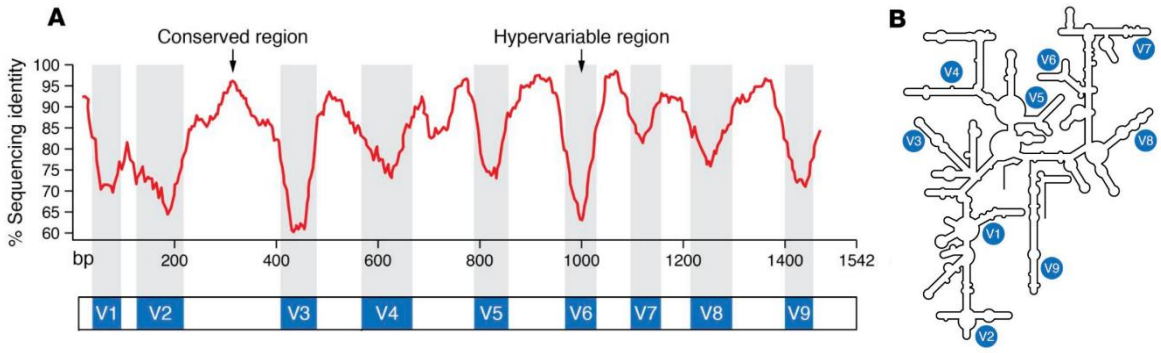


Figure 3: The prokaryotic 16S rRNA gene.
(**A**) Site-specific sequence identity between bacterial 16S rRNA genes, with the nine hypervariable regions indicated by shaded ranges in the graph. The degree of variability differs between V1-V9 regions, such that different subsets of those regions are well-suited for different taxonomic resolutions.
(**B**) Secondary structure of the 16S rRNA molecule, with the positions of V1-V9 regions indicated. Adapted from: [Wensel et al. 2022].

The functional aspect of gene sequence annotation also relies on comparisons with known genes. Indeed, the nucleotide sequence of a gene that codes for a protein dictates its amino-acid sequence (see Figure 1), which in turn influences its three-dimensional conformation. Many features of a protein's spatial conformation, such as its flexibility or rigidity, or which of its residues are exposed at its surface, can be essential for a protein to perform its 'intended' function by interfacing and interacting with other biomolecules. As such, some functional information about an unknown gene sequence can be extracted in a number of ways. If a gene has clear homology with another sequence that has already been functionally annotated, then this likeness in sequence may extend to a likeness in function. However this is not automatically true, as relatively small changes in a protein's sequence can result in much greater structural variation [Tokuriki and Tawfik 2009]. For this reason, other approaches prefer to look directly for structural features in the sequence to annotate. In particular, the computational prediction of 3D protein structures from their primary sequence, which has long been considered a major challenge for bioinformatics, has seen dramatic advances in the last few years with the emergence of deep-learning methods (with DeepMind's AlphaFold at the forefront).

[2] Carl Woese and George E. Fox notably pioneered the use of 16S rRNA to reconstruct phylogenies, leading to their discovery of the archaeal Domain [Woese and Fox 1977].

Nearly all proteins in the UniProt sequence database now have AlphaFold-predicted structures, greatly improving inferences of functional features for unknown sequences. At a more granular level, specific structural features can also guide functional inferences. Protein domains, in particular, are structural units that are encoded by one contiguous region in the coding sequence and translate to a contiguous region of the protein's polypeptide chain that folds onto itself independently from the rest of the protein. Because these domains often correspond to specific functions, identifying domain-coding regions in a gene sequence can also provide indications about the functional role of its protein. Nonetheless, the functional annotation of gene and protein sequences is far from a solved issue. Many predicted protein domains, for instance, are not associated with known functions (as of October 2024, domains of unknown function (DUFs) appear in approximately four thousand Pfam protein families), and some proteins are not covered by domains at all [Paysan-Lafosse et al. 2023]. Likewise, as we discuss in more detail in a following chapter, large numbers of detected protein families have no functional labels, especially in metagenomic data.

As a contextual note: because protein-coding DNA sequences (CDS) can be deterministically translated to amino-acids (see Figure 1), it is common to use protein sequences for evolutionary comparisons at larger taxonomic scales and, likewise, to assimilate gene families and protein families. This offers an immediate computational advantage, because sequences are then three times shorter (since each codon is encoded in one character instead of three). Moreover, sequences of amino-acids are better conserved than nucleic ones, meaning that more ancient relationships between sequences can be detected. This omits information about synonymous mutations in CDS, i.e. substitutions that occur in a codon without changing the corresponding amino-acid, due to the redundancy in the genetic code. Synonymous substitutions are likely to be selectively neutral, but they can also affect gene expression and protein folding [Bailey, Alonso Morales, and Kassen 2021] and thus be adaptive. Still, the approximation of ignoring synonymous mutations is acceptable for the larger-scale studies that we focus on, especially in light of the gain in computational efficiency. On the other hand, indel mutations inside a gene can result in a translational frameshift (if their length is not a multiple of three), which will greatly alter the resulting amino-acid sequence. Using translated sequences to compare genomic data can therefore potentially lead to erroneous results, because related sequences that differ by a frame-shifting indel will then not be recognised as similar. In the case of protein-coding genes however, this is highly likely to produce a dysfunctional protein, which may be abnormally short or long because the initial stop codon is now out-of-frame. In a way, this could be assimilated to a gene loss, since the well-formed protein is not encoded anymore. Here, because we are mostly focused on studying the evolution of protein-coding families, we predominantly use amino-acid

sequences for our computations, as the drawbacks mentioned here are only of minor concern for our purposes.

2.2 – Comparison of biological sequences with pairwise alignments

By 'comparing' pairs of sequences, what is generally meant is that we try to identify common regions between the two sequences that have many identical or similar letters at the same positions. This practice makes sense when comparing sequences that derive from a common ancestry: the sequences were initially identical, then modifications appeared over time in one or the other (a letter could have been lost by one but kept by the other, or have appeared in only one of them, or changed for a different letter), but enough positions may have stayed unchanged to recognise a common root. Aligning these sequences then consists in matching their corresponding positions two by two, to represent which positions are conserved and which ones have diverged.

Let us look, as an example, at the English name Peter, and its Spanish equivalent Pedro. Both come from the Greek *Petros* (Πέτρος), and from this root they share a similar consonant structure *p-[t/d]-r*, as well as an *-e-* in the leading syllable. On the other hand, the terminal *-s* has disappeared in both, and the Spanish version of the name turned the *-t-* into a *-d-*, whereas the English one lost an *-o-* and gained another *-e-*. These positional similarities and differences can be summarised by writing one name above the other, in a way that matches pairs of corresponding letters vertically:

```
P E D - R O
| | :   |
P E T E R -
```

Alignments are by far the most common way to compare sequences together, and are routinely performed for a wide range of biological studies. Aligning two sequences relies on the hypothesis that they share some level of homology, such that their differences can be attributed to an accumulation of mutations since their divergence, rather than a convergent acquisition of the same features. Each insertion, deletion or substitution of a letter in the sequence is considered to happen with a given probability. For instance, when comparing DNA sequences, transition mutations (A↔G or C↔T) are more likely than transversions (A/G ↔ C/T) because of the two different molecular classes of nucleotides (purines, A and G; pyrimidines, C and T). Likewise, in protein sequence alignments, the frequency of each substitution is estimated from sets of sequences with known homology. Based on these probabilities, a numeric score can be assigned to a sequence alignment, using a substitution matrix (PAM and BLOSUM matrices being the most common ones) and a gap penalty function: matches between identical positions or frequent substitutions increase this score, whereas rare mismatches and insertions or deletions that add gaps to the alignment are penalised negatively. The

score of an alignment between two sequences is then obtained by adding up the scores at each position. In these scores, the penalties assigned to gaps generally follow a linear function, where the first position 'opening' a gap is more severely scored than following positions that 'extend' the gap (Figure 4).

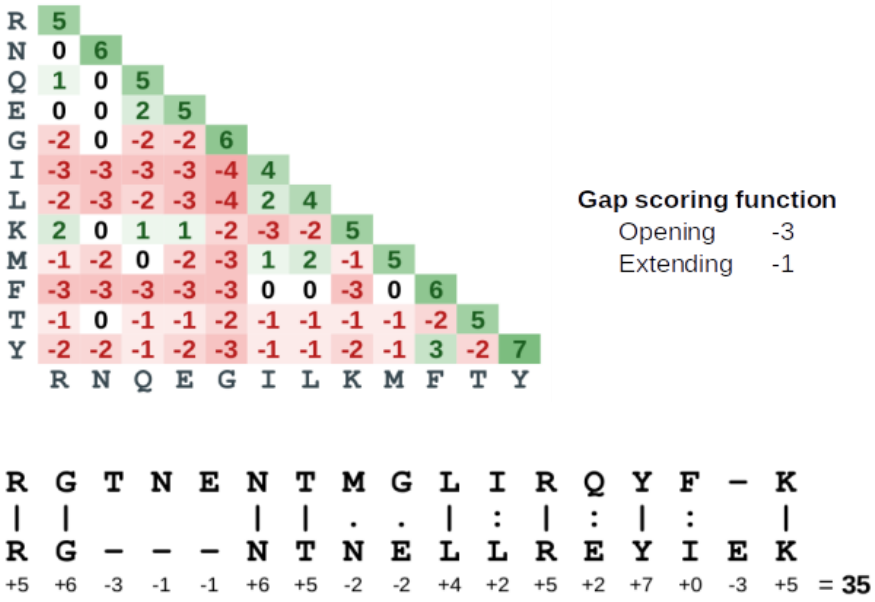


Figure 4: Process of scoring a pairwise sequence alignment.
Top left: An extract of the BLOSUM62 substitution matrix, computed from sets of proteins with less than 62% of identical positions. **Top right**: The gap scoring system, which penalises gap openings more than extensions, to reflect that a single indel event may produce gaps over more than one position. **Bottom**: The alignment between two amino-acid sequences is scored by adding up the score of each position, according to the chosen substitution matrix.

2.3 – Alignment algorithms, BLAST, and the speed-accuracy trade-off

The theoretical number of possible alignments between two sequences grows extremely large as soon as sequences exceed a few dozen letters in length. Algorithmic methods have therefore been developed to identify the optimal alignments of two input sequences efficiently. Some of those algorithms are designed to provide the exact solution to this problem (i.e. the highest-scoring alignment possible), but this optimality results in a higher complexity that greatly slows computations. Notable exact algorithms include the Needleman-Wunsch algorithm, which uses dynamic programming to identify the optimal global alignment between two sequences [Needleman and Wunsch 1970], and the Smith-Waterman algorithm, which adapts the Needleman-Wunsch process to find local alignments between subregions of the sequences [Smith and Waterman 1981]. Probabilistic algorithms that rely on heuristics and approximations, on the other hand, can run much more efficiently to find alignments that get close to, but not always exactly on, the optimal solution. Most

notable among them is BLAST (Basic Local Alignment Search Tool), which has become a veritable staple to anyone working with sequence data [Altschul et al. 1990]. Most modern sequence databases published online now integrate some BLAST implementation to let users search an input sequence within the database records, leading to BLAST being sometimes presented as the 'search engine' for DNA and protein sequences. Perhaps unsurprisingly, the work we present here made extensive use of BLAST alignments too, and warrants a few explanations about the way it functions.

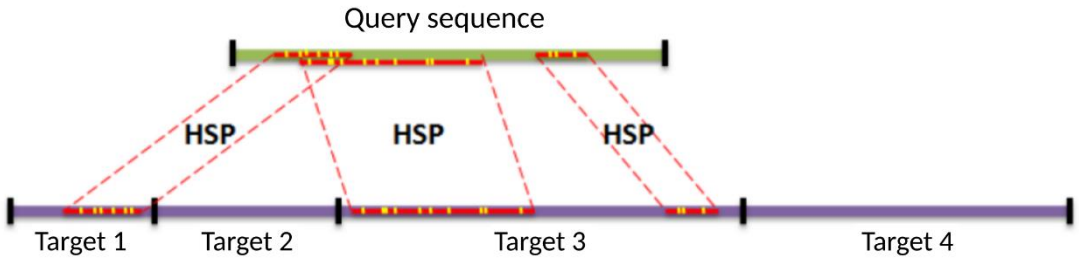


Figure 5: Seed-and-extend search of local alignments using BLAST.
BLAST identifies short segments of (near-)perfect identity between the query sequence and the targets (yellow points). These seeds are then extended into High-scoring Segment Pairs (HSPs), highlighted in red. Notice that more than one HSP can exist between the same sequences. From: [Jachiet 2014].

Like Smith-Waterman, BLAST belongs to the family of local aligners, which implement the notion that sequences can be similar for only some portion of their length, instead of their full span. It uses a particular heuristic called 'seed-and-extend', which assumes that high-scoring alignments must contain short segments of identical or near-identical letters (Figure 5). The first step of a BLAST alignment thus consists of finding 'seeds', i.e. identical or nearly identical segments between the two sequences (typically three letters long for protein alignments, and 11 letters for nucleotide sequences). Local matches called HSPs (High-scoring Segment Pairs) are then extended from each seed, towards the left and the right, until dropping significantly in quality, and the highest scoring alignment encountered during this extension is retained. HSPs from two consecutive seeds can sometimes overlap, in which case BLAST merges them together when beneficial. BLAST then returns all HSPs with a better score than a user-defined threshold. This output can be provided in a variety of formats, which can either make the exact alignment explicit or only specify the endpoints of each HSP along each sequence.

Due to the rapid growth of the amount of sequence data now available and of the increased computing power of modern processors, BLAST is now generally used to align many sequences together at once, rather than simply two. In its implementation, there is therefore a distinction between query and target sequences. Typically, if one has obtained new gene sequences and wants

to check if they are any similar to genes already published in a database, the new sequences form together the query set, and sequences from the database are the targets. A single execution of BLAST can then search for alignments between any query sequence and any target. On the other hand, if one wants to compare between all pairs of sequences in a single set (in what we refer to as an all-against-all alignment), then every sequence is both a query and a target. Each pair of sequences will then be compared twice (once in each 'direction'), and it should be noted that results of bidirectional alignments can sometimes differ slightly, due to algorithmic optimisations that improve calculation times but disrupt the symmetry of the comparison.

2.4 – Understanding an alignment output: what are the relevant metrics?

A

qseqid	sseqid	evalue	pident	qstart	qend	sstart	send
seq1	seq1	6e-222	100	1	293	1	293
seq1	seq3	1e-53	32.7	1	293	1	279
seq1	seq2	4e-49	31.7	8	287	6	274
seq1	seq4	2e-32	28.1	9	267	1	239
seq1	seq5	2e-25	29.0	1	293	1	276
seq2	seq2	1e-214	100	1	282	1	282
seq2	seq3	2e-57	37.3	1	280	1	279
seq2	seq1	4e-49	31.7	6	274	8	287
seq2	seq5	6e-32	30.8	5	281	7	277

B

qseqid	**q**uery **seq**uence **id**
sseqid	**s**ubject **seq**uence **id**
evalue	E-value
pident	**p**ercentage of **ident**ical positions
qstart	**start** position of alignment on **q**uery
qend	**end** position of alignment on **q**uery
sstart	**start** position of alignment on **s**ubject
send	**end** position of alignment on **s**ubject

C

```
Query= seq2
>seq3

 Score = 174 bits (442),  Expect = 2.65e-57
 Identities = 107/287 (37%), Positives = 166/287 (57%), Gaps = 15/287 (5%)

Query   1  CKCSKDLDPIYSTDYKVPTVGDKPEGCITKLDKGGEE-DFKYFNREVLNKKGLGESEYIE 59
           CK S D  P++S +  +   GD PEG   KL+K   + D K+   + +  K   + EY++
Sbjct   1  CKFSDDWQPLHSLEAPLDICGDIPEGARAKLEKANAQADIKHSTTKFVPGKVFADVEYVD 60

Query  60  DGPGSPQYKTKIKFELLGLIIQSPMIELLLEGAPWRLSLRCEWWFGRNVDFADNGEHQQP 119
           D PG  QY+  + +E +GLI Q+P +EL+ + +PW L++R E W  R+ +F D   H QP
Sbjct  61  DFPGKAQYEYGV-YESIGLIFQNPELELVRDASPWGLTVRSECWSDRSGNFFDGAAHAQP 119

Query 120  SPVILEALSVRFTKSLSRNDFS--DVPPASSQKPHYVEGVDTVIPYSDKDRTDIELWINQ 177
           +P  L      FT++L + D +  ++P A+S+ P++++ ++  +PYSD D  +I LWI+
Sbjct 120  TPWAL------FTRALVKYDIACNEMPHANSKTPNFIKVLNNGLPYSDFDNLEITLWISD 173

Query 178  NSESVSCRSF-LAFSRLGPESLTAAPLTYYSESKPS-SADETP--GSVSFFQRVAVYMFA 233
           N  + SCRSF L   S +GPE+LT A + Y SES PS SADE     GS +FF  +AV++
Sbjct 174  NGRTASCRSFSLDLSHIGPEALTLA-IKYTSESDPSQSADEAGEMGSTAFFLHMAVFIMC 232

Query 234  MRKLGKEDRAQRKDKETNTGVEQTYVLYHDKDAAEVKGRQVNNERKR 280
           M+          ++ K  N GV  +YVL   ++    KG QVN+  +R
Sbjct 233  MKTRNFGRFEPKEAKWANQGVPNSYVLLATRERIAAKGSQVNHSNER 279
```

Figure 6: Example of a protein sequence alignment with BLAST.
(**A**) Tabular BLAST output of an all-against-all alignment between a toy set of 5 sequences. (**B**) Alignment descriptors used in a typical BLAST output. Note that these only qualify the aligned region: for instance, the pident column only counts the positions within each HSP. (**C**) Detailed output of BLAST showing the explicit alignment between seq2 and seq3. Highlighted areas of text show the correspondence between fields in the tabular and the full alignment outputs.

Raw alignment score values intrinsically depend on the choice of a scoring matrix and gap penalties, making it impossible to compare between search results that used different scoring systems. Other scoring metrics that can be applied more uniformly are therefore preferred when it comes to comparing alignments. In particular, the E-value is very frequently used to assess the validity of BLAST outputs (Figure 6). The E-value of an alignment quantifies how many hits of similar or better quality could be expected (hence, E) by chance between two random sequences of similar sizes. More precisely, an alignment between two sequences of length **m** and **n** with a raw score **S** (computed as

above) will be assigned an E-value of $\mathbf{E = K{\cdot}m{\cdot}n{\cdot}e^{-\lambda S}}$ – here K and λ, called the Karlin-Altschul parameters, normalise the influence on the E-value of the selected scoring system and its underlying assumptions about amino-acid frequencies in the protein space [Karlin and Altschul 1990]. The E-value thus decreases exponentially with the raw alignment score, with lower values indicating more reliable alignments that are less likely to come from a spurious similarity between sequences. When the target set is not a single sequence but a larger database, its size is factored into the E-value calculation, which then represents the expected number of similar hits against a random database of comparable size.

The influence of the selected scoring system on raw alignment scores disappears when converting to E-values, but the resulting values are still a function of the query and target sizes. This means that comparing alignments based on E-values only formally makes sense when they are performed against the same target: looking up two genes in a same database can tell us which one 'fits' the database the most, but looking up the same gene in two different databases (of different sizes) should not inform us on which one it fits best. In practice, an E-value grows proportionally with the size of the target, but since it decreases exponentially with the raw alignment score, that score remains the main deciding factor in expected values. Comparing across databases can thus be permissible with a proportional adjustment of E-values to take unequal database sizes into account.

Choosing an E-value threshold under which to consider alignments as significant depends on the stringency that is required for the specific purposes of each analysis. The limit of 10^{-5} is commonly used, with sometimes even lower orders of magnitude for stricter filters. Even at E-values of 10^{-5}, similarities between two sequences are sometimes difficult to discern visually, and the E-value threshold is often coupled with other criteria to evaluate alignments. The percentage of identical positions within the alignment, and the fraction that it covers on the entire length of each sequence, are in particular often used in tandem with an E-value threshold. Again, the choice of a threshold here depends on what kind of sequence similarities we are looking to find. Strict limits on alignment coverage will help identify full-length similarities between sequences, but lower thresholds are more adapted to local similarities, for instance when looking for common domains between complex multi-domain proteins. Likewise, looking for near-identical sequences can warrant percentages of identity above 90% or even 95% (for instance, two organisms are generally considered to belong to the same species when their 16S rRNA genes have more than 97% identical nucleotides), whereas lower values are relevant for finding more distant similarities. There is, however, a lower limit of sequence similarity that can be detected in practice by sequence aligners. For amino-acid sequences, in particular, proteins that have less than 25-30% of identical positions (a range known as the "twilight zone" of protein alignment) are aligned by BLAST only with some difficulty [Rost 1999], meaning that distantly

related homologues may be missed by alignment searches. This issue of remote homology detection is discussed in further detail in a subsequent chapter, as it is one of the focuses of this thesis.

2.5 – Different kinds of homology and sequence similarity

Establishing significant similarities between the sequences of different genes or proteins usually serves to answer one question: do these genes/proteins stem from the same ancestor, i.e. are they homologous. Indeed, modern genomes are close to the only source of information at our disposal to study evolution, so we must rely on contemporary data to infer past events in the history of life[3]. The similarity between two sequences is thus used as a proxy to infer their homology, based on the strength of the alignment between the two (according to the metrics discussed above). In this subsection, we briefly review the different types of homology and their significance in the study of evolution.

The perhaps 'canonical' scenario for sequence homology is that in which a gene, present in an ancestor organism, perpetuates itself in multiple extant organisms following a chain of speciation events. The term *orthology* was coined to describe such homology relationships in 1970, in opposition to the then-new concept of evolution by gene duplication, which was in turn called *paralogy* [Ohno, Wolf, and Atkin 1968, Fitch 1970] (Figure 7). Orthology is a particularly important notion for evolutionary studies, because the divergence of a set of orthologous genes presumably reflects the divergence of their hosts, and therefore an accurate definition of orthologous families is crucial to reconstruct phylogenies between species. Moreover, orthologues generally fulfil identical (or biologically equivalent) functions in different organisms, whereas paralogues are more likely to diverge in function after duplication[4] (a process known as sub- or neo-functionalisation). Although this "orthology-function conjecture" is more a statistical genomic trend than an immutable law [Gabaldón and Koonin 2013], orthology is nevertheless an important resource for the functional annotation of

[3] Of course this is not strictly true: evolution was already studied before the advent of genetics, e.g. by analysing fossils (i.e. paleontology). However, fossils only provide morphological data, and are thus mostly relevant for animal and plant evolution, *de facto* overlooking microorganisms. Another source of evolutionary data is ancient DNA (aDNA), sampled in preserved specimens: naturally or artificially mummified remains, paleofeces, frozen material, etc. Because aDNA is subject to degradation, we can only sequence samples up to 2 million years old [Willerslev et al. 2004, Kjær et al. 2022], meaning that aDNA will mostly be able to yield insights into comparatively recent evolution.

[4] When gene duplication was first described, its ability to produce functional innovations at a faster rate than local mutations led some scientists to view duplication as the driving force of evolution (see Introduction of [Ohno, Wolf, and Atkin 1968]: "Gene duplication now emerges as the prime factor of evolution"). Although duplications are clearly an important factor of evolution, further investigations have found that neo-functionalisation was not the main outcome for duplicated genes [Shakhnovich and Koonin 2006].

new genetic sequences and is the basis for numerous databases of functional clusters (including eggNOG and OMA, to name but a few).

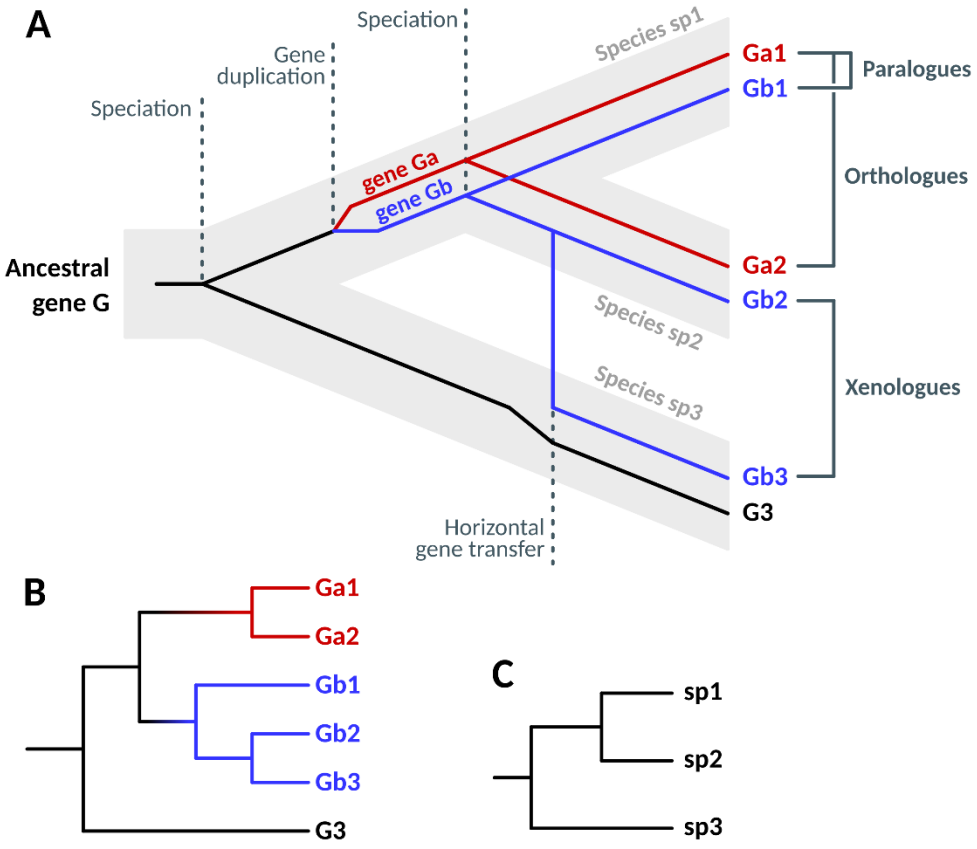


Figure 7: Different kinds of sequence homology.
(**A**): A hypothetical scenario of evolution for one gene family. Successive events of speciation, gene duplication, and horizontal gene transfer give rise to paralogues, orthologues and xenologues in three extant species. As a result, the phylogenetic tree of the gene family (**B**) is incongruent with the species phylogeny (**C**).

The two main types of homology, orthologues and paralogues, both represent 'internal' processes of gene inheritance that follow an arborescent model of divergence. However, genomes also evolve by more reticulate mechanisms, and horizontal gene transfer (HGT), wherein a gene from one organism is recruited into the genome of another one, is another significant process driving evolution. Two genes that are related in this manner are termed xenologues (Figure 7). Far from marginal, this process is actually ubiquitous in prokaryotes, which routinely exchange genetic information with their counterparts, even when only very distantly related. Several mechanisms contribute to HGT in prokaryotes, including gene transductions mediated by viruses, exchanges of plasmids between cells in membrane-to-membrane contact, and direct uptakes of extra-cellular genetic material from the environment (Figure 8A). As a result, the prokaryotic world is often

described as one huge interconnected gene pool, sub-compartmentalised by genomes of distinct lineages but with little to no rigid barriers [Bapteste et al. 2009]. This process is essential for microbial adaptation and survival: due to their asexual mode of reproduction that does not involve genetic admixture, other sources of genetic innovation are vital to prevent a gradual accumulation of deleterious mutations that eventually leads to extinction[5]. In eukaryotic lineages, HGT is a far less frequent occurrence, partly because of the existence of the nucleus that segregates chromosomes from the rest of the cellular milieu. Still, multiple cases have been documented, including gene uptakes from bacteria, but also between plants and fungi. Another particular case concerns the transfer of genetic material between the nucleus of a eukaryotic cell and its organelles. Mitochondria, present in all eukaryotes, and chloroplasts, responsible for the photosynthetic ability of plant and algal lineages, both result from endosymbiosis events with bacteria, and thus possess genomes. The 'protection' offered by an endosymbiotic lifestyle allowed major reductions in the gene content of mitochondria and plastids, and instances of gene flow between organelles and the nucleus have been recorded in just about every direction (Figure 8B). In all these instances, from the prokaryotic 'web of life' to eukaryotes and their organelles, gene transfers result in homology relationships that challenge the strictly tree-like view of evolution.

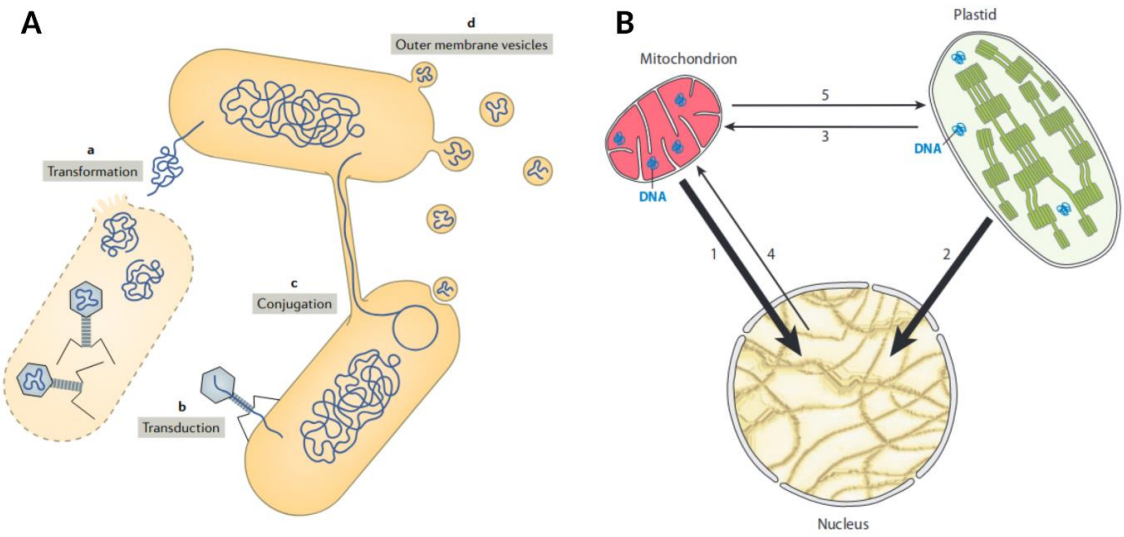


Figure 8: Horizontal gene transfer between prokaryotes and within eukaryotic cells.
(**A**) Prokaryotic organisms can exchange genetic material in a variety of ways: direct environmental intake (a), cell-cell conjugation (c), mediation by mobile genetic elements (b, d). From: [Brito 2021].
(**B**) Gene flow routes between organelles and nuclei in eukaryotic cells. The thickness of arrows represents the frequency of each exchange route. From: [Kleine, Maier, and Leister 2009].

[5] This process, known as Muller's ratchet, was first introduced by the American geneticist Hermann Muller in a 1932 talk titled "Some genetic aspects of sex" [Muller 1932]. The term itself was coined by Joseph Felsenstein in his 1974 paper "The evolutionary advantage of recombination" [Felsenstein 1974].

When discussing homology relationships thus far, we have mostly focused on genes and proteins as units, and considered sequence similarities that span the majority of the length of each sequence. However, sequences may also resemble each other only in some part of their length – this is, for instance, implicitly stated by the design of local sequence aligners. Two proteins can share the same domain, but can each have a second domain that is absent in the other. In this case, should we consider these proteins to be homologues, if some part of them descends from a common ancestor but others do not? To clarify this relationship, we can say that the shared regions are homologous, and that the two proteins as a whole are *partial* homologues – here "partial" does not qualify the strength or the quality of the homologous region, but rather that it is indeed only a subregion of the whole sequence. This also implies that gene sequences can be composites of different subunits from different origins: in the same way that HGT introduces mosaicism to genomes, genetic remodelling introduces mosaicism to genes. This is far from a fringe phenomenon, as multi-domain proteins are estimated to represent 65-80% of the proteome in Eukaryotes, as well as 40-60% in prokaryotes [Apic, Gough, and Teichmann 2001, Ekman et al. 2005]. Along with remote homology, partial homology and gene recombination is the second focus of this thesis, and we return to these notions in greater detail in a future chapter.

In summary, sequence similarity is an extremely important descriptor to infer homology, but it is not an infallible one. First, distantly homologous sequences can be difficult to align past a certain point of divergence; second, several evolutionary processes (e.g. gene duplication, HGT, gene remodelling) can result in incongruences between the evolution of a gene family and that of its hosts. Additionally, gene families can take complex evolutionary trajectories involving processes that are sometimes incompatible with the usual arborescent representation of its history. As such, other modelisations of a gene family can complement this tree-centric view in a useful way. The present work, in particular, relies heavily on sequence similarity networks, which attempt to provide a more holistic representation of the gene-to-gene (or protein-to-protein) relationships within families. Networks in general, and sequence similarity networks in particular, are the focus of the next section.

3. Sequence similarity networks

3.1 – What do we talk about when we talk about networks

The term "network" is a recurring buzzword that has permeated common parlance around a broad variety of topics. The word is perhaps most frequently used in relation to technology and telecommunications (as in social networks, neural networks, and implicitly in internet), but also appears in interpersonal contexts (attending networking events, for instance, is often recommended

to people looking to grow their professional network). Under this jargon lies a specific approach to model certain types of data, which is extensively studied by mathematicians and computer scientists, and applied to numerous scientific areas such as biology, physics, sociology or economics. Networks provide intuitive visual representations of relationships between data points, backed by a robust theoretical framework that provides quantitative ways to describe these relationships.

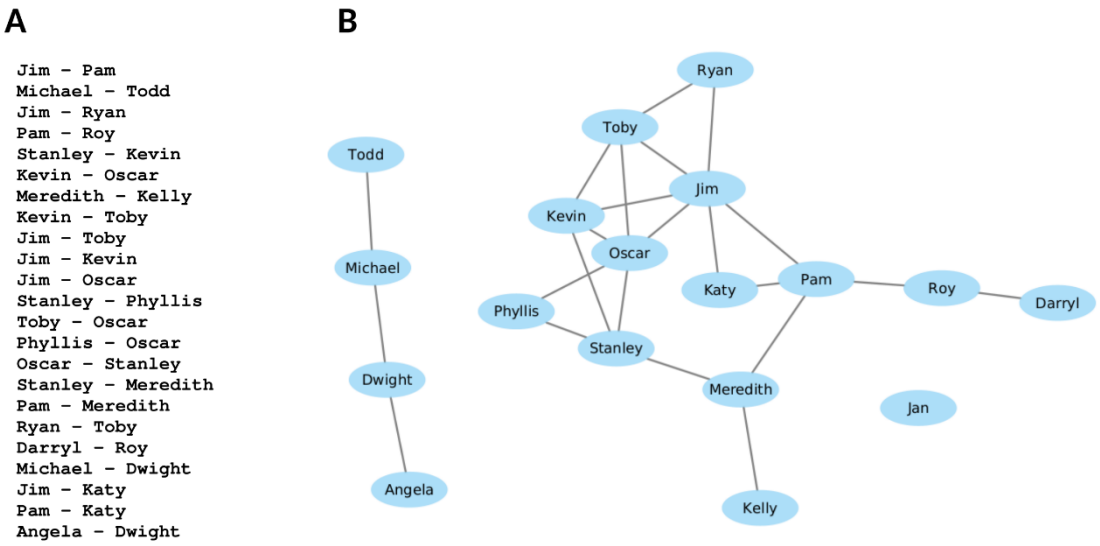


Figure 9: Using the network view to represent a list of social interactions.
(**A**) A list of all friendly interactions represented on screen in the first series of *The Office (US)* (subject to personal interpretation).
(**B**) The same interactions represented as a network: each node corresponds to a character, and edges are drawn between characters that have an interaction listed in (A). Note that Jan is absent from the list in (A), as she is never shown to be more than neutral with any other character despite appearing in 4 out of 6 episodes.

Let us take advantage of this visual quality to introduce some basic notions of network science with an example, based on one of the most popular comedy TV series of recent years: the American sitcom *The Office* (2005-2013). This TV show follows the day-to-day lives of employees in the Scranton regional branch of the (fictional) paper distribution company Dunder Mifflin. Over nine series, viewers can follow the evolution of interpersonal relationships between all the employees, ranging from romantic to friendly or cordial, all the way to hostile. As the series goes on, the overall social structure of the Scranton office is heterogeneous and dynamic. To represent the current state of relationships at a given point in the show, we could for instance list all the friendships and all the animosities that the relevant characters entertain with one another. This would, of course, be a slightly reductive depiction of the more complex social dynamics that the authors portray on screen, but could still give a fairly accurate idea of affinities between characters. However, this list would perhaps be ineffective

at depicting the global office relationship picture in a convenient visual way, given the cast of 18 recurring characters in the first series alone. This is where the network view comes in, providing a descriptive figure that depicts the same information as the friendships list in a perhaps clearer way (Figure 9). In this figure, each character in *The Office* is represented by a node, and links (or edges) connect characters that are friends, or at least friendly, in the pilot series.

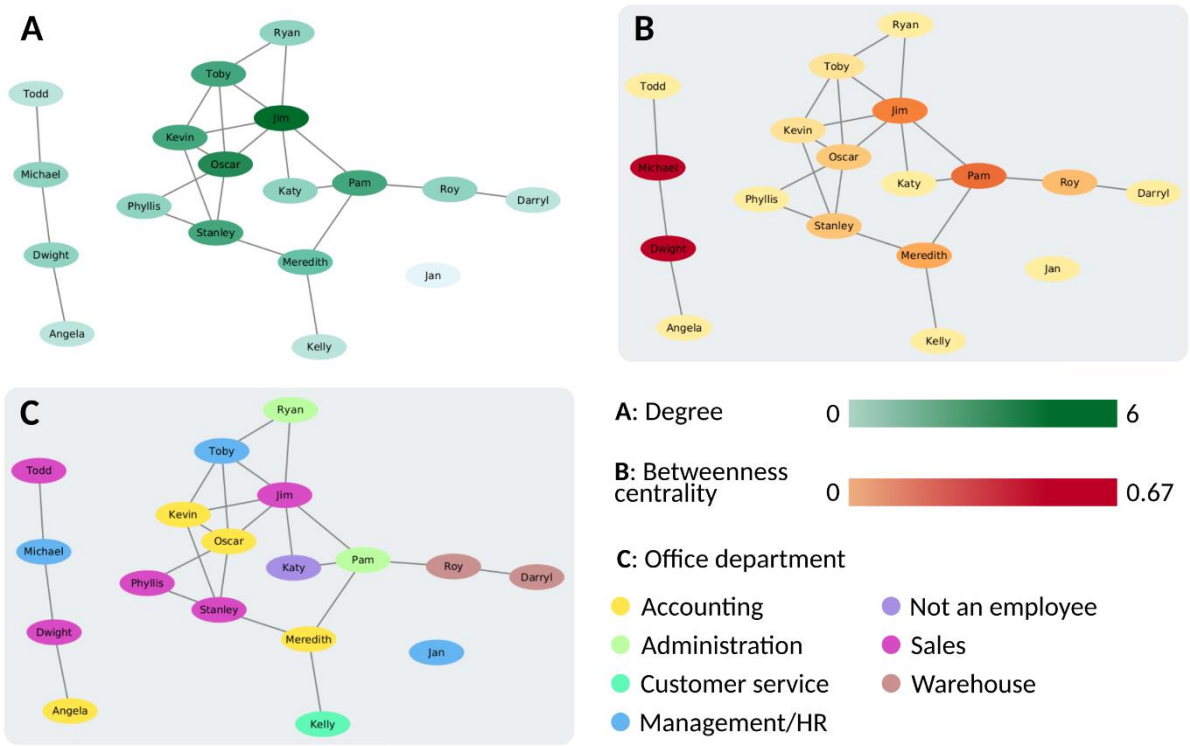


Figure 10: Visual representation of different characteristics of a dataset.
(**A**) The Office network with nodes coloured according to their degree, indicating how many other nodes they are connected to.
(**B**) The same network with nodes coloured according to their betweenness centrality, indicating the importance of each character for the overall cohesion of the social structure.
(**C**) The same network with nodes coloured according to the department each character works in. This shows both the size of each department, and the fact that coworkers form relationships with people across specialisms.

Structural features and patterns of connectivity in the resulting network reflect certain features of the social community. Some people concentrate many positive relationships, such as Jim the friendly salesman or Pam the quirky receptionist, while others seem generally disliked; secondary characters, usually working in a different place to the main office (e.g. Todd and Katy), only relate to a small number of employees in the main cast, and are mainly indifferent to the rest. In network terms, popular characters are represented by nodes with a high degree (the number of adjacent edges), and less popular ones with lower degree nodes (Figure 10A). Distinct social groups can also be identified as disconnected clumps, i.e. connected components: the largest component gathers most characters

depicted rather positively, whereas four ‘antagonists’ are grouped separately, and Jan the upper manager forms a singleton (an unconnected node) as she is not shown to be friendly with anyone.

Other network properties can illustrate other social dynamics in interesting ways. For instance, we can test whether “the friend of my friend is my friend” holds true, by looking at trios of nodes: Jim is friends with Pam, who is friends with Roy, but Jim and Roy do not seem to be friends (we can say that they form a non-transitive triplet). This notion of (in)transitivity also suggests that if people can belong to the same social group only because of common friends, or even friends-of-friends, then some people (intuitively among the more popular ones) may act as ‘social glue’, centralising the cohesion of their group that would perhaps be more fragmented in their absence. Such people are represented by nodes with a high betweenness centrality, also called hubs, in the social network (Figure 10B). For instance, Dwight has a betweenness of 0.67 because he is in the middle of two out of three interactions between other people in his social cluster (he mediates Angela-Michael and Angela-Todd interactions, but not Michael-Todd).

Analysing the topology of a network, i.e. the structure of its nodes and edges, can therefore provide significant quantitative information about the underlying data. But another strength of network analyses is that this information can be conveyed visually by different graphical features of the network’s image representation. Playing with specific visual properties, such as node size, colour, or shape, can help to highlight specific properties. Making nodes bigger when they have higher degrees, for instance, is a common way to bring attention to the most ‘active’ agents of a network. Qualitative information can also be mapped to the network representation, to enrich the image with additional data. For instance, colouring nodes in the *Office* network based on the department of their character (sales, accounting, customer service, management, etc.) shows that these departments vary in size, and that bonds between coworkers are not necessarily restricted to one’s own job type (Figure 10C).

The layout of a network (i.e. the spatial distribution of nodes in the 2D plane) is also an important visual vector of information. Because our minds rely on pattern recognition for visual cues, we intuitively expect co-located nodes to be strongly connected (and vice-versa), with central nodes in the middle of the grouping and peripheral nodes closer to the edge. This visual proximity bias can sway our understanding of the data being depicted: artificially placing the node representing Angela at the centre of the network, for instance, could fool us into thinking that she is unanimously liked by her coworkers. Selecting an appropriate layout algorithm that accurately depicts the information we wish to convey from a network analysis is therefore essential. Most network layouts used for large datasets try to minimise the distance between closely connected nodes, usually by mimicking the

stabilisation of a physical system: edges are modelled as springs that can pull together nodes or push them away, and the layout algorithm simulates this system until reaching a stable conformation.

The graphical depiction of a network can be a useful tool to guide the analysis of a dataset, as it offers visual support to make hypotheses based on observed trends. It can reveal patterns that were not initially anticipated and that may not have been considered in the initial collection of data. Still, for large datasets, the visual representation of a network may not be able to convey the full complexity of the data: in a similar way to a principal component analysis, a network layout is only a 2D projection of a dataset with a potentially much higher dimensionality. In many cases, the qualitative approach must be complemented by computational and statistical work to draw more objective conclusions from the network. The underlying paradigm is that relationships between elements of a dataset are relevant to the characteristics of these elements, and contribute to shaping the data itself. Therefore, more than being simple surface-level descriptors of a dataset's contents, network representations can provide pertinent insights into its specificities and dynamics. This is the conceptual basis of network science, which is the quantitative study of relational data, as well as the main lens through which we approach biological data in this work, primarily by constructing and analysing sequence similarity networks [Watson et al. 2019].

3.2 – Constructing sequence similarity networks

As we have established in previous sections, sequence comparisons are now ubiquitous in modern biology. Whenever a new set of sequences is generated, for instance, one of the first steps of analysis usually consists of comparing them in an all-against-all pairwise alignment. This is generally done in order to pool sequences into groups of high similarity [Zou et al. 2020]. This is allows data inspection at a higher level of abstraction, for instance by constructing Operational Taxonomic Units to work at the level of species (or higher) rather than individual sequences [Blaxter et al. 2005]. This clustering can also be useful to minimise the computational load by dereplication: groups of identical or near-identical sequences are reduced to a single representative, under the assumption that the informational loss of removing this redundancy is negligible [Fu et al. 2012].

Underlying each of the examples above is the issue of the general organisation of a sequence dataset: how are sequences similar and distinct from one another? Am I working with a mere handful of major archetypes, or a constellation of small unrelated sequence groups? These questions pertain to the overall structure of the data at hand, and as we have argued earlier, network representations are well suited to address these – and specifically, because we are talking about pairwise comparisons between biological sequences, sequence similarity networks.

Sequence similarity networks (SSNs) consist of nodes representing sequences from a given dataset, and of edges linking pairs of sequences that meet a predetermined criterion for similarity. This criterion can in theory be anything, but will most frequently correspond to thresholds on one or more metrics that can be applied to pairwise sequence alignments (e.g. E-value, alignment identity, alignment coverage). Indeed, SSNs are usually constructed by performing an all-against-all BLAST alignment on the dataset (using the same set as both query and target sequences). This produces a list of pairs of sequences that were successfully aligned (Figure 11A), with (if specified prior to execution) the corresponding metrics for each alignment. Any alignment reported by BLAST that fails to meet the thresholds determined for the specific purpose of the SSN is then discarded.

A

qseqid	sseqid	evalue	pident	qcov	scov
seq1	seq1	6E-222	100	100	100
seq1	seq3	1E-53	32.7	100	100
seq1	seq2	4E-49	31.7	95.6	95.4
seq1	seq4	2E-32	28.1	88.4	91.6
seq1	seq5	2E-25	29	100	99.3
seq2	seq2	1E-214	100	100	100
seq2	seq3	2E-57	37.3	99.3	100
seq2	seq1	4E-49	31.7	95.4	95.6
seq2	seq5	6E-32	30.8	98.2	97.5
seq2	seq4	1E-20	27.4	81.9	83.9
seq3	seq3	5E-218	100	100	100
seq3	seq2	1E-60	37.3	100	99.3
seq3	seq1	7E-56	32.7	100	100
seq3	seq5	1E-26	24.9	99.3	97.8
seq3	seq4	1E-25	27.2	80.6	81.6
seq4	seq4	6E-201	100	100	100
seq4	seq1	1E-34	28.1	91.6	88.4
seq4	seq3	4E-25	27.8	80.1	79.2
seq4	seq5	1E-22	28.5	96.2	92.1
seq4	seq2	2E-22	27	99.2	95.7
seq5	seq5	2E-214	100	100	100
seq5	seq2	6E-31	30.8	97.5	98.2
seq5	seq1	1E-24	29	99.3	100
seq5	seq3	4E-24	25.7	93.2	94.6
seq5	seq4	2E-20	28.5	92.1	96.2

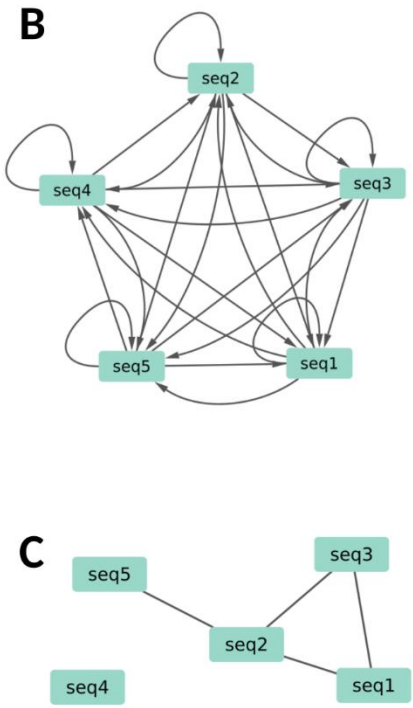


Figure 11: Constructing a sequence similarity network from a tabular BLAST output.
(**A**) An example tabular BLAST output, as in Figure 6, with added columns representing the coverage of each alignment (in %) on the query (**qcov**) and subject (**scov**) sequences.
(**B**) Taking the raw BLAST output as a list of edges produces a network with duplicate edges (bidirectional hits) and self-loops (self-hits).
(**C**) The network in (B) is filtered to remove self-loops, bidirectional edges and hits with less than 30% amino-acid identity, resulting in a 'clean' sequence similarity network.

At this stage, all remaining alignments correspond to pairs of sequences that we do wish to be linked in the SSN, with a couple of caveats. First, because the entire sequence dataset is used as both query and target by BLAST, each sequence will be compared to itself once in the process, resulting in self-alignments with perfect scores that obviously pass any criterion for similarity (Figure 11B). These

'self-loops' in the BLAST output are entirely uninformative from a biological standpoint, and should therefore be discarded. Second, because the BLAST alignment process is asymmetrical, each pair of (distinct) sequences is actually compared twice: when sequence X is used as query, it is compared to all other sequences including sequence Y, and conversely when Y is the query it is also compared to X. This produces, in the 'unfinished' SSN, two directed edges between X and Y, one in each direction if raw alignment outputs are read as being oriented query-to-target. These two alignments can have slight differences of E-value or base identity due to the nature of the BLAST algorithm – in rare cases, the similarity criteria can even retain one edge but not the other, when the gap between their scores overlaps the thresholds considered. Again, this does not have any biological meaning, as we intuitively want similarity relationships to be reciprocal, and so duplicate edges should also be removed, usually by keeping the highest-scoring alignment out of the pair (Figure 11C).

In short, the construction of a SSN from a raw BLAST output can be carried out following three main steps once the similarity criteria have been set: (i) apply the criteria to remove alignments of insufficient quality, (ii) delete self-loops, and (iii) remove duplicate edges by discarding the weakest alignment of a bidirectional alignment pair. Of course, the key step in this process is the definition of similarity criteria that are suitable for the eventual purpose of the SSN, as their stringency or leniency dictate entirely the density of edges in the network as well as their signification. Extremely strict thresholds on sequence identity, for instance, will create sparser networks (possibly with many connected components) where edges represent remarkable similarities between closely related sequences, whereas more relaxed thresholds will yield dense networks connecting sequences more distantly related. In real-world biological data, there is rarely a clear-cut threshold of sequence identity or E-value guaranteeing that all sequences above the threshold, and only those, are homologous. It is therefore for the biologist to decide how strict the criteria for constructing the SSN should be, to best mitigate the risks of including false-positive similarities, and of excluding tenuous but real homologies (Figure 12).

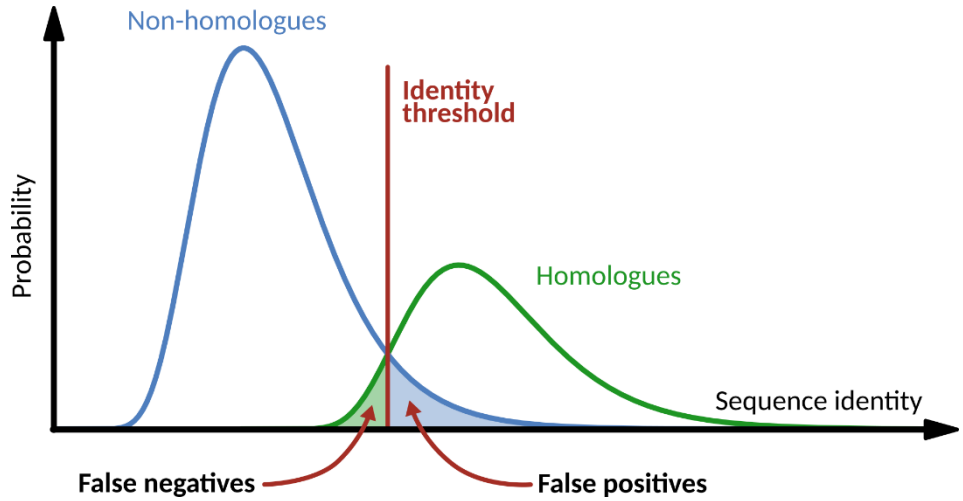


Figure 12: Selection of a threshold for filtering sequence alignments.
The overlap in sequence similarity between homologues and non-homologues makes the choice of an appropriate threshold crucial.

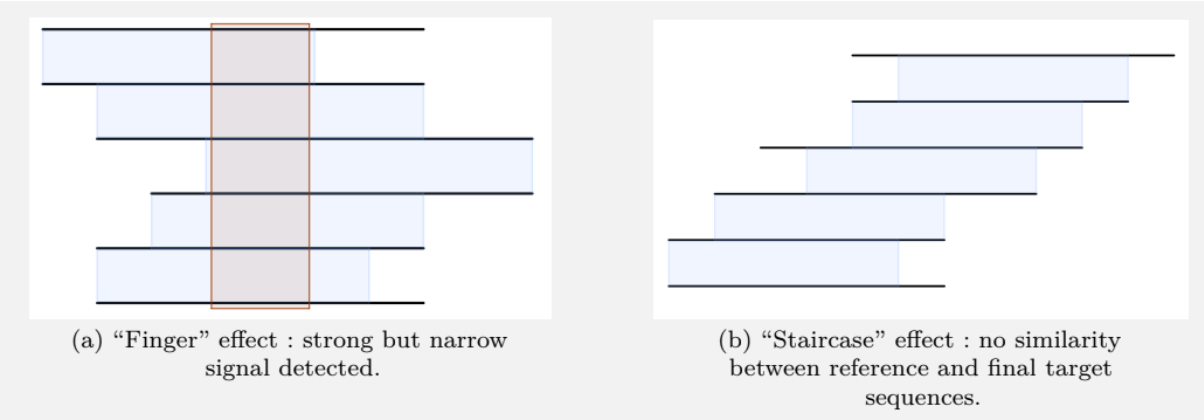
3.3 – Similarity networks and sequence families

In an earlier section, we discussed the connection between sequence similarity and sequence homology, and we came to the conclusion that although one does not necessarily imply the other, similarities between sequences could generally be considered as a sign of homology provided they rely on adequate thresholds of alignment quality. Therefore, under this assumption, by constructing similarity sequence networks, are we then not constructing sequence homology networks too? In that case, if we apply criteria coherent with sequence homology when constructing a SSN, what can it then tell us about evolutionary relationships between the sequences it represents?

The implicit expectation behind the definition of homologous gene families (as sets of genes that share a common ancestor) is that all genes in a family should have the same evolutionary history. In other words, those genes should all be homologous to each other even if they cannot be aligned together directly: the similarity network of that family might not be a fully connected clique, but the underlying, hypothetical homology network is. Even if two homologous genes have diverged too much to be readily aligned, they might both still have some similarity with a common neighbour, a sort of intermediate sequence bridging the gap between the distant homologues. Taking this idea further, distant homologues might not even have a direct common neighbour but may be linked by a longer chain of intermediate homologues. Thus, in an SSN, any two sequences connected by either a direct edge or a longer path would be considered homologous: connected components of the SSN therefore delineate exactly the different gene families in the dataset. This approach to reconstructing gene

families is called *single-linkage*, and aims to permit the identification of homology relationships beyond the detection scope of sequence alignment.

In practice, applying a single-linkage protocol to reconstruct sequence families is likely to connect sequences that are barely homologous, if at all. A number of pathological patterns can emerge from chains of sequence alignments that do not conform to the expectation stated above for a common evolutionary history (Figure 13). This is true even with strong constraints on the proportion of each sequence that must be covered by an alignment, such that this mode of reconstructing gene families is only really suitable for strongly conserved, evolutionarily stable families. Indeed, we used this compatibility with conserved sequences to expand gene families with distant homologues, by applying additional constraints to link together sequences in SSNs. This is the object of the next chapter of this thesis.



(a) "Finger" effect : strong but narrow signal detected.

(b) "Staircase" effect : no similarity between reference and final target sequences.

Figure 13: Anomalous patterns in chained sequence alignments.
Each black line represents a sequence, and the aligned regions between two adjacent sequences are shaded in blue.
(**A**) Sequences can align over different regions, such that only a narrow portion of each sequence is really common to all.
(**B**) Successive alignments can result in a sliding or staircase pattern, connecting distant sequences without any actual correspondence between them.

The aforementioned imposition of a high mutual coverage in sequence alignments (and, even before that, the definition of gene families as sequences with the same evolutionary path) *de facto* excludes partial homology relationships: if two gene families share a portion of their length but are otherwise unrelated, then the single-linkage approach to constructing gene families will simply not reflect this partial homology information. A part of the evolutionary history of these gene families is therefore overlooked, assuming that this shared region (e.g. a common protein domain) is indeed descended from the same ancestor in both families. To take into account such relationships, the method of constructing SSNs must be adapted to reflect the plurality of sequence similarities, with

partial or full coverage of the aligned sequences. Using such similarity networks that account for partial homology, in order to study the dynamics of gene remodelling and combinatorial evolution, is the focus of the third chapter of this thesis.

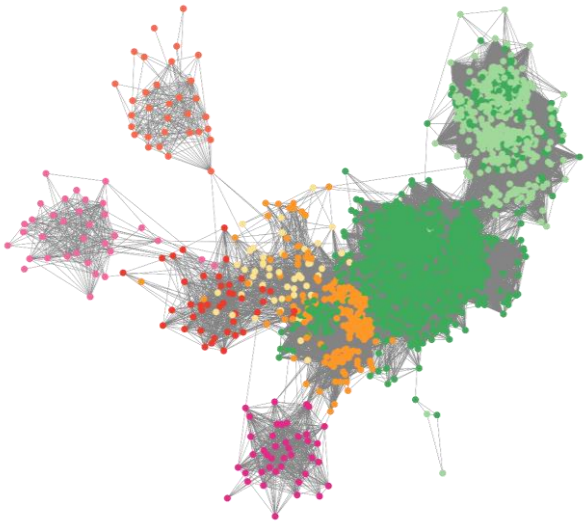


Figure 14: A sequence similarity network representing the SMC protein family.
Nodes in the network are coloured according to their Domain of life. Pale green indicates sequences of CPR bacteria, and other bacterial sequences are shown in dark green. Similarly, DPANN and other Archaea are represented in pale and bright yellow respectively. Lastly, shades of red correspond to four different SMC paralogues in Eukaryotes.
The topology of the network reflects the evolution of SMC proteins: CPR and non-CPR bacteria are grouped separately, reflecting the divergence between these two bacterial groups; eukaryotic sequences clearly cluster with the same paralogous copies; the small clump of green near Archaea represents sequences from Cyanobacteria, which are suspected to have acquired their SMC gene in a HGT event of archaeal origin.

As discussed earlier (in the context of OTUs and sequence dereplication), analysing large amounts of data will often call for a reduction of the complexity and dimensionality of the dataset in order to operate at higher levels of abstraction. With the massive accumulation of biological data in the past few years, SSNs (like any other bioinformatic tool) are now being used to handle large sets of sequences that can contain tens of thousands of nodes, and thus edges numbering in the (tens of, hundreds of) millions. A higher-order view of relationships in the dataset is therefore imperative, which brings us back to the importance of identifying highly cohesive groups of sequences in the network. The single-linkage principle is one such approach, but it can often fail to reach the necessary level of granularity, especially because large SSNs frequently contain a 'giant' connected component that concentrates a majority of the nodes [Newman, Strogatz, and Watts 2001, Halary et al. 2010]. Therefore, less blunt methods can also be fruitful, by recognising that attachment within connected components is not random, such that several tightly knit groups of sequences can exist in the same component. In this way, the overall structure of an SSN reflects the underlying organisation of its gene

set, beyond simple connectivity (Figure 14). Clustering algorithms can identify coherent groups of preferential attachment in large networks, allowing the subdivision of SSNs into clusters of similarity that can, for instance, distinguish between gene families that were connected together in a giant component. In later chapters of this thesis, we discuss several uses of network clustering in SSNs, in particular to identify divergent variants within gene families, and to delineate families in networks of partial homology.

4. Reconstructing the evolutionary history of poorly characterised proteins

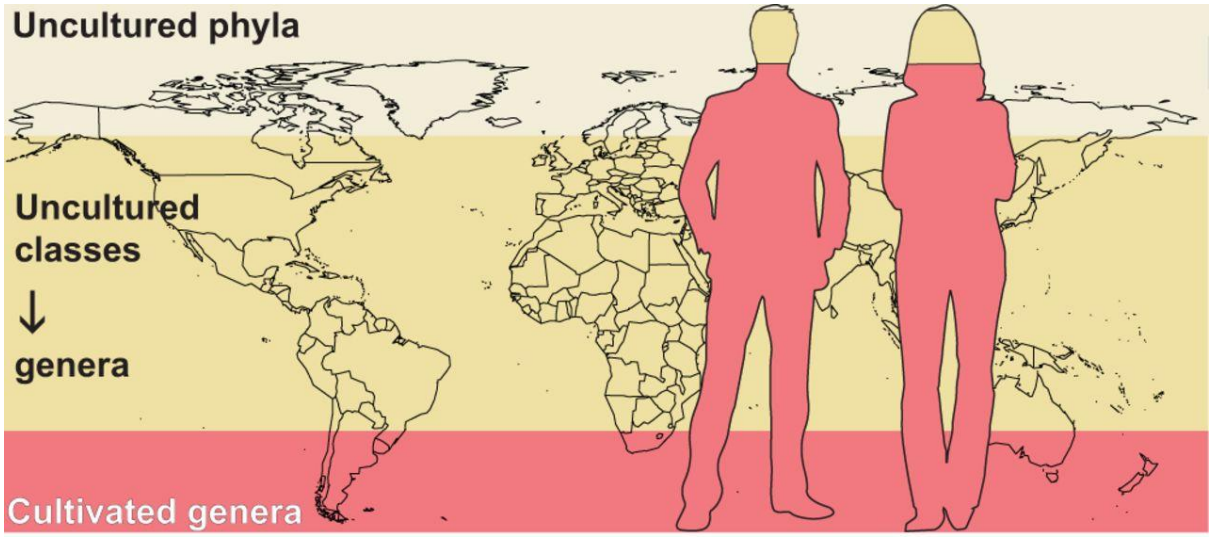


Figure 15: Proportion of uncultured cells in natural and human-associated environments.
Only 19% of Earth's microorganisms belong to cultivated lineages, with 25% being from phyla with no cultured representative. In human and human-associated microbiomes, more than 80% of cells are from cultivated genera.
Data from [Lloyd et al. 2018], illustration from [Hug 2018].

Although the genomics revolution that ushered in a boom in the number of gene and genome sequences available allowed unprecedented insights into the diversity and evolution of living organisms and their function, it also revealed a perhaps unsuspected complexity of the gene universe. In particular, vast amounts of gene sequences remain poorly characterised to this day, even in the core genomes of intensively studied organisms such as the Escherichia coli K12 strain that is used in countless microbiology labs [Cummins et al. 2022]. In addition to ORFans, i.e. genes encoded in only one genome with no apparent orthologue, sequence databases are rife with hypothetical proteins that have been predicted by computational methods in several genomes but never confirmed experimentally, often containing DUFs that hinder attempts to understand their function. Even more striking is the abundance of uncharacterised sequences in environmental metagenomes. While host-

associated microbiomes (such as the human gut microbiome that is highly popular in the media) are relatively well understood, at least in their composition, studies of microbiomes in natural environments paint a different picture, in which genes of unknown taxonomical origin and/or biological function make up the overwhelming majority of the microbial genetic space (Figure 15).

These uncharacterised genes, performing undescribed functions in microorganisms well known or otherwise, constitute a major lacuna in our understanding of the diversity of life and its molecular processes. The difficulties encountered by canonical annotation methods (generally based on sequence orthology inferences) to qualify these sequences suggest that maybe their divergence stems from unusual evolutionary trajectories beyond the scope of those methods. If that hypothesis is true, then alternative approaches targeted to address specific causes of sequence divergence may be well suited to complement more general models.

4.1 – Distant homologues fly under the radar of sequence alignment

In previous sections, we discussed the links between sequence similarity and homology, and in particular how sequence alignments are the main empirical data used to infer evolutionary relationships between genes. However, we also mentioned that genes can have low sequence similarity (resulting in failed or low-scoring alignments) but still be distantly homologous.

Perhaps one of the best ways to illustrate distant homology is to consider the case of ancestral gene families that appeared early in the history of life on Earth, typically prior to LUCA (the Latest Universal Common Ancestor) and the separation of Archaea and Bacteria, some 3.5 billion years ago (to use a conservative estimate). A large portion of these genes are present in all major lineages of cellular life, and involved in fundamental biological processes such as information processing (transcription, translation) and DNA maintenance. Sequences of such key genes generally evolve under strong forces of purifying selection and consequently diverge more slowly than the average of gene families. Yet, because of their remarkable evolutionary age, the gradual accumulation of mutations in their sequences can diminish the sequence similarity between ancestral genes in distantly related organisms, beyond the scope of detection by sequence alignment.

The erosion of sequence similarity within a gene family can also occur when genes on a specific branch of the family develop faster rates of mutation than their counterparts. This is commonly observed, for instance, in the aftermath of a gene duplication event, where the increase in copy number relaxes the purifying pressures on a gene sequence, allowing one of the paralogues to accumulate mutations and develop a new function while the other copy retains its original role. This

rapid divergence in primary sequence can also appear between orthologues present in different lineages, e.g. when changes in selective pressure can quickly favour beneficial adaptive mutations.

Remote homology can be problematic when trying to explore a new sequence dataset that is yet to be annotated. Most annotation processes in biology rely on previous knowledge, e.g. a reference database of functionally resolved genes, and target genes are compared directly against references. Direct homologues of reference genes are easily identified and given an annotation, but the target database may also contain genes that are only distant homologues of the reference set – for instance, because they correspond to unknown paralogues or come from divergent lineages. These indirect homologues might not be picked up by simple alignment-based searches from the reference database, despite their evolutionary connection to them. This is especially problematic when considering that our current knowledge of the extant diversity of genomes and life forms on Earth is far from complete. In most environmental metagenomes, for instance, sequences that can confidently be mapped back to well-characterised lineages and functions only represent a small fraction of the entire microbial diversity. Many important discoveries have come from exploring this "microbial dark matter" [Marcy et al. 2007, Rinke et al. 2013], but many unknowns remain about the full nature of microbial life on Earth [Bernard et al. 2018]. Therefore, if this partial knowledge is our basis for studying and understanding the biological world, then the insights we gain from it may not be fully comprehensive. Methods that are able to address this distant homology in an efficient and reliable way could therefore improve biological knowledge, by providing a comprehensive picture of the diversity of gene families.

4.2 – Remodelled genes and the combinatorics of evolution

In addition to tree-like evolutionary processes that occur within the boundaries of gene families, such as vertical modifications, duplications or horizontal transfers, genes can also undergo more combinatorial processes that involve 'subunits' of several genes from unrelated families. In particular, genes can merge with others, split into several independent genes, or recombine regions of their sequence into new arrangements during the course of evolution. When thinking about genes as assemblages of protein domains [Forslund, Kaduk, and Sonnhammer 2019], for instance, this idea of cross-combination between the contents of different genes can help to explain the modular nature of multi-domain proteins as well as the sometimes patchy phyletic distribution of domains in distantly related proteins. In this work, we focus specifically on gene fusions and gene fissions, which we group together under the term of gene remodelling events, although we recognise that other kinds of combinatorial processes exist in gene and protein evolution.

Events of gene remodelling can occur in different ways following a change in genomic organisation [Marsh and Teichmann 2010, Leonard and Richards 2012]. The loss of a stop codon and/or a larger intergenic region, for instance, can merge the sequences of two adjacent genes that will now be transcribed as one (an event called gene fusion). Conversely, the emergence of a stop codon and a new transcription initiation sequence inside a gene can split it into two new distinct genes (gene fission). In addition to these 'local' events, gene fusions and fissions can also result from broader scale chromosomal rearrangements, such as translocations that can bring together genes that were formerly sitting at distant places in the genome.

The primary lens through which gene fusions and fissions are perceived is that of protein domain rearrangements. Indeed, multi-domain proteins represent the majority of proteins both in Eukaryotes and prokaryotes, and their functional coherence allows for many insights into the role that those proteins play in different biological processes. On the other hand, domains do not represent the full extent of biological sequences, and many CDS are not covered by any domain [Mistry et al. 2021]. For this reason, they provide a good but only partial picture of the evolutionary significance of gene remodelling. More comprehensive models, taking a more systematic approach to the characterisation of partial homology, can complete this picture and describe in further detail the dynamics of combinatorial processes as a whole. There also exists another source of bias in a number of gene remodelling studies, which consists in a heavy focus on gene fusion events, sometimes to the detriment of gene fission. Genes that have partial homology to two separate gene families are sometimes automatically considered as fused, even though in reality some may have been split by a gene fission event, which gave rise to the other two families. Avoiding this pitfall is necessary in order to present an accurate view of combinatorial evolutionary processes.

5. Aims of this doctoral thesis

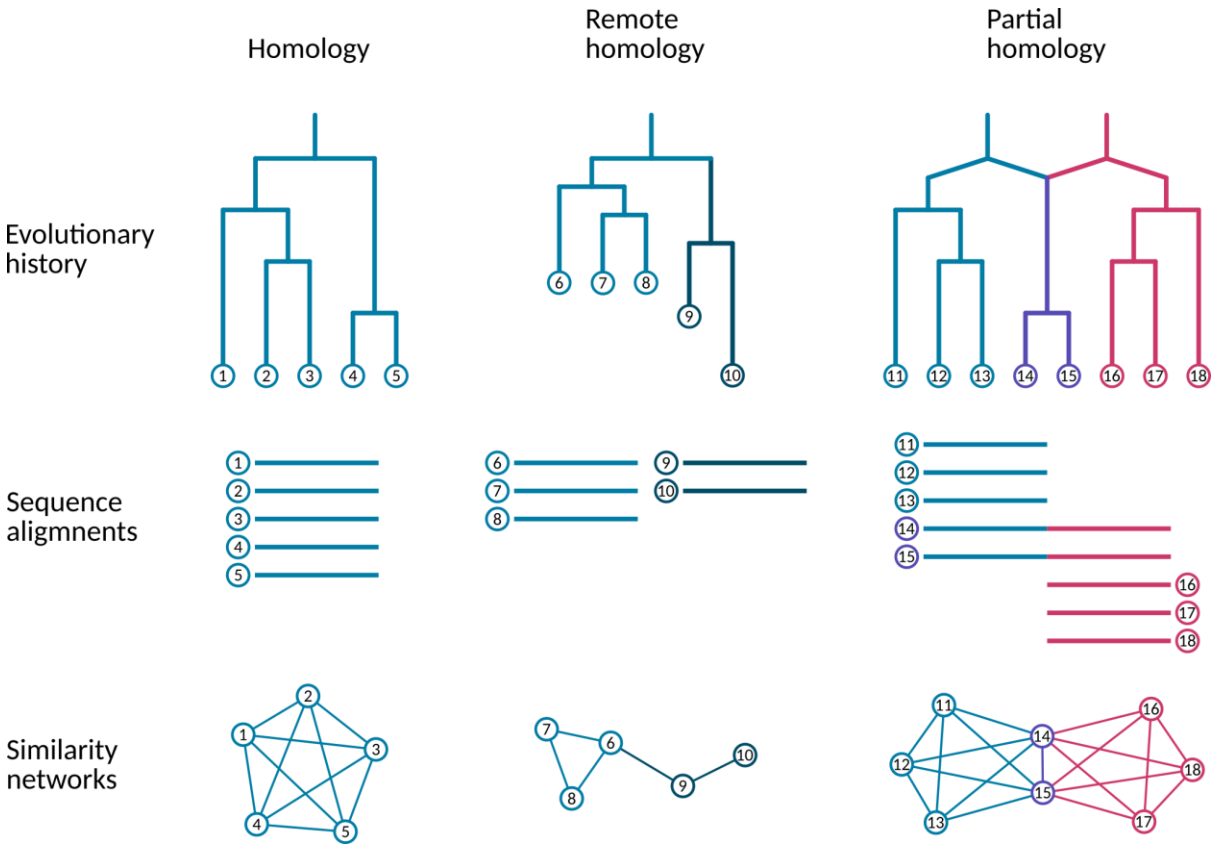


Figure 16: Canonical, remote and partial homology.
The three main types of homology that we discuss in this thesis produce different alignment patterns and therefore different motifs in sequence similarity networks.

During the three years of my doctoral research, I have developed and applied new methods, based on network science and in particular sequence similarity networks, to study two main types of evolutionary relationships between genes: distant homology, and partial homology from gene remodelling events.

In Chapter II of this thesis, I detail the work that we conducted on using sequence similarity networks to identify and describe distant homologues of known gene families in an environmental dataset. We sought to measure the genetic and phylogenetic diversity of highly conserved gene families, typically as old as cellular life, when uncultured organisms are taken into account. In particular, we were interested in finding divergent groups of sequences compatible with new microbial lineages branching near the root of the tree of life. To that end, we performed iterative homology searches, from a set of reference ancestral gene families, in a large oceanic metagenome. We showed that many of these families have important groups of divergent homologues in the global ocean microbiome, and that new major discoveries remain possible from microbial dark matter. We

found, in particular, a new putative paralogue of SMC proteins in Actinobacteria, with divergent structural features that are likely to indicate an alteration of the way this protein interacts with DNA. We also identified vast amounts of uncharacterised genetic diversity in DNA clamp-loading subunits, as well as in recombinases. In addition to reporting these results, we also prepared the publication of the computer programme that performs these distant homology searches.

Furthermore, we used sequence similarity network analyses to detect gene fusions and fissions, which is detailed in Chapter III of this thesis. We studied these remodelling events in two different lineages of eukaryotes that independently evolved a multicellular life (brown algae and animals), and we sought to understand the effects of gene remodelling on the emergence of complex multicellularity. In both of these studies, we combined network information and phylogenetic signal to 'polarise' the inferred remodelling events, distinguishing between gene fusion and fission. We found that fusions were slightly more frequent than fissions in brown algae, and that the majority of these events occurred in the early stages of their evolution. The genetic products of fusions and fissions only represented a small portion of all brown algae genes, but they tended to be more retained than non-remodelled genes in extant genomes. In animals, we found that fusions were significantly more prevalent than fissions, and that bursts of gene fusions occurred at key nodes of animal evolution. Additionally, many gene fusions appeared convergently in several places of the animal phylogeny, in a pattern of repeated evolution of successful innovations. These results on gene remodelling in two different lineages allow us to draw comparisons between how remodelling might have contributed to each of their independent emergences of multicellularity.

Overall, this research highlights the multifactorial nature of evolutionary processes beyond conventional models of gradual and arborescent evolution, and demonstrates the importance of taking this diversity of processes into account when trying to understand biological sequences in a more comprehensive manner.

Chapter II. Remote environmental homologues of conserved protein families

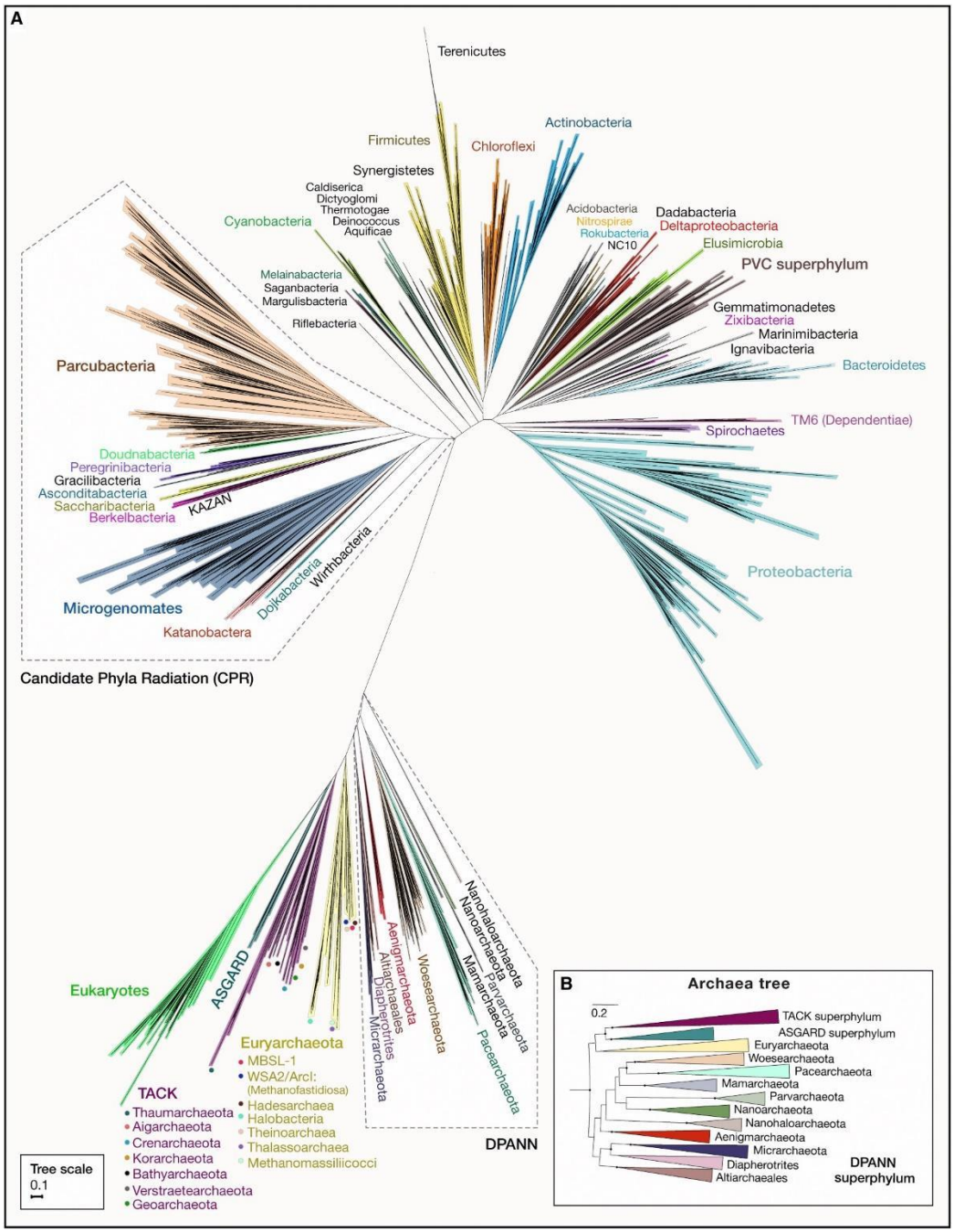


Figure 17: Uncultivated lineages enrich the modern Tree of life.
(**A**) Phylogenetic tree of all known major lineages, with CPR bacteria and DPANN archaea highlighted, constructed from a concatenated set of 14 ribosomal proteins.
(**B**) Reconstructing a phylogenetic tree with only archaeal sequences restores the monophyly of the DPANN clade, which was not monophyletic in the full tree.
From: [Castelle and Banfield 2018].

1. The great unknowns of environmental genomics

1.1 – Microbial dark matter: the unseen majority

Microbial life forms are virtually everywhere, populating just about every single corner of Earth, from our own bodies to the deepest waters of the oceans. Despite their extremely low individual mass, microbes as a whole (which includes archaea, bacteria, viruses and unicellular eukaryotes) represent the second largest share of global biomass, far behind plants but far ahead of animals and fungi [Bar-On, Phillips, and Milo 2018]. The tree of life is also dominated by microbial lineages, relative to which multicellular organisms only represent a tiny fraction of the overall phylogenetic diversity.

Our knowledge of microorganisms still, to this day, derives primarily from strains that can be isolated and grown in laboratory conditions. The historical reasons for this are numerous, starting with the fact that the precursors to modern microbiology were largely concerned with the study, prevention and healing of infectious diseases. In the late 19th century, Robert Koch formulated a series of principles for establishing a causal link between a microbe and a disease, in which he stipulated the need to isolate and grow the microbe in pure cultures [Koch 1877]. Unquestionably, microbial cultures remain entirely relevant to today's biology, as they provide unparalleled insights into the functioning of microorganisms. However, the advent of environmental genomics, starting in the 1990s, revealed a strikingly large diversity outside the scope of culture-based studies, and led to

the realisation that traditional cultivation techniques are only compatible with a small minority of microbial life forms [Staley and Konopka 1985, Whitman, Coleman, and Wiebe 1998]. Surveys of 16S rRNA diversity in natural environments have repeatedly found large numbers of OTUs that could not be confidently assigned to any of the known bacterial and archaeal phyla. From the early 2000s, the improvements and democratisation of cultivation-independent shotgun sequencing and high-throughput sequencing allowed biologists to sequence the entire DNA contents of an ecosystem at once, rather than single-gene amplifications. In these metagenomes too, the vast majority of genes proved difficult to annotate, taxonomically and/or functionally. This is especially true of sequences derived from natural environments, although host-associated and man-made microbiomes also harbour significant amounts of genes of unknown origin and function [Lloyd et al. 2018]. This colossal repertoire of untapped microbial diversity is collectively referred to as "microbial dark matter" (MDM), in direct analogy to the dark matter of the cosmological kind. The term is somewhat loosely defined, and can refer either to the set of microorganisms that do not belong to any well-established lineage (MDM in the cellular sense), or to the set of genome sequences with elusive taxonomical origin and biological function (MDM in the molecular sense). Still, microbial dark matter is a useful shorthand for the vast diversity of unknown microbes and microbial genes that may contribute to ecosystems in unsuspected ways.

The genomic content of MDM can be unravelled to access the genes and genomes that are at play in microbial ecosystems. From raw metagenomic reads (typically 100-500 bases long), the sequencing data is filtered and processed in order to discard low-quality sequences, as well as possible contaminant DNA (e.g. host DNA in human gut microbiomes). The remaining sequences are then assembled into longer contigs, based on overlapping regions between reads. Coding DNA sequences (CDS) can be detected from these contigs to gain insights into the ecological composition and function of the sequenced microbial community. The accuracy of this assembly step is therefore of particular importance, especially to avoid producing chimeric contigs that merge sequences from different organisms. The identified CDS can then be curated into a clean metagenome that contains all microbial genes detected[6] in the sample. Although already informative in itself, the gene pool of a sampled biome can be further studied by binning contigs in order to reconstitute the genomes of sequenced organisms. In addition to bringing additional hierarchy to otherwise unstructured metagenomes, these metagenome-assembled genomes (or MAGs) allow for a deeper understanding of *in situ* microorganisms and their diversity, for instance by enabling phylogenetic reconstructions from

[6] This generally represents an underestimation of the genetic diversity present in the sample: genes of low-abundance organisms can go undetected if the sequencing coverage is insufficient, and assembly algorithms can blend intraspecific variations in gene sequence.

concatenated gene sets. MAGs of uncultured microbes are of particular significance for MDM research, as they represent (along with single-cell amplified genomes) one of the only ways to bypass the prerequisites of cultivation for genomic analysis.

1.2 – The tree of life in the light of uncultured organisms

With the developments of environmental genomics, MDM is now far from entirely inscrutable, and has proven to be a formidable source of biological discovery in the past two decades. By turning MDM from a data-poor to a data-rich field of research [Jiao et al. 2021], cultivation-free sequencing has allowed for new major perspectives on our fundamental knowledge of life on Earth (Figure 17).

Some twenty years ago, it was largely undisputed that the range of prokaryotic cell sizes hardly overlapped that of viral capsids [Koonin and Yutin 2019]. This assumption, however, was disproved by the discovery of CPR bacteria [Brown et al. 2015] and DPANN archaea[Baker et al. 2010, Rinke et al. 2013], two novel prokaryotic lineages with ultra-small cell diameters of around 0.2 µm. Although they belong to different Domains of life, CPR and DPANN share a number of common features in addition to their nanoscopic sizes [Castelle and Banfield 2018]. Both are remarkably diverse, and are largely accepted as forming distinct superphyla in their respective kingdoms (although the monophyly of DPANN is somewhat less evident than that of CPR), with CPR representing somewhere between 15-50% of all bacteria. The genomes of both CPR and DPANN have undergone significant reduction, and are typically only 0.5 to 1 Mbp long, in line with other prokaryotes that live obligate symbiotic or parasitic lifestyles [Castelle et al. 2018]. In comparison, the alphaproteobacterium "*Candidatus* Pelagibacter communis"[7] has one of the smallest known genomes of free-living organisms at 1.3 Mbp, which is already considered an advanced level of genome streamlining [Giovannoni et al. 2005]. In this genome reduction, most CPR and DPANN have lost metabolic pathways that are essential for self-sufficient lifestyles, including *de novo* biosynthesis of amino-acids, nucleotides, and fatty acids (key components of cellular membranes). The extent of loss in metabolic capacity varies between different groups, but it is expected that most CPR and DPANN are reliant on other microorganisms for a number of essential biochemical resources, mediated via episymbiotic lifestyles (i.e. an attachment to the outer membrane of a host) (Figure 18A-B).

[7] Incidentally, "Ca. P. communis" also has some of the smallest cell dimensions for non-symbionts, with a rod-like shape of roughly 0.8 µm in length and 0.2 µm in diameter. It is alternatively known as "*Ca*. P. ubique" due to its extreme abundance in both salt and freshwater environments worldwide, making up 25% to 50% (in summer) of all microbial cells in temperate ocean surface layers.

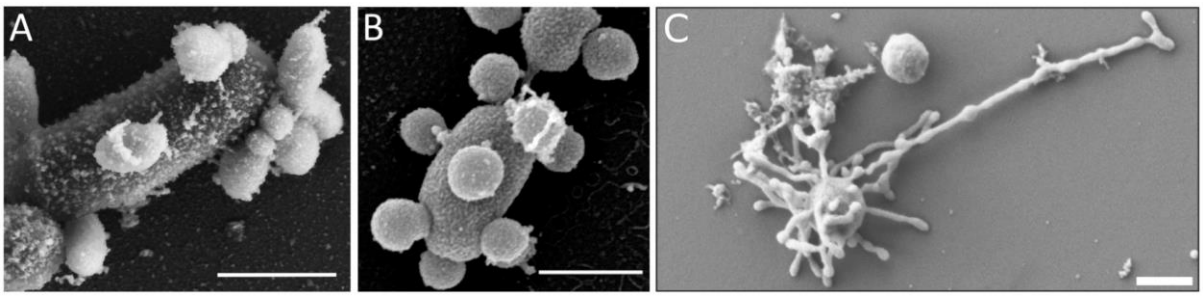


Figure 18: Electron microscopy images of CPR bacteria, DPANN archaea and Asgard Archaea.
(**A**) CPR bacterium *Nanosynbacter lyticus* parasitising its bacterial host *Actinomyces odontolyticus*.
(**B**) DPANN archaeon *Candidatus Nanohalobium constans* parasitising its archaeal host *Halomicrobium* sp.
(**C**) Asgard archaeon *Ca. Lokiarchaeum ossiferum* showing multiple membrane protrusions.
On all images, the scale bar indicates a size of 500 nm.
From: [López-García and Moreira 2021] (A, B) and [Rodrigues-Oliveira et al. 2023] (C).

The overlap between prokaryotic and viral sizes is further amplified by the discovery of giant viruses that can reach up to 1 µm in diameter in the phylum *Nucleocytoviricota*, or NCLDV (nucleocytoplasmic large DNA viruses) [Iyer et al. 2006]. These include the *Poxviridae*, responsible for multiple human and animal diseases such as smallpox and mpox, though by far the largest viruses are found in the families *Mimiviridae*, *Pandoraviridae* and *Pithovidirae*, mainly infecting amoebae. NCLDV viruses have remarkably large genomes that can exceed those of free-living archaea and bacteria, with the record being held by *Pandoravirus salinus* and its 2.5 Mbp genome. Moreover, these genomes were found to encode multiple proteins that are universal to cellular organisms but rarely present in other viruses, including key proteins of the translational apparatus – although ribosomes are notably absent from known giant viruses genomes [Schulz et al. 2017]. These virus-encoded cellular genes were initially thought to branch between bacterial and eukaryotic clades, leading to hypotheses about NCLDV possibly representing either a fourth domain of life [Colson et al. 2012] or a degraded variant of some reduced eukaryotic lineage [Claverie and Abergel 2013], but it has since been shown that these genes were instead acquired from eukaryotic hosts. The horizontal acquisition of genetic material between virus and host is commonplace in the virosphere, especially in double-stranded DNA viruses, but even within these NCLDV viruses stand out as particularly frequent vectors of HGT, both as donors and receivers [Irwin et al. 2022]. The breadth and frequency of these exchanges likely contributed to the expansion of NCLDV genomes and viral particles, and suggest remarkable co-evolutionary relationships between NCLDV and eukaryotes. The viral acquisition of eukaryotic genes promotes infection via the development of new host-manipulation strategies, for instance by alleviating the reliance on host machinery that can be shut down by immune responses. Conversely, some important transitions in the evolution of eukaryotes may have been facilitated by genes acquired

from giant viruses, including the development of cell wall structures in algae and cellular aggregation in opisthokonts [Irwin et al. 2022].

Environmental genomics can also help to shed light on our own origins. The identification of a new superphylum of archaea, the Asgard archaea, was a groundbreaking discovery for the research done on eukaryogenesis. Genomes of these archaea, first assembled from metagenomic surveys of deep-sea sediment samples, formed a monophyletic group with eukaryotes in phylogenomic analyses [Spang et al. 2015]. This indicates that the first eukaryotic cell likely evolved from an archaeal ancestor, with features similar to those of extant Asgard archaea. In accordance with this phylogenetic placement, Asgard genomes encode many eukaryotic signature proteins, previously thought to be exclusive to eukaryotes. These include functions related to information processing and regulation, such as a ubiquitin-based system for post-translational protein regulation and modification. Asgards also encode an actin-based cytoskeleton system that is highly analogous to that of eukaryotes. Although these close relatives of eukaryotes are predominantly known from environmental MAGs, two strains of Asgard have recently been successfully isolated in cultures (namely, *Ca. Prometheoarchaeum syntrophicum* and *Ca. Lokiarchaeum ossiferum*) [Imachi et al. 2020, Rodrigues-Oliveira et al. 2023]. Microscopy imaging performed on these isolates revealed a rather intricate cellular architecture, with multiple tentacle-like protrusions budding from the membrane of Asgard archaea cells, supported by the actin filaments of their cytoskeleton (Figure 18C). If eukaryotic cells did evolve from an Asgard-like ancestor, these membrane protrusions may have played a role in the recruitment of the alphaproteobacterium that would eventually become the mitochondrion. Indeed, such extrusions could have mediated trophic interactions by direct cell-cell contact between an Asgard (obligate anaerobe) and an aerobic partner, and led to the progressive engulfment of that bacterium within the Asgard host. This proposed entangle-engulf-endogenise model [Imachi et al. 2020] provides a mechanistic explanation for the revived eocyte hypothesis [Archibald 2008] on the origin of eukaryotes, in opposition to the once-preferred three-domain system.

1.3 – Challenges facing the exploration of microbial dark matter

If the past couple of decades have been particularly prolific in major discoveries from cultivation-independent genomic surveys, this trend seems to have slowed down somewhat in the past few years. Gradually, new conjectures emerged, predicting that we may soon have discovered all the major divisions of life that were unknown to us before cultivation-free sequencing came of age [Castelle and Banfield 2018]. After all, as more and more metagenomic projects are undertaken over time, covering an increasing share of the world's ecosystems, it appears logical that we would eventually exhaust all the possible ecological niches for divergent life forms. In 2020 for instance, a

large meta-survey has been performed on a broad collection of metagenomes, and "found little evidence of new deep-branching lineages representing new phyla" in prokaryotes [Nayfach et al. 2020]. Despite this, metagenomes still harbour vast amounts of predicted genes that lack taxonomical and/or functional annotations to this day [Bernard et al. 2018]. There is therefore still a great potential for biological discovery in the microbial world, certainly at least in sub-phylum taxonomic scales. Many of these elusive genes, for instance, could be specific to bacterial lineages that are incompatible with lab growth requirements, but nonetheless belong to well-documented taxa. Still though, one could argue that if such large portions of metagenomes escape characterisation attempts, then entirely ruling out the possibility of discovering new highly divergent life forms in the future is perhaps a pessimistic perspective. More generally, this vast pool of elusive environmental sequences remains a major blind spot in our understanding of the microbial world, both its inhabitants and their internal and collective processes.

This discrepancy between, on one hand, the massive amount of environmental sequences that remain uncharacterised and, on the other, the apparent ebb of biological discoveries that are made from them, is reflective of a number of challenges that MDM research is currently facing. The first, and perhaps the most fundamental, is the fractal-like structure of the space of microbial unknowns[8]. At every stage of unravelling the MDM of an ecosystem, only a portion of it is effectively addressed: rarely occurring organisms may not be accurately represented in a metagenome, only a fraction of any metagenome can be assembled into MAGs, MAGs often fail to cover the entirety of an organism's actual genome, and most MAGs contain many genes that cannot be assigned to known families. Methodological developments are thus required to improve the reconstruction and annotation of metagenomes and MAGs, in order to increase their descriptive and discovery power. A second challenge resides in the inference of metabolic and ecosystemic functions from MAGs. Genomic information can provide only limited insight into the internal function of a microorganism or its interaction within an ecological community, and complementary methods used in tandem with metagenomics have proven useful to lift these limitations. These include other meta-omics methods (metatranscriptomics, metaproteomics, metabolomics), as well as bioimaging and mass spectrometry techniques [Jiao et al. 2021]. Lastly, improvements to cultivation protocols could lead to new microbial strains being grown and studied in lab conditions, which would provide biologists a much more

[8] This is an analogy to Koonin's "fractality of the prokaryote gene space-time" in *The Logic of Chance* (p. 75) [Koonin 2012]. He describes prokaryotic pangenomes as having a distinct structure consisting of a reduced core, a larger shell and an even larger cloud. He then goes to show that this structure exists at all levels of prokaryotic lineages, from the pangenome of a single multi-strain bacteria to that of prokaryotes as a whole: zooming in or out on prokaryotic evolution does not affect the core-shell-cloud picture of the current lineage's pangenome.

detailed knowledge of their physiology. To achieve this goal, functional information gained from meta-omics studies could be leveraged towards the production of specifically well suited growth strategies. Co-cultivation strategies could also be devised for pairs or groups of microbes that form obligate syntrophic partnerships, similar to the two successful isolations of Asgard archaea mentioned above. The different challenges listed here are naturally interconnected, which highlights the benefit of a plurality of complementary approaches to improve our understanding of the microbial world.

In this chapter, we present the work we conducted in an attempt to address the persistent issue of unravelling the uncharacterised fraction of metagenomes. In particular, we sought to measure the diversity of ancient, highly conserved gene families in natural environments. Our hypothesis was that identifying divergent homologues of genes that evolve notoriously slowly, and are recorded in most (if not all) extant lineages, would result in particularly interesting candidates for potential biological novelty. Highlighting these potential sources of novelty could guide further MDM investigations, especially in the search for new basal groups of microbial lineages. We opted for the OM-RGC (Ocean Microbiome Reference Gene Catalog) metagenome as the target environmental dataset for our analyses, due to its highly comprehensive sampling of marine environments across the planet [Sunagawa et al. 2015]. Starting from a reference dataset of highly conserved gene families, we performed iterative alignment-based searches in the OM-RGC database to gather increasingly distant homologues around the references.

In the following section, we explain the motivations for implementing an iterative process to retrieve remote homologues, as well as the specific method we developed and how it performs on a benchmark dataset. We then return to the real-world analysis mentioned above, and detail the datasets we used to explore the marine microbiome.

2. Iterative detection of distant homologues

2.1 – What motivates a propagative approach?

Sequence alignment algorithms, such as BLAST, perform best above a certain threshold of similarity between sequences. For protein sequences, in particular, the accuracy of aligners remains high above the 30% mark for sequence identity, but drops drastically once the identity dips below 25%, defining a critical range of sequence similarity known as the “twilight zone” for protein alignment [Rost 1999]. Below this range, the homology signal between proteins is increasingly blurred by fortuitous local matches, making the detection of distant homology a persistent challenge in bioinformatics, which calls for a change in strategy. Instead of relying on a single alignment search

with over-relaxed similarity cut-offs to identify remote homologues, some of these strategies attempt to establish this homology indirectly, in a propagative approach. Starting from a "seed" sequence (or group of sequences), for which we want to identify distant homologues, successive search steps are performed, each time updating the search criteria to reflect the newly retrieved homologues.

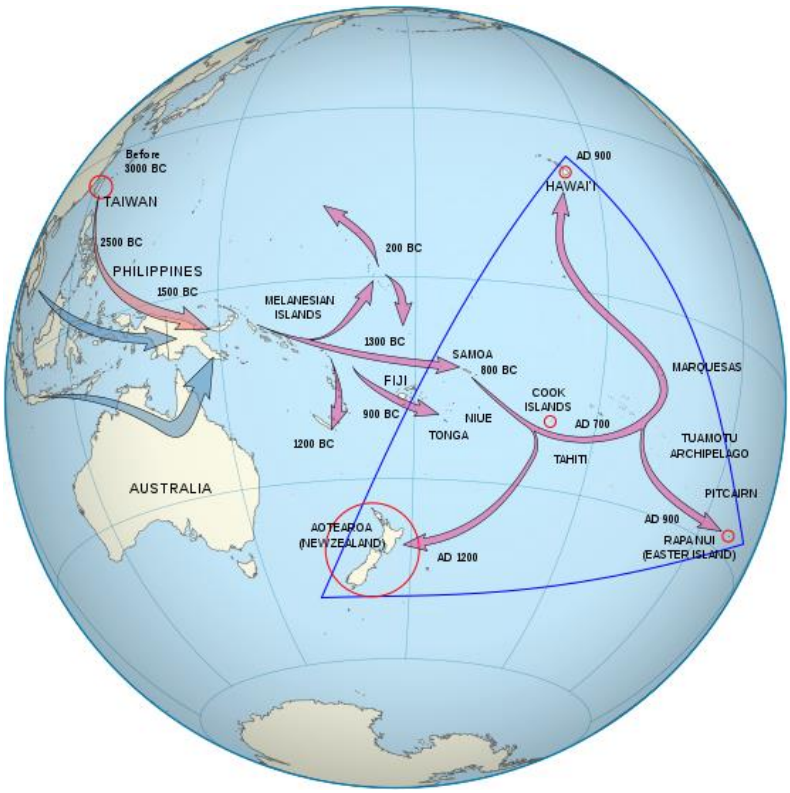


Figure 19: Migration routes of Polynesians from mainland Asia to the Pacific Islands.
From: [Eccles, n.d.].

The underlying assumption is that gaps that are too big to be cleared by a single step (of search) could still be crossed with a series of shorter steps. Consider, for the sake of comparison, the indigenous population of Easter Island, one of the most remote inhabited locations in the world. The first humans to settle on the island could likely not have reached it directly from mainland Eurasia. Instead, it is believed that the Rapa Nui people descend from Polynesians, and that the spread of *Homo sapiens* in the Pacific islands of Oceania followed a chain of shorter migrations, from South East Asia to New Guinea, then to the Islands of Salomon, Vanuatu, Fiji, Polynesia and eventually Easter Island [Hunt and Lipo 2006] (Figure 19). Had Melanesia and Polynesia been sparser, with fewer islands separated by longer distances, *H. sapiens* may have never been able to settle on Easter Island. Similarly, the iterative approach to detecting remote homology presupposes that between two sequences that cannot be aligned directly, there exists a chain of intermediate sequences placed at

regular enough intervals that each can be reached from the previous one, collectively bridging the evolutionary gap between the start and end sequences.

2.2 – Different classes of iterative algorithms

The main algorithm implementing such a strategy to identify remote homologies is PSI-BLAST [Altschul et al. 1997]. From an initial set of query proteins, PSI-BLAST constructs a position-specific scoring matrix (PSSM), which represents a statistical profile of the input sequences. This matrix is then used to identify similar proteins in a database, which are then taken into account to update the weights of the profile, prior to the next search phase. The iterations end when no new sequence has been found by the last search step. This allows the retrieval of more homologues than a regular BLAST search, and reduces the variability of the results based on the initial choice of query proteins. The time performances, however, can make the use of PSI-BLAST cumbersome on larger sequence datasets, as all target sequences are queried at each step of the procedure.

Some methods for sequence comparison rely on other approaches than sequence alignments. Several algorithms based on hidden Markov models (HMMs), in particular, have been developed as alternatives to BLAST. Fundamentally, these methods consist in using HMMs as statistical descriptors to condense the information contained in a multiple sequence alignment (MSA), in a similar way to PSSMs but allowing for a finer level of detail by taking into account insertions and deletions. Sequences can then be scored against a HMM to check their similarity with the underlying MSA, and HMMs can even be compared together by pairwise HMM-HMM alignments [Söding 2005]. These comparisons are usually more sensitive than those based on profiles or direct sequence alignment, simply because Markov models are finer descriptors of sequence data that can take into account more parameters than PSSMs (e.g. by implementing site-specific gap penalties, rather than uniform values). As a result, a single HMM-based search is generally able to identify some homologues beyond the twilight zone of protein similarity. An iterative version of HMM search has been developed for detecting remote homologies, dubbed HHblits, which relies on pairwise HMM alignment [Remmert et al. 2012]. Simply put, the query sequences are abstracted into a query HMM, and sequences in the target database are clustered by similarity, before constructing one target HMM per cluster. The query HMM is then compared to each target HMM, and target HMMs with hits below a certain E-value threshold are retained. A new query HMM is then built using query sequences as well as those of matched target HMMs. This protocol leverages the statistical power of Markov models to produce fast and sensitive searches, but it does come with the requirement of having queries and databases already formatted as HMMs, or paying the computational cost of formatting them *de novo*.

We developed an alternative approach to PSI-BLAST and HMM-based models that also relies on iterative searches to identify distant homologues, this time based on chains of direct sequence alignments. We called our implementation of this method SHIFT, for Sequence Homology Iterative Finding Tool. From an initial set of seed sequences, a target database is queried by a BLAST alignment, and the direct homologues of seed sequences (above given thresholds of E-value, sequence identity and alignment coverage) are retained. This group of first-degree homologues are then used in a second round of search against the remainder of the target dataset, and their homologues (thus second-degree homologues to the initial seeds) are retained. A new cycle of search then begins, and so on, each time using as queries the sequences newly retrieved at the previous step (Figure 20A-B). The resulting set of homologues is therefore layered around the initial queries, like the layers of an onion: the seed sequences occupy the central position, and are direct homologues to the first layer of target sequences, which are themselves homologous to sequences in the second layer, and so on. As with PSI-BLAST, the execution is interrupted once no new match is found at a given search step (HHblits, on the other hand, requires its user to specify *a priori* the number of iterations to perform).

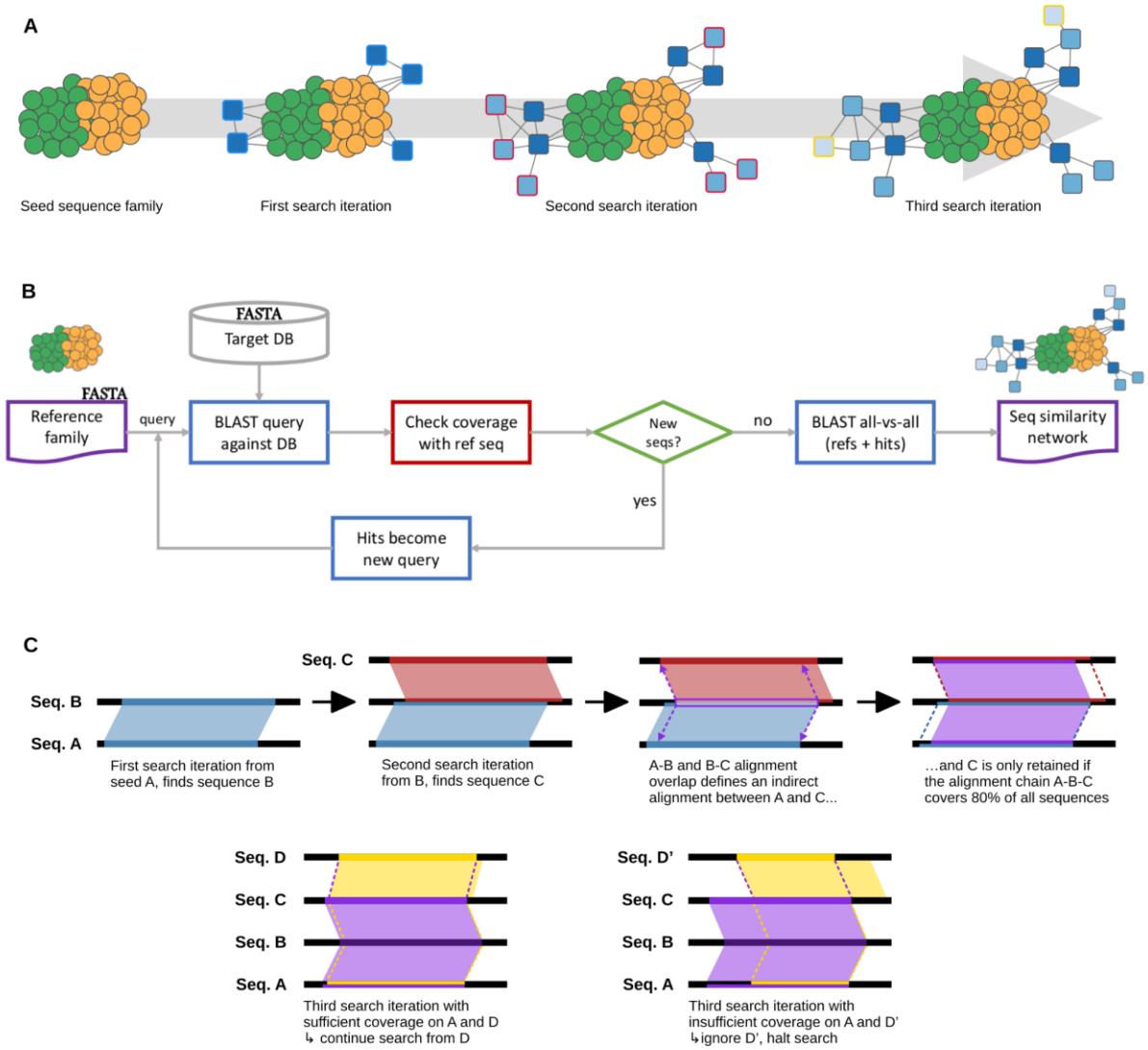


Figure 20: Iterative aggregation of remote homologues with SHIFT.
(**A**) From a set of seed sequences belonging to a given protein family, a first search iteration finds environmental homologues (dark blue) for some of the seeds. A second search iteration then uses these environmental sequences as queries to find more homologues (medium blue, red outline), which are themselves used as queries for a third search iteration finding further environmental homologues (light blue, yellow outline).
(**B**) Schematic representation of the main steps of SHIFT. Taking a FASTA file for a reference family as input, sequences are repeatedly aligned against the target database to find increasingly distant homologues. A sequence similarity network with reference sequences and homologues is then produced.
(**C**) At each iteration of the search, newly found homologues are only retained if their aligned region can be mapped back onto a seed sequence in a way that ensures at least 80% coverage on all sequences along the chain of aligned sequences.

When creating chains of direct pairwise alignments, two opposite kinds of anomalies can occur and challenge the validity of the inferred distant homology between connected sequences. Both of

these pitfalls come from the fact that slightly different regions of a sequence can be aligned to different homologues, even if each resulting alignment covers a sufficient portion of each sequence.

Firstly, consecutive alignments can extend each other to the left or the right seemingly randomly, such that only a small interval is shared by all sequences, with a strong but narrow homology signal (Figure 13A). This can arise, for instance, if all sequences in the alignment chain share a short and frequently occurring protein domain, but are otherwise largely unrelated. Conversely, each aligned region along a chain of alignments can extend the previous one horizontally in the same direction, which produces a sliding or 'staircase' pattern, wherein the aligned parts of the start and end sequences are eventually completely mismatched (Figure 13B). To avoid both of these pitfalls, SHIFT imposes a significant reciprocal coverage between all sequences: after each step that retrieves new potential homologues, all candidates are checked to map back onto a seed sequence, typically with at least 80% of coverage on each. All sequences along a valid alignment chain thus have >80% of their length indirectly aligned to the seed sequence set (Figure 20C).

2.3 – Stringent models are required to preserve valid results

To test the reliability of SHIFT for retrieving distant homologues, we performed a benchmark on a set of 3402 simulated protein families. We generated a collection of toy phylogenies, with a balanced binary topology on 64 leaves, but with asymmetrical branch lengths on opposing sides of the root. Thus, in each phylogeny, branches on one side all had a uniform unit length, whereas some internal branches on the other side were elongated by a multiplicative factor between 1 and 8. To generate our artificial protein families, each tree was assigned a randomly generated protein sequence of 300 amino-acids, which was numerically evolved along the tree, resulting in 64 different sequences, half of which had diverged faster than the rest (in accordance with the elongated internal branches on one side of the tree). 'Slow-evolving' sequences (at the tips of non-elongated branches) within the same toy family shared together an average of 42.7% amino-acid identity. We then conducted SHIFT searches for each family, each time using the 32 slow-evolving sequences to retrieve their 32 fast-evolving homologues among all the sequences of all other protein families (64 sequences from each of 3401 other families, i.e. 217,664 non-homologous sequences to filter through).

The performance of SHIFT in retrieving these homologues was evaluated against two metrics: precision and recall. A high precision indicates a low rate of erroneous positive calls, meaning that reported homologues can be trusted to indeed belong to the seed protein family, whereas a high recall indicates low rates of false negative calls, meaning that most of the existing homologues for that family were successfully retrieved. In general, there is a trade-off between precision and recall for

classification tasks, and increasing one at the expense of the other can be relevant depending on the relative severity of false positive or false negative errors. In our case, a false positive error would be to include as homologue a sequence that is unrelated to seeds. When this happens in SHIFT, an unrelated sequence that is mistakenly retained will then be used to further query the target database for its own homologues, which might eventually result in the inclusion of many more sequences that are not actually homologous to the starting sequences. There is therefore a risk of a snowballing effect for false positives, due to the very nature of iterative searches. This problem is also well-documented in PSI-BLAST, known as "model corruption": once an unrelated sequence is retained, it will skew the weights of the PSSM in a way that enables other unrelated sequences to be matched as well [Schäffer et al. 2001]. For all iterative methods, this model corruption not only undermines the reliability of the search results by creating incorrect outputs, but can also inflate the time and memory costs of their execution with irrelevant and unnecessary computations. Limiting the frequency of false positives is therefore highly desirable, as long as some recalling power is preserved, including for non-trivial cases (i.e. still being able to detect some non-direct homologues). In our simulations, we observed a perfect precision score across all instances, meaning that unrelated sequences were never marked as homologues erroneously. The recall strength of SHIFT, on the other hand, varied based on the rapidity at which fast-evolving sequences diverged from their regular counterparts (quantified by the elongation factor applied to internal branches). When the divergence speed was up to 2.5 times the regular rate, distant homologues were nearly systematically retrieved; then the recall power gradually fell, down to a near-zero for six- and eight-fold branch length increases. The method implemented in SHIFT thus retrieves remote homologies rather conservatively, minimising the risk of model corruption from spurious homology calls, although this comes at the expense of an inability to retrieve many homologues in cases of extreme divergence.

3. Distant homologues of ancestral gene families in the ocean microbiome

3.1 – Highly conserved gene families

The speed of sequence evolution and the phyletic distribution across taxa are both highly variable properties of gene families. In practice, evolutionary biologists are often predominantly interested in families that are well-distributed across the tree of life and show a relative stability in their sequence, as they preserve a greater amount of phylogenetic signal than fast-evolving genes. These genes can be called highly conserved, both in the sense that their sequences accumulate mutations at a slower pace, and that they are rarely lost from genomes altogether. This latter part is

of particular significance, because genomes have a higher plasticity than genes, and the gene content of a genome can evolve much more rapidly than the actual sequence of most genes. In fact, many genes in any given genome can be relatively ancient (typically, as old as their host's domain of life), but only a very narrow set of core genes are reliably found across taxonomic scales with limited exceptions [Wolf et al. 2009]. Many of the gene families among the most conserved are involved in fundamental molecular processes that are shared by all forms of cellular life and likely originated at the time of LUCA or before. These include, for instance, numerous functions related to transcriptional and translational machineries (such as ribosomal RNAs and proteins, frequently used as phylogenetic markers), as well as transporter proteins that mediate the import and export of biochemical products across cellular membranes. In a sense, those core genes can be thought of as the most basic prerequisites for sustaining cellular life, a hypothesis that can be confirmed by gene knockout experiments to observe the consequences of their loss on the survival of their host. As such, it can be reasonably expected of any undiscovered lineage to also rely on these core genes for the same functions as observed in known organisms. In other words, detecting divergent variants of universally conserved genes in the microbial dark matter could potentially be indicative of new groups of currently undescribed organisms, or at least suggest divergent modes of operation in the fundamental processes of cellular life.

We thus sought to explore the environmental diversity of highly conserved gene families, as well as the potential evolutionary implications that divergent variants of these families may have. We assembled an initial dataset of gene families, and from this we extracted a small selection of particularly conserved families to use as seeds for our distant homology search. The initial dataset was constructed by Romain Lannes, former PhD candidate in the lab, by gathering a representative sample of public genomes across all major groups of life, and performing a large all-against-all BLAST search of all genes present in these genomes. The resulting SSN consisted of hundreds of thousands of connected components, and we extracted a subset of 53 clusters corresponding to the most evolutionarily conserved families in the dataset. These included 12 families of ribosomal proteins, as well as a number of families involved in transcription, chromosome stability, amino-acid biosynthesis, and protein translocation.

3.2 – The *Tara Oceans* metagenome

Natural aquatic environments across the globe harbour an unparalleled diversity of microorganisms. Each millilitre of water can contain between 10^4 and 10^6 microbes, which account for up to two thirds of total oceanic biomass [Bar-On, Phillips, and Milo 2018]. As the primary contributors of organic carbon to the marine food web, microorganisms are an essential component of aquatic

ecosystems, sustaining the larger occupants of higher trophic levels. The marine microbiome also plays a vital part in the capture of carbon dioxide and the release of oxygen into the atmosphere, equalling terrestrial forests and wetlands. In the current context of anthropogenic climate degradation, understanding the composition, organisation and function of the marine microbial biosphere is therefore extremely important for the preservation of biodiversity and ecosystems at large.

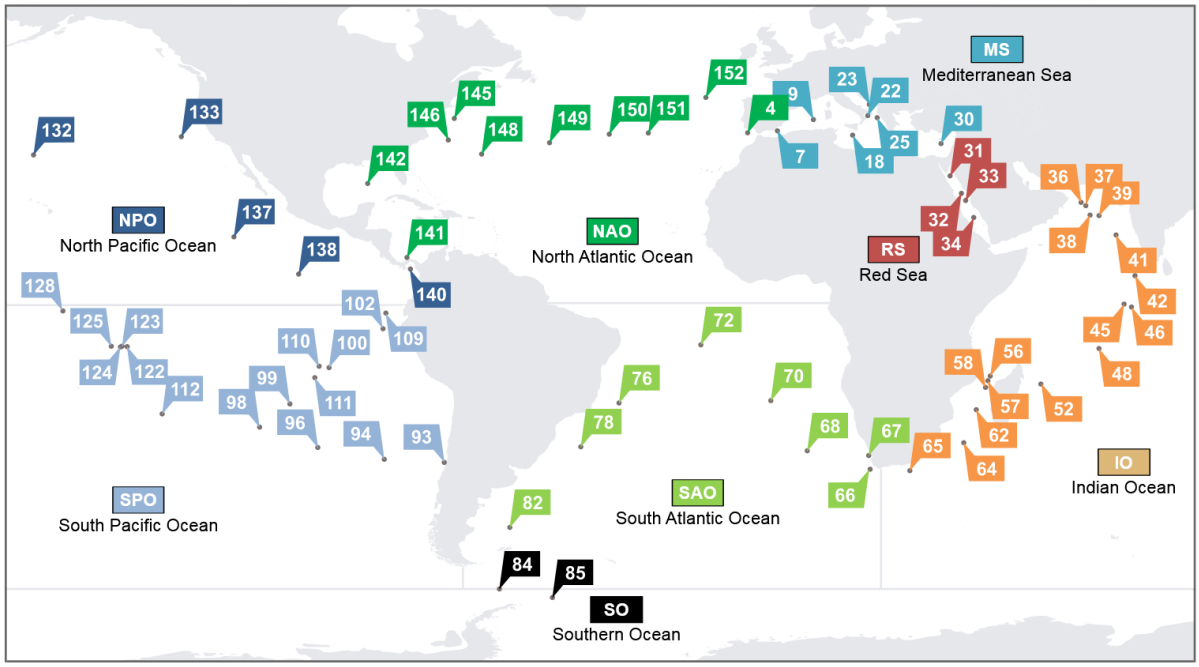


Figure 21: Sampling sites of the *Tara Oceans* expedition.
From: [Sunagawa et al. 2015].

Global-scale efforts to sequence the oceanic microbiome have been carried out in the last few decades. One of these projects, conducted by the *Tara Oceans* foundation and scientific consortium, collected hundreds of water samples in 68 marine locations around the Earth, at varying depths and times of day (Figure 21). Metagenomic sequencing of these samples led to the constitution of the Ocean Microbiome Reference Gene Catalog, a considerable dataset of over 40 million genes from marine microorganisms, and an unprecedented window into the diversity of the microbial world [Sunagawa et al. 2015]. One of the first discoveries made from this genetic record, for instance, highlighted the central role of temperature in shaping microbial community composition in the sunlit layer, more so than geographical distance. Gene rarefaction analyses showed that almost no new genes were detected by the end of the sampling, suggesting that this dataset constitutes a virtually exhaustive picture of the microbial gene space in the ocean, at least in the locations sampled. In this *Tara Oceans* metagenome, at time of initial publication, 45% of sequences lacked a taxonomical annotation at or below the Domain level, and 43% were unassigned to a functional orthologous group.

A large part of the genetic diversity in the global ocean is therefore still undercharacterised, and unravelling this mysterious fraction could thus be an abundant source for discoveries of exciting new biology.

In the following article, we used SHIFT to mine the OM-RGC metagenome in order to measure the environmental diversity of our selected protein families. This resulted in a seven-fold increase of their sequence content once environmental homologues are identified. We found that a fifth of those sequences diverged more from any gene in the entire diversity of well-characterised organisms than bacterial and archaeal homologues diverged on average in our reference dataset. These highly divergent variants were present in comparable proportions in all sampling sites, suggesting that MDM still persists in many marine environments. In particular, we investigated the significance of divergent environmental homologues in three key protein families. In DNA polymerase clamp loaders, we found groups of divergent variants spread throughout the phylogenetic diversity of the family, suggesting that a diversity of uncultivated marine organisms replicate DNA using various unusual proteic machineries. We also detected a new variant of SMC proteins, responsible for chromosome conformation and stability in all Domains of life [Hirano 2002, Cobbe and Heck 2004], with unusual structure and domain architecture in Actinobacteria. Specifically, this divergent SMC clade has lost the *hinge* domain responsible for interfacing with DNA to initiate DNA binding [Gruber et al. 2006], which indicates that these proteins may either perform a different function than usual SMC, or use a different mechanism to achieve this function. These *hinge*-less SMC could be encoded by known members of Actinobacteria (which would be the first description of a duplication of SMC in prokaryotes), or by a novel lineage within this phylum with a unique SMC variant. Lastly, we identified clusters of divergent recombinases that were enriched in super-small cell size fractions, typical of CPR and DPANN but phylogenetically distinct from recombinases of those phyla. These recombinases might belong to unknown bacteriophages, or perhaps to unknown groups of ultra-small organisms, and in any case highlight this size fraction as a particular source of potential biological novelty. Together, these results support the notion that significant gaps remain in our understanding of microbial life, and provide examples of possible discoveries to be made regarding new types of biology in the ongoing unravelling of microbial dark matter.

New groups of highly divergent proteins in families as old as cellular life with important biological functions in the ocean

Duncan Sussfeld[1,2,*], Romain Lannes[1], Eduardo Corel[1], Guillaume Bernard[1], Pierre Martin[1], Eric Bapteste[1], Eric Pelletier[2,3], Philippe Lopez[1]

[1]Institut de Systématique, Evolution, Biodiversité (ISYEB), Sorbonne Université, CNRS, Museum National d'Histoire Naturelle, EPHE, Université des Antilles, Paris, France

[2]Génomique Métabolique, Genoscope, Institut François-Jacob, CEA, CNRS, Université d'Evry, Université Paris-Saclay, 91000 Evry, France.

[3]Research Federation for the Study of Global Ocean Systems Ecology and Evolution, FR2022/Tara Oceans GOSEE, 75016 Paris, France.

***Corresponding author**: Duncan Sussfeld, duncan.sussfeld@gmail.com

Abstract

Background: Metagenomics has considerably broadened our knowledge of microbial diversity, unravelling fascinating adaptations and characterising multiple novel major taxonomic groups, e.g. CPR bacteria, DPANN and Asgard archaea, and novel viruses. Such findings profoundly reshaped the structure of the known tree of life and emphasised the central role of investigating uncultured organisms. However, despite significant progresses, a large portion of proteins predicted from metagenomes remain today unannotated, both taxonomically and functionally, across many biomes and in particular in oceanic waters, including at relatively lenient clustering thresholds.

Results: Here, we used an iterative, network-based approach for remote homology detection, to probe a dataset of 40 million ORFs predicted in marine environments. We assessed the environmental diversity of 53 gene families as old as cellular life, broadly distributed across the Tree of Life. About half of them harboured clusters of environmental homologues that diverged significantly from the known diversity of published complete genomes, with representatives distributed across all the oceans. In particular, we report the detection of environmental clades with new structural variants of essential genes (SMC), divergent polymerase subunits forming deep-branching clades in the polymerase tree, and variant DNA recombinases of unknown origin in the ultra-small size fraction.

Conclusions: These results indicate that significant environmental diversity may yet be unravelled even in strongly conserved gene families. Protein sequence similarity network approaches, in particular, appear well-suited to highlight potential sources of biological novelty and make better sense of microbial dark matter across taxonomical scales.

Keywords: Microbial dark matter, Sequence similarity networks, Distant homology, Microbiome

Background

Over the last decades, novel sequencing methods have allowed microbiologists to appreciate the ubiquity and abundance of uncultured organisms [1–5], and access microorganisms' genomes beyond the isolation-cultivation dogma issued from the Koch principles [6] that underpinned microbiological studies for decades. Metagenomic studies [7] have led to an unprecedented broadening of our knowledge of microbial diversity [8], from the unravelling of microbial adaptations and interactions in numerous environments [9–12] to the characterisation of multiple novel major taxonomic groups [13–17] – most notably CPR bacteria [13, 18, 19], DPANN archaea [18, 20, 21] and Asgard archaea [22–24], profoundly reshaping the structure of the tree of life. Large groups of novel viruses [25–27] and mobile elements [28] have also been unearthed. Together, these major discoveries emphasise the central role of investigating yet uncultured organisms, believed to constitute the majority of overall microbial lineages [3, 29], in addressing many fundamental questions of biology and evolutionary biology.

Over time, as cultivation-independent sequencing efforts are carried out in an increasing range of ecosystems, discovery events of novel branches near the base of the tree of life are predicted to become less frequent [8, 17]. In accordance with this perspective, an extensive study of over 50,000 MAGs, assembled from a vast ensemble of metagenomes and including 12,556 novel candidate species-level OTUs, found no reliable evidence of novel prokaryote phylum content [30]. It may therefore seem that whatever biodiversity remains to be discovered should yield few more "major unknowns".

However, contrasting with these observations, it still persists that across most biomes, large portions of environmental metagenomes remain taxonomically and functionally unannotated, even at relatively permissive clustering thresholds [31]. This vast pool of uncharacterised sequences remains a significant blind spot in our grasp of the extant biological diversity on Earth. Some may yet belong to genomes of unknown organisms that have so far escaped detection efforts, for instance due to

accelerated evolution rates or an ancestral divergence from known organisms. Novel genes of well-characterised organisms with “open” pangenomes, divergent paralogues of known genes, and unusual mobile elements may also be expected to contribute to this “microbial dark matter” [4]. In any case, the persistence of those biological unknowns highlights the need for novel approaches complementing the current techniques to mine metagenomes for highly divergent groups.

Various network-based approaches [32], in particular, have been developed to address these concerns. Co-occurrence networks, for instance, can help assessing ecological roles of unknown taxa [33]. Sequence similarity networks, wherein pairs of primary sequences are connected according to set similarity criteria, can also be employed to compare sequences from cultured and uncultured organisms [34, 35]. In 2012, Lynch et al. used sequence similarity networks to identify several candidate new lineages from environmental 16S rRNA [36]. In 2015, Lopez et al. designed a network-based exploratory analysis to probe metagenomes for distant homologues of well-distributed gene families [37]. 86 clusters of genes broadly distributed across Domains of life were used as seeds for a two-step BLAST search inside a metagenome collection. Seed sequences were then gathered in sequence similarity networks together with their direct and indirect environmental homologues, and environmental sequences gathered in the second alignment step were more divergent from their cultured relatives than those gathered in the first round. The authors found several hundred groups of highly divergent environmental variants, some of them potentially compatible with novel major divisions of life. Consequently, (i) iterative explorations of environmental datasets may allow the retrieval of increasingly divergent variants (Fig. 1A), and (ii) network-based methods may be well-suited to handle this type of data, by integrating sequences with various levels of divergence within homologous gene families. Sequence similarity networks have also been used recently to assess how the deep-learning breakthrough in protein structure prediction may be leveraged to shed light into “functionally dark” regions of the natural protein space [38].

In this work, we conducted an exploratory search of ocean metagenomic data to identify potential sources of novel diversity in highly conserved, near-universal gene families. Our search mined the environmental diversity of the Ocean Microbial Reference Gene Catalog (OM-RGC) dataset [39]. This extensive, non-redundant record contains sequences for over 40 million bacterial and archaeal genes, predicted from metagenomic sequencing of a large variety of marine environments across the world. At the time of initial publication, around 45% of these sequences lacked taxonomical annotation at or below the Domain level, and 43% lacked functional annotation to an eggNOG orthologous group (OG), highlighting the existence of a vast, undescribed diversity in the global oceanic microbiome, as well as the necessity of additional efforts to improve its characterisation. To perform this search, we further developed the iterative explorative strategy of environmental datasets initiated by Lopez et al. [37], by allowing distant homologue search iterations to continue indefinitely until convergence. Specifically, we focussed our search on ancestral gene families that showed particular conservation across their taxonomic distribution in the face of evolution. Retrieving highly divergent variants in such families could indeed carry an increased biological significance, given their stability in primary sequence for many reference genomes, and potentially guide future searches for novel putative taxonomical groups or biological functions involving these nearly universal gene families. We thus used a custom dataset of 53 ancient, conserved gene families with key biological functions to initiate our iterative probing of OM-RGC. We identified highly divergent variants of multiple gene families, uncovering new putative structural and sequence variants of biologically essential proteins across taxonomical scales.

Results and Discussion

Oceanic metagenomes harbour distant homologues of highly conserved protein families

We developed an iterative mining procedure to accumulate highly divergent environmental variants for families of genes or proteins of interest. From an initial set of nearly ten million protein sequences gathered from prokaryotic, eukaryotic, viral and plasmidic complete genomes (Table SI-1), we selected a set of 53 protein clusters, highly conserved and at least as old as cellular life. Most clusters corresponded to single protein families, though a few of them comprised proteins from two or more closely related families (we hereafter refer to those clusters as families for simplicity, and will make multiplicity cases explicit when discussing such clusters specifically). These families spanned a total of 125,774 sequences and included 12 families of ribosomal proteins (Table SI-2). On average, bacterial sequences in these families had 34.9% amino-acid identity to their closest archaeal homologue (and vice-versa), roughly illustrating the level of divergence to expect between sequences from different Domains of life.

Each selected family was used as the seed for a deep homologue-mining procedure in the OM-RGC dataset [39]. This iterative search aimed at aggregating around each seed family the diversity of its environmental homologues, including variants too divergent to produce a significant direct alignment to any seed sequence. For each family, direct oceanic homologues of seed sequences were identified in a first round of search. The OM-RGC dataset was then further queried for homologues of those homologues, and so forth until the procedure converged to find no additional environmental homologues (See Fig. 1A-E and Methods for details).

We tested the performance of our method by conducting homology searches on a simulated dataset, and found that our protocol was particularly resistant to false-positive homology calls. More specifically, we sought to evaluate (i) how reliably our iterative procedure successfully retrieved distant homologues of seed sequences, and (ii) whether this retrieval was prone to false-positive

calls, where sequences would be attained from seeds that did not share a homologous origin. To that end, we generated a collection of phylogenetic trees, based on a common balanced binary tree structure, where branches along the path from the root to another given node (internal or terminal) were elongated to represent various levels of divergence. Each tree was then assigned a randomly generated amino-acid sequence, which was evolved numerically along its branches, resulting in some slow- and some fast-evolving terminal sequences. Slow-evolving sequences within the same "family" shared an average of 42.7% sequence identity. Finally, the slow-evolving subsets were each used as seeds for iterative homology searches to retrieve their own fast-evolving homologues amongst all sequences generated from all phylogenies. Across the 3402 test cases that were performed in total, we detected no instance of false-positive homology hit, i.e. homology searches only ever retrieved sequences genuinely related to the seeds. In cases where fast-evolving sequences diverged up to 2.5 times faster than their slow counterparts, the search procedure was nearly systematically able to retrieve all divergent sequences (Fig. SI-1). When the evolution rate difference was four-fold, about half of the test instances successfully retrieved all divergent homologues. Finally, above a six-fold increase, seed sequences were largely unable to retrieve any divergent sequence at all. These results on simulated data show that the procedure we developed to identify remote homologies aggregates new sequences in an efficient but conservative manner that resists spurious homology calls, although the higher complexity of real-world biological sequence data may be expected to yield aberrant results on occasion.

Our iterative metagenome mining procedure expanded the selected 53 seed families by a total of 826,717 environmental sequences from OM-RGC (Fig. SI-2). All seed families had their own set of environmental homologues, requiring an average of 7 rounds of iterative search before exhaustion. Despite metagenomic sequencing sometimes yielding shorter gene sequences than what is anticipated from genomes in culture, sequences retrieved from OM-RGC were only marginally shorter than their reference counterparts (Pearson r=0.96, p-value 3.5×10^{-30}), further confirming that their divergence was not related to a systematic bias associated with sequence size.

OM-RGC homologues of the 53 selected seed families were then compared against proteins from the NCBI non-redundant (*nr*) database to find their closest relative amongst all published sequences with taxonomically resolved annotations (Fig. 1F; Supplementary Text SI-1). Only 6.7% of all retrieved environmental sequences were >90% similar to their closest characterised relative, implying that a large majority of environmental proteins cannot be accurately represented by genomes captured by current cultivation or isolation techniques. Furthermore, 20.5% of environmental variants had less than 34.9% similarity with their closest *nr* relative, i.e. they diverged more from any proteins of well-characterised organisms than bacterial and archaeal homologues diverged from one another on average in the reference dataset. Environmental homologues of ribosomal protein families had generally higher similarity to their closest characterised relative than non-ribosomal environmental sequences (one-sided Kolmogorov-Smirnov test, p-value $<1.6\times10^{-22}$; Fig. SI-3), possibly owing to their reputedly high evolutionary conservation. Still, even ribosomal protein families included very divergent oceanic variants (Fig. SI-3). Moreover, all sampled oceanic sites revealed similar proportions (but uneven absolute numbers) of divergent and highly divergent prokaryotic sequences (Fig. SI-2). Any location in the global ocean could therefore be a prolific reserve of new microbial gene variants, including temperate surface-layer habitats. Some of the retrieved environmental sequences show levels of divergence to the known diversity that are comparable with the difference between archaeal and bacterial homologues. These variants could potentially belong to uncharacterised lineages that branched away from well-known taxa long ago, although alternative hypotheses can be offered: divergent environmental homologues could, for instance, be distant paralogues of seed sequences, that evolved faster than their known counterparts due to relaxed selective pressure after duplication, and appear environmentally conserved but not described in cultured organisms.

Highly divergent clusters of environmental variants expand the diversity of multiple universal protein families

Seed sequences and their (direct and indirect) oceanic homologues were then gathered in family-specific sequence similarity networks (SSNs). Similar sequences in these networks are expected to gather in coherent, well-connected groups, thus reflecting the structure of protein families in the network topology. Sequences within each SSN were therefore partitioned into network communities using Louvain clustering [40] (Fig. 1G). This higher-level view of network structures allows an easier assessment of the environmental diversity, including identifying potential sources of biological novelty in these protein families. In particular, clusters consisting exclusively or predominantly of environmental sequences (>90% of environmental sequences), with little similarity to published sequence records (<40% sequence identity to any non-environmental sequence in the nr database), and containing enough proteins to be unlikely the result of sequencing inaccuracies, are intuitively the most likely to correspond to genuinely novel groups of environmental homologues.

691 clusters of sequences were inferred in total across the 53 SSNs, of which we retained 80 clusters of proteins fitting the above criteria for significant novelty potential. These 80 clusters of highly divergent sequences were distributed across 25 ancient, conserved protein families. Remarkably, no cluster with such a high level of divergence was found in networks of ribosomal proteins, possibly due to a superior level of conservation or a higher coverage of their diversity in public sequence databases. Still, the fact that clusters of divergent environmental homologues were identified in nearly half of our selected protein families suggests that numerous key biological processes are carried out by a currently underestimated diversity of protein primary structures. In other words, the "functional dark matter" of proteins likely consists of both unknown functions and unknown actors of known functions [31, 41] .

To assess how these groups of divergent sequences may relate to their reference counterparts, we reconstructed phylogenetic trees regrouping seed and environmental sequences from each of the 80

selected highly divergent clusters. This selection exposed an additional phylogenetic diversity in conserved protein families when environmental contributions are considered. In particular, in some families, sequences representative of certain divergent network clusters branched between or beside the main groups of archaeal and bacterial sequences. Such phylogenetic placements indicate substantial potential for novelty in the sequence space of those protein families. We detail findings of particular interest for three families in the following subsections.

High environmental diversity in oceanic DNA polymerase clamp loaders

One of the selected seed families in our study consisted of several AAA+ ATPases [42], mostly involved in clamp-loading systems for DNA replication. In the environmental diversity of this family, we identified large contingents of highly divergent variants across the phylogeny of the family.

In a mechanism conserved across all cellular life forms, DNA polymerases process and replicate DNA by binding onto circular clamps that encircle and slide along the template DNA strand. Sliding clamps are embedded onto DNA by a pentameric clamp-loading system, which exhibits a universally conserved structure in archaea, bacteria and eukaryotes despite differences in subunit composition [43]. All clamp loaders consist of one "large" subunit (δ in bacteria, RfcL in archaea, Rfc1 in eukaryotes) complemented by four "small" subunits: three γ and one δ' subunits in bacteria (also respectively called DnaX and HolB), four RfcS subunits in archaea, one each of Rfc 2-5 subunits in eukaryotes. All subunits are homologous to one another within and across all three Domains of life [44–47].

Our seed family consisted of sequences for the clamp loader "small" subunits (CLSSUs) described above (i.e. bacterial DnaX and HolB, archaeal RfcS, and eukaryotic Rfc 2-5), as well as sequences for the bacterial replication-associated recombination protein RarA. This protein, present in bacteria and eukaryotes but not in archaea [48], is involved in homologous recombination and DNA repair, both in the context of DNA replication and outside [49]. The RarA protein sequence is highly conserved and

also substantially homologous to DnaX, and as such was grouped alongside it in the construction of our seed families.

The iterative retrieval of environmental homologues for this protein family resulted in a nearly five-fold increase of its sequence content (Table SI-2). In particular, the resulting SSN harboured 10 new clusters of highly divergent environmental homologues (Fig. SI-4). Owing to their high divergence in primary sequence, not all clusters translated to perfectly monophyletic groups in the phylogeny we produced (Fig. 2), though they still generally maintained some level of coherence. Amongst the ten environmental clusters, one had its representative sequences branch within reference archaeal and eukaryotic Rfc sequences (cluster 26), and another translated to a new clade within reference HolB/DnaX bacterial sequences (cluster 23). Additionally, one environmental cluster branched next to bacterial RarA sequences (cluster 27), and its sequences were annotated as belonging to the B subunit of the Holliday junction resolving complex RuvABC, already shown to cluster near clamp-loading proteins in sequence networks [50]. Finally, sequences from seven divergent clusters resulted in groups outside the bacterial and archaeal/eukaryotic seed sequence clans [51] in the phylogeny (clusters 2, 14, 15, 16, 19, 24, 25). Eggnog annotations for these sequences mapped them predominantly to HolB (COG0470), though it should be noted that one particular cluster contained 96% of functionally unassigned sequences (cluster 24).

Protein structures were predicted for representatives of seed and divergent environmental CLSSUs using ColabFold [52, 53], and gathered in a dendrogram depicting their similarities (Fig. 3). Most seed proteins used for this comparison showed similar structures, although HolB, DnaX, RarA and archaeal/eukaryotic Rfc still formed distinct groups in the structure dendrogram. Structures inferred from environmental variants followed a pattern similar to the sequence phylogeny, with representatives from clusters 2, 15 and 23 branching near HolB references, and most other clusters translating to structures sitting outside of the main reference groups. In other words, the

environmental HolB variants that we identified on the basis of primary sequence divergence also exhibited a divergence in 3D structure consistent with their phylogenetic placements.

Oceanic homologues of CLSSUs therefore diverge from their known counterparts in primary sequence, and exhibit tertiary structures comparable, but not identical, to canonical CLSSU structures. Such structural differences and sequence-based phylogenetic placements for these highly divergent environmental CLSSU homologues could reflect the existence of undetected divergent paralogues in these gene families, which would raise interesting questions about their possible contribution in such a conserved subprocess of DNA replication. These could also hypothetically be indicative of some unknown microbial lineage(s), though much more conclusive data would be required before firmly asserting this. In any case, these results hint at a diversity of uncultivated marine organisms replicating DNA using various unusual proteic machineries, possibly resulting in unusual replication mechanisms operating in the ocean.

Novel abundant clade of SMC proteins with unusual structure in Actinobacteria

Another remarkable seed family consisted of SMC (structural maintenance of chromosomes) proteins, and we identified a small but abundant group of environmental SMC variants with strikingly singular structures within Actinobacteria.

SMC proteins are present in all Domains of life and act (as part of the SMC complex) as regulators of high-order chromosome organisation [54]. Eukaryotic genomes encode six paralogous SMC proteins (SMC1-6), due to a sequence of duplications around the time of the last eukaryotic common ancestor. Indeed, a single copy of the *smc* gene is present in nearly all archaea and bacteria, with a few exceptions. In some γ-proteobacteria a different proteic complex, MukBEF, is responsible for these functions instead [55]. Bacteria from various phyla can also harbour another complex, MksBEF, alongside their SMC or MukBEF machinery [56]. MksBEF is believed to be evolutionarily related to MukBEF, and both are structurally analogous to the SMC complex, but primary sequence comparisons have ruled this structural similarity as convergent rather than due to distant homology

[54]. SMC complexes are also notably absent from Crenoarchaeota, resulting in distinctive chromosomal dynamics and cell cycle logics [57, 58].

A typical SMC protein consists of five domains: an N-terminal domain containing a Walker A motif; a first helical chain of roughly 300 amino-acids; a central “hinge” domain; a second α-helix of comparable length to the first; a C-terminal domain containing a Walker B motif [59, 60]. This linear structure self-folds by linking the N- and C-terminal motifs into an ATPase “head”, with the two α-helix domains forming an antiparallel coiled-coil between this head and the hinge domain. This hinge then serves as a dimerisation site for a second SMC monomer, with accessory proteins binding to the ATPase heads to complete the ring-shaped SMC complex [54]. The hinge region of the SMC complex subsequently plays the essential role of mediating DNA binding, and allows the loading of SMC rings onto chromosomes [61, 62].

From seed sequences in this family, we retrieved a rather limited amount of environmental homologues (0.97 environmental homologue per seed sequence in this family, compared to a median value of 2.6 across all families, see Table SI-2), but one small cluster of distant environmental homologues was still identified (cluster 9 in Fig. SI-5). In the phylogeny produced from seed SMC sequences and oceanic variants from this cluster (Fig. 4), environmental sequences formed a monophyletic clade branching close to the base of seed actinobacterial sequences. These divergent environmental sequences were functionally annotated as SMC proteins (COG1196), and were strikingly abundant in the sequencing data, nearly seven times more so than other OM-RGC SMC homologues. Moreover, this novel oceanic clade harbours SMC-related proteins that are critically different in structure from canonical SMC proteins (Fig. 5A; average TM-score between two proteins in the divergent cluster: 0.828; average TM-score between a protein in the divergent cluster and a reference SMC protein: 0.440). Namely, these oceanic variants lack the hinge domain which is normally essential to SMC assembly and function (Fig. 5B). As such, they may be considered more similar to bacterial SbcC and archaeal and eukaryotic Rad50 proteins, thought to be distant

evolutionary relatives of SMC [54]. Indeed, proteins from this ancestral family also consist of an SMC-like head and an antiparallel coiled-coil with no hinge domain, dimerising instead through a zink-hook structure induced by a CXXC motif [63]. However, FoldSeek structural comparisons clearly discriminate between reference Rad50/SbcC proteins on one side, and SMC proteins (reference or divergent OM-RGC variants) on the other (Fig. SI-6). The zink-hook CXXC motif conserved in Rad50/SbcC is also absent from our environmental cluster sequences, confirming them as divergent variants within the SMC diversity rather than beside it.

Several evolutionary scenarios could explain this new bacterial cluster of “hinge-less” SMC. Firstly, it could be indicative of some paralogue of SMC existing in Actinobacteria. This would then be, to the best of our knowledge, the first description of SMC duplication in prokaryotes [64]. Alternatively, this divergent cluster could indicate the existence of an unknown lineage, supposedly branching within Actinobacteria, where the SMC hinge domain would have been lost. In any case, the substantial divergence of these environmental sequences to any gene published from a well-characterised organism, together with the loss of the essential hinge domain and their remarkably high abundance in the sampling data, suggests that we identified a new kind of biology within the SMC family. By the absence of their expected interaction site with DNA, one would speculate that these hinge-less SMC-related proteins must either perform a different function than known SMC or bind DNA through different mechanisms. The broad distribution of hinge-less SMC variants across the oceans, their monophyly and their relative abundance in the ocean microbiome suggest that they play an important, underappreciated function in this oceanic clade.

Divergent recombinases from potentially novel groups in sub-micrometre size fractions

In a third family, consisting of RecA/RadA DNA recombinases [65], we identified other possible sources of novel diversity, including within ultra-small cell size fractions.

During the course of DNA replication, accidental double-strand breaks (DSBs) in the DNA molecule can have detrimental effects on genome stability and cell viability [66]. Recombinase proteins in the

RecA/RadA family are central to homologous recombinational repair, a key replicative stress-reduction pathway that can correct DSBs as well as other types of DNA damage. This family contains the extensively studied bacterial recombinase RecA (also present in eukaryotic organelles) as well as its archaeal and eukaryotic homologues, respectively RadA and Rad51 [65, 67–69].

Identifying distant environmental homologues of this seed family increased its total size five-fold (Table SI-2). Amongst this added diversity, four clusters of environmental sequences were retained as highly divergent, totalling 1700 sequences. A phylogenetic tree was produced from seed sequences as well as representative sequences for these divergent environmental clusters (Fig. 6). In this phylogeny, sequences from a first cluster branched near the root of archaeal seed sequences, and was functionally categorised as RadA (COG1066) in accordance with this placement (cluster 20 in Fig. SI-7). A second cluster of divergent environmental sequences (cluster 12) branched within the environmental ultra-small cluster in Bacteria. Interestingly, this cluster was predominantly annotated as ArlH (COG2874), an archaeal protein involved in the biogenesis of the archaellum, a cellular motility structure analogous to bacterial flagella [70]. Structure and sequence similarities between ArlH and bacterial RecA have previously been described [71] but, to the best of our knowledge, no evolutionary hypothesis has yet been put forth to explain this surprising homology. Finally, one cluster of distant environmental RecA homologues (COG0468) branched within bacterial sequences (cluster 5), and a final cluster, also annotated as RecA, saw its representative sequences sit between the archaeal and bacterial references (cluster 19). Interestingly, both of these clusters were composed of >50% of sequences from the “ultra-small” size fraction of cells with diameters <0.2 µm. Such cellular sizes are akin to those of CPR bacteria and DPANN archaea [72]; however, seed sequences from these ultra-small superphyla branch clearly within the clans of their respective Domains of life. Additionally, environmental sequences from these ultra-small clusters bore no remarkable similarity to viral sequences recorded in the NCBI Virus sequence database (accessed in February 2023) and just 11 of them (out of 1700) matched to a single oceanic virus from the GVMAG database [73].

351 The origin of these divergent RecA proteins therefore remains open: they could, for instance, belong
352 to unknown bacteriophages or mobile elements populating the global ocean. They might also result
353 from a duplication and divergence of the *recA* gene in CPR bacteria, although they should in that case
354 be expected to appear in published genomes for members of this lineage. They may yet genuinely
355 belong to new uncharacterised, deep-branching cellular lineages of sub-micrometre cell size, though
356 significantly more evidence would again be required to support this hypothesis. Nevertheless, our
357 finding of new very deep-branching groups related to RecA is consistent with the description of new
358 basal groups of metagenomic RecA sequences formerly proposed [74], and highlights the ultra-small
359 size fraction as a notable source of novelty in this essential protein family. Uncovering divergent
360 forms of RadA in metagenomes is also exciting, because even some forms of RadA previously
361 described as inactivated have been demonstrated to be functionally relevant for their host cells, and
362 putatively attached to an alternative mechanism of replication initiation or in the regulation of origin
363 recognition [75]. Moreover, sequence divergence, typically in the non conserved region of intein-
364 containing RadA, may be functional, as it may affect the temperature-induced splicing of the intein of
365 RadA, a phenotype that has been described in *Thermococcus sibericus* [76].

Conclusion

The prevalence of biological unknowns in environmental metagenomes remains, to this day, vast; vast indeed to the extent that “known unknowns” and “unknown unknowns” constitute a relevant distinction to address genes, organisms, processes and interactions at play in the uncultured microbial world. With our network-based, multi-marker iterative approach, we sought to understand the structure of environmental genetic variation for a range of ancient, conserved gene families with functions essential to cellular life. We found that environmental variants for those gene families could exist in marine microbiomes with considerable divergence to the known diversity. Moreover, these highly divergent sequences organised in (sometimes vast) cohesive groups of homology, supposedly harboured in (sometimes vast) groups of related genomes, as illustrated by the oceanic variants of DNA polymerase clamp loaders, hinge-less SMCs, and deep-branching divergent RecA/RadA variants from the ultra-small size fraction.

A common issue surrounding metagenomic data is to know whether predicted genes and proteins actually exist in the sampled environment or result from aberrations in the assembly process. To avoid this pitfall, we purposefully limited our analyses to larger clusters of (similar but non-identical) sequences, from the already non-redundant OM-RGC dataset. Furthermore, the nature of our retrieval process imposes at least 80% of the length of any retrieved sequence to map back to at least 80% of a seed sequence (Fig. 1B-C). As such, recombined proteins mixing sequence fragments from several protein families are unlikely to be matched to our “canonical” seed families if exogenous regions cover more than 20% of their length. Lastly, the benchmarks we performed on simulated protein families show that sequences unrelated to the search seeds are seldom retrieved by erroneous homology calls. For these reasons, we believe that the groups of oceanic variants we discussed correspond to genuine environmental homologues of reference sequences, rather than assembly artifacts, protein recombinants, or non-homologous proteins from unrelated families.

390 Still, various competing scenarios of evolution and diversification could explain highly divergent
391 homologues such as those we detected. We list some of them here, understanding that a single
392 “one-size-fits-all” explanation to all the divergent groups we identified is highly unlikely.

393 A first hypothesis could be that an environmental cluster represents deep paralogues resulting from
394 an ancestral duplication in the gene family. Though not impossible, this hypothesis does require an
395 explanation as to why these paralogues do not appear more broadly in the wide range of public
396 genomes currently available, save for some unlikely event of widespread parallel gene loss across the
397 tree of life. Alternatively, these divergent sequences could have been spawned by more recent gene
398 duplications at narrower taxonomic scales, after which they would have diverged rapidly from their
399 “original” copy. This is entirely possible, predominantly for clusters clearly branching inside the
400 phylogenetic clade of established taxa. The divergent SMC proteins we identified within
401 Actinobacteria are perhaps an example of this (this would then be the first description of an SMC
402 duplication in prokaryotes), though once again it would leave unexplained why most actinobacterial
403 genomes do not seem to carry these “hinge-less” variants. Cases like this are also interesting from a
404 functional standpoint, as the rapid divergence in primary sequence following gene duplication raises
405 questions of neo- or subfunctionalisation for the novel paralogue.

406 Divergent homologues of highly conserved, ancestral families could also stem from uncharacterised
407 genomes bearing these variants. Marine viruses, or other mobile elements, could be carrying such
408 variants, especially those identified in smaller organism size fractions, such as the divergent forms of
409 recombinase A we reported. It is possible that the divergence of these homologues could then point
410 to radical gene changes, driven by specific selective pressures associated with non-cellular organisms.
411 Conversely, unknown cellular lineages that diverged recently (e.g. from known genera or families)
412 could also harbour unusual gene variants. In the functions we specifically targeted, strong constraints
413 on sequence evolution are expected, meaning that drastic changes in intracellular processes or
414 external selective pressure may have prompted those high levels of sequence divergence over short

evolutionary timeframes. Lastly, the levels of divergence observed from some environmental groups could be compatible with novel major taxonomic groups that diverged from the established diversity some hundreds of millions, or even billions of years ago. This last hypothesis would, of course, require a lot more evidence to substantiate such a claim, and full genomes with high levels of divergence across their length would have to be produced and analysed thoroughly. Still, however remote, the possibility for new basal branches in the tree of life should not be fully discarded in the absence of conclusive evidence favouring other hypotheses.

All in all, the detection of divergent variants in key protein families, that have likely existed since cellular life began, supports the notion that major gaps remain in our knowledge of biological diversity, and that various forms of exciting new biology may be expected from unravelling this microbial world. To that end, future methodological extensions that rely less on primary sequence comparisons still appear warranted to address the whole natural diversity. The recent breakthroughs in protein structure prediction, in particular, could greatly benefit microbial dark matter analyses, as 3D structures tend to be more conserved than primary sequences during evolution. As such, the development of 3D similarity networks, connecting protein structures from cultured organisms to structures predicted from metagenomes, could offer unprecedented insights into the evolution and the functional landscape of environmental microbiomes, with possible applications to fields such as ecological, biotechnological or biomedical sciences.

Materials & Methods

Constitution of a conserved protein families dataset

We constituted a dataset of 9,737,821 proteins, from 4403 bacterial (including CPR), 567 archaeal (including DPANN and Asgard), 120 eukaryotic, 18,020 viral and 1586 plasmidic genomes, acquired from public NCBI databases [77] (Table SI-1). The sequence similarity network (SSN) of this protein collection was reconstructed by an all-against-all DIAMOND blastp alignment [78] (version 2.0.9, thresholds: E-value $\leq 10^{-5}$, sequence identity ≥30%, mutual coverage ≥80%). This SSN contained 891,459 protein clusters (connected components). The assortative mixing between Domains of life within each cluster was computed using the Python package networkx [79] (version 2.8.8). We retained 53 protein clusters meeting thresholds of (i) Domain assortativity ≥0.65 and (ii) 150 or more sequences from both archaea and bacteria. These 53 protein families comprised a total of 125,774 sequences.

Iterative retrieval of environmental homologues

40,154,822 gene sequences from the Ocean Microbial Reference Gene Catalog (OM-RGC v1) [39] were collected, alongside corresponding sampling metadata and eggNOG [80] annotations, and translated into amino-acid sequences. An iterative search for environmental homologues in the OM-RGC dataset was conducted for the selected 53 protein families independently (building upon [37]). For each family, seed sequences were aligned against the OM-RGC protein sequences with DIAMOND (thresholds: E-value $\leq 10^{-5}$, sequence identity ≥30%, mutual coverage ≥80%). Environmental sequences retrieved were used as a base for a new round of DIAMOND alignment (identical parameters) against OM-RGC. This procedure was iterated, each round using as queries the environmental sequences retrieved in the previous round, until no additional sequence was found (Fig. 1A). At each step, the aligned regions of matched sequences were checked to project back to a region covering at least 80% of a seed sequence, to maintain the plausibility of distant homology

between indirectly linked sequences (Fig. 1B-C). Sequences not meeting this criterion were discarded before the next search iteration. 826,717 sequences in the OM-RGC dataset were assigned to the selected protein families in this way (Table SI-2).

Precision and accuracy of our iterative retrieval protocol on simulated protein families

From a balanced binary tree with 64 leaves, we generated a collection of toy phylogenies. For each non-root node in the starting tree, new trees were created by elongating branches between the root and this node, by a factor of 1 ("null" case), 1.5, 2, 2.5, 3, 3.5, 4, 6 or 8, yielding 126 non-root nodes × 9 possible elongation factors = 1134 (non-unique) tree instances. Random sequences of 300 amino-acids were then generated and numerically evolved along the branches of these trees using pyvolve [81] (version 1.0.3, LG model). Doing three replicates per tree instance, we thus simulated a total of 3402 artificial protein families with 64 members each.

In each tree we generated, branches were only elongated from the root to one target node, and therefore only on one side of the root, leading to leaf nodes on that side being further away from the root than the leaves on the opposite side. Sequences simulated along those trees could therefore be classified as slow- or fast-evolving depending on their side in the tree. 3402 iterative homology searches (same parameters as for real-world data) were thus conducted, each time using the slow-evolving sequences from one simulated family to find their fast-evolving homologues within the entire set of generated sequences. The precision (percentage of true positive homology calls amongst all retrieved sequences) and recall (percentage of fast-evolving homologues successfully retrieved) of the search protocol were determined from these results, for each possible factor of divergence, and each possible depth in the tree this divergence spanned (from 1, stopping at a node directly under the root, to 6, all the way to a leaf node).

Comparison of retrieved environmental sequences to cultured diversity

Environmental sequences retrieved for each of the 53 selected seed families were compared to published sequences from taxonomically-resolved organisms in the NCBI *nr* database (downloaded in March 2020) via a DIAMOND alignment search (E-value $\leq 10^{-5}$). Similarity values between environmental sequences and their closest published relative were calculated as the product of the amino-acid identity in the aligned region times the alignment coverage on the shortest sequence.

Sequence similarity network reconstruction and analysis

SSNs were computed for each environmentally expanded protein family by conducting all-against-all DIAMOND blastp alignments of seed and environmental sequences (E-value $\leq 10^{-5}$, sequence identity ≥30%, mutual coverage ≥80%). We then inferred, using Louvain clustering (implemented in networkx, v2.8.8) [40], node communities in those networks, i.e. groups of sequences tightly connected by homology links. This clustering defined 691 communities across the 53 families in our dataset. We further selected clusters containing at least 30 sequences, of which at least 90% were from the environmental dataset, and with environmental sequences averaging 40% identity or less with their closest published counterpart. 80 such clusters were identified across 25 families.

SSNs were rendered using Cytoscape (version 3.9.1) [82]. However larger networks, typically with millions of edges, made visualisations intractable. Synthetic "meta-networks" of those SSNs were created instead (Fig. SI-4, SI-5, SI-7). Rather than showing interconnections between all sequences, these represented connections between sequence clusters (as defined above): each Louvain cluster inferred in an SSN was condensed to a single "meta-node", and two meta-nodes were linked by a "meta-edge" if the corresponding clusters were adjacent in the SSN. Meta-edges were also given a numeric weight representing the proportion of edges between clusters, relative to the total possible number of edges if the clusters had been fully connected together.

Phylogenetic analysis of divergent clusters

Sequences from divergent clusters were gathered in phylogenetic trees along with seed sequences. We used CD-HIT (90% identity threshold, version 4.8.1) [83] to dereplicate sequences from each of the 80 selected clusters, as well as seed sequences from each of the 25 corresponding families. Up to 100 sequences per environmental cluster and 200 seeds per family were selected as representatives. We then first computed cluster-specific maximum likelihood phylogenies. Sequences from each divergent cluster were aligned with corresponding seed sequences using Mafft (version 7.520, 1000 iterative refinement cycles) [84]. These alignments were then trimmed using trimAl (version 1.4.1) [85], and phylogenies were produced using IQ-TREE (version 1.6.12, 1000 bootstrap replicates) [86–88]. Next, we inferred family-wide alignment-free phylogenies, grouping together (representatives of) seed sequences and all divergent clusters from each family [89, 90]. *k*-mer-based distance matrices were computed between all representative sequences of a family using jD2Stat (version 1.0, *k*=7) [91], and used to infer Neighbour-Joining trees with RapidNJ (version 2.3.2) [92]. All trees were rendered and annotated in iTOL (version 6.9) [93].

Inference and comparison of protein tertiary structures

3D structures were inferred for a selection of representative sequences in the SSNs of SMC proteins and DNA clamp-loading subunits.

For clamp loaders, one sequence was selected as representative for each cluster in the SSN. Divergent environmental clusters were represented by the environmental sequence with the highest degree (number of edges in the SSN) to other environmental sequences within the cluster; other clusters were represented by the reference sequence with the highest degree to other references in the cluster. For SMC proteins, which have a significantly longer primary sequence (around 1200 amino-acids), we sought to reduce the number of structures to infer *de novo.* Six sequences from the divergent cluster of environmental SMC variants were chosen arbitrarily (all had maximal degree, because the cluster was fully connected), and public AlphaFold structures [53, 94] were acquired

527 from UniProt [95] to represent reference SMC sequences (UniProtKB accessions: P9WGF2, Q5N0D2,
528 A3PMS2, A9BZW2, P51834, Q69GZ5, Q8TZY2) and their Rad50/SbcC homologues (UniProtKB
529 accessions: A0A7I7YPX7, A5GLL1, O68032, A0A210VWK9, A0A640H0H1, P62134, P58301).

530 Structures were inferred for selected clamp loaders and environmental SMC sequences using
531 ColabFold (v1.5.2, default parameters) [52]. Then, reference and environmental clamp loader
532 structures were compared using FoldSeek (version 7-04e0ec8, all-against-all, easy-search mode, no
533 pre-filter, alignment by TM-Align) [96]. Inferred environmental SMC structures were compared with
534 UniProt reference SMC structures following the same protocol. For both protein families, these
535 comparisons were used to construct dendrograms with RapidNJ [92], taking as distance metric
536 between two structures the average local distance difference test (lDDT) score of the corresponding
537 bidirectional structural alignment. Dendrograms were plotted in iTOL [93] and annotated with 3D
538 models of the protein structures rendered by PyMOL (version 2.5.5).

Declarations

Ethics approval and consent to participate: Not applicable.

Consent for publication: Not applicable.

Availability of data and materials: The OM-RGC dataset analysed in this study is available from this webpage: http://ocean-microbiome.embl.de/companion.html. The dataset of conserved protein families constructed for this analysis is available from this Figshare project: https://doi.org/10.6084/m9.figshare.24893910.v1. Source code for the iterative retrieval of environmental homologues can be found on the following repository: https://github.com/TeamAIRE/SHIFT.

Competing interests: The authors declare that they have no competing interests.

Funding: RL, GB, PM and EB were supported by the European Research Council under the European Community's Seventh Framework Programme (FP7/2007-2013 Grant Agreement # 615274, category LS8).

Authors' contributions: DS conducted the formal analyses, with RL contributing to the creation of the seed dataset. DS, RL, EC, GB and PM contributed to methods design and software implementations. EB, EP and PL designed and supervised the study. DS wrote the original draft, with the help of EB, EP and PL to edit and finalise the manuscript. All authors read and approved the final manuscript.

Acknowledgements: The authors are grateful to Charles Bernard for his insights on designing the iterative search procedure.

558 **References**

1. Staley JT, Konopka A. Measurement of in situ activities of nonphotosynthetic microorganisms in aquatic and terrestrial habitats. Annu Rev Microbiol. 1985;39:321–46.

2. Amann RI, Ludwig W, Schleifer KH. Phylogenetic identification and in situ detection of individual microbial cells without cultivation. Microbiol Rev. 1995;59:143–69.

3. Whitman WB, Coleman DC, Wiebe WJ. Prokaryotes: The unseen majority. Proc Natl Acad Sci. 1998;95:6578–83.

4. Marcy Y, Ouverney C, Bik EM, Lösekann T, Ivanova N, Martin HG, et al. Dissecting biological "dark matter" with single-cell genetic analysis of rare and uncultivated TM7 microbes from the human mouth. Proc Natl Acad Sci U S A. 2007;104:11889–94.

5. Alain K, Querellou J. Cultivating the uncultured: limits, advances and future challenges. Extremophiles. 2009;13:583–94.

6. Koch R. Untersuchungen uber bakterien V. Die aetiologie der milzbrand-krankheit, begrunder auf die entwicklungegeschichte Bacillus anthracis. Beitrage Zur Biol Pflanz. 1877;2:277–310.

7. Handelsman J, Rondon MR, Brady SF, Clardy J, Goodman RM. Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. Chem Biol. 1998;5:R245–9.

8. Castelle CJ, Banfield JF. Major New Microbial Groups Expand Diversity and Alter our Understanding of the Tree of Life. Cell. 2018;172:1181–97.

9. Delmont TO, Robe P, Cecillon S, Clark IM, Constancias F, Simonet P, et al. Accessing the Soil Metagenome for Studies of Microbial Diversity. Appl Environ Microbiol. 2011;77:1315–24.

10. Ventosa A, de la Haba RR, Sánchez-Porro C, Papke RT. Microbial diversity of hypersaline environments: a metagenomic approach. Curr Opin Microbiol. 2015;25:80–7.

11. Behzad H, Gojobori T, Mineta K. Challenges and Opportunities of Airborne Metagenomics. Genome Biol Evol. 2015;7:1216–26.

12. Sunagawa S, Acinas SG, Bork P, Bowler C, Eveillard D, Gorsky G, et al. Tara Oceans: towards global ocean ecosystems biology. Nat Rev Microbiol. 2020;18:428–45.

13. Hugenholtz P, Pitulle C, Hershberger KL, Pace NR. Novel division level bacterial diversity in a Yellowstone hot spring. J Bacteriol. 1998;180:366–76.

14. Chouari R, Le Paslier D, Dauga C, Daegelen P, Weissenbach J, Sghir A. Novel major bacterial candidate division within a municipal anaerobic sludge digester. Appl Environ Microbiol. 2005;71:2145–53.

15. Pelletier E, Kreimeyer A, Bocs S, Rouy Z, Gyapay G, Chouari R, et al. “Candidatus Cloacamonas Acidaminovorans”: Genome Sequence Reconstruction Provides a First Glimpse of a New Bacterial Division. J Bacteriol. 2008;190:2572–9.

16. Hug LA, Baker BJ, Anantharaman K, Brown CT, Probst AJ, Castelle CJ, et al. A new view of the tree of life. Nat Microbiol. 2016;1:1–6.

17. Parks DH, Rinke C, Chuvochina M, Chaumeil P-A, Woodcroft BJ, Evans PN, et al. Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. Nat Microbiol. 2017. https://doi.org/10.1038/s41564-017-0012-7.

18. Rinke C, Schwientek P, Sczyrba A, Ivanova NN, Anderson IJ, Cheng J-F, et al. Insights into the phylogeny and coding potential of microbial dark matter. Nature. 2013;499:431–7.

19. Brown CT, Hug LA, Thomas BC, Sharon I, Castelle CJ, Singh A, et al. Unusual biology across a group comprising more than 15% of domain Bacteria. Nature. 2015;523:208–11.

20. Huber H, Hohn MJ, Rachel R, Fuchs T, Wimmer VC, Stetter KO. A new phylum of Archaea represented by a nanosized hyperthermophilic symbiont. Nature. 2002;417:63–7.

21. Baker BJ, Comolli LR, Dick GJ, Hauser LJ, Hyatt D, Dill BD, et al. Enigmatic, ultrasmall, uncultivated Archaea. Proc Natl Acad Sci. 2010;107:8806–11.

22. Spang A, Saw JH, Jørgensen SL, Zaremba-Niedzwiedzka K, Martijn J, Lind AE, et al. Complex archaea that bridge the gap between prokaryotes and eukaryotes. Nature. 2015;521:173–9.

23. Zaremba-Niedzwiedzka K, Caceres EF, Saw JH, Bäckström D, Juzokaite L, Vancaester E, et al. Asgard archaea illuminate the origin of eukaryotic cellular complexity. Nature. 2017;541:353–8.

24. Imachi H, Nobu MK, Nakahara N, Morono Y, Ogawara M, Takaki Y, et al. Isolation of an archaeon at the prokaryote–eukaryote interface. Nature. 2020;577:519–25.

25. Paez-Espino D, Eloe-Fadrosh EA, Pavlopoulos GA, Thomas AD, Huntemann M, Mikhailova N, et al. Uncovering Earth’s virome. Nature. 2016;536:425–30.

26. Zhou Y, Zhou L, Yan S, Chen L, Krupovic M, Wang Y. Diverse viruses of marine archaea discovered using metagenomics. Environ Microbiol. 2023;25:367–82.

27. Gaïa M, Meng L, Pelletier E, Forterre P, Vanni C, Fernandez-Guerra A, et al. Mirusviruses link herpesviruses to giant viruses. Nature. 2023;616:783–9.

28. Al-Shayeb B, Schoelmerich MC, West-Roberts J, Valentin-Alvarado LE, Sachdeva R, Mullen S, et al. Borgs are giant genetic elements with potential to expand metabolic capacity. Nature. 2022;610:731–6.

29. Lloyd KG, Steen AD, Ladau J, Yin J, Crosby L. Phylogenetically Novel Uncultured Microbial Cells Dominate Earth Microbiomes. mSystems. 2018;3.

30. Nayfach S, Roux S, Seshadri R, Udwary D, Varghese N, Schulz F, et al. A genomic catalog of Earth’s microbiomes. Nat Biotechnol. 2020;39:499–509.

31. Bernard G, Pathmanathan JS, Lannes R, Lopez P, Bapteste E. Microbial Dark Matter Investigations: How Microbial Studies Transform Biological Knowledge and Empirically Sketch a Logic of Scientific Discovery. Genome Biol Evol. 2018;10:707–15.

32. Liu Z, Ma A, Mathé E, Merling M, Ma Q, Liu B. Network analyses in microbiome based on high-throughput multi-omics data. Brief Bioinform. 2021;22:1639–55.

33. Zamkovaya T, Foster JS, de Crécy-Lagard V, Conesa A. A network approach to elucidate and prioritize microbial dark matter in microbial communities. ISME J. 2021;15:228–44.

34. Forster D, Bittner L, Karkar S, Dunthorn M, Romac S, Audic S, et al. Testing ecological theories with sequence similarity networks: marine ciliates exhibit similar geographic dispersal patterns as multicellular organisms. BMC Biol. 2015;13:16.

35. Arroyo AS, Iannes R, Bapteste E, Ruiz-Trillo I. Gene Similarity Networks Unveil a Potential Novel Unicellular Group Closely Related to Animals from the Tara Oceans Expedition. Genome Biol Evol. 2020;12:1664–78.

36. Lynch MDJ, Bartram AK, Neufeld JD. Targeted recovery of novel phylogenetic diversity from next-generation sequence data. ISME J. 2012;6:2067–77.

37. Lopez P, Halary S, Bapteste E. Highly divergent ancient gene families in metagenomic samples are compatible with additional divisions of life. Biol Direct. 2015;10:64.

38. Durairaj J, Waterhouse AM, Mets T, Brodiazhenko T, Abdullah M, Studer G, et al. Uncovering new families and folds in the natural protein universe. Nature. 2023;622:646–53.

39. Sunagawa S, Coelho LP, Chaffron S, Kultima JR, Labadie K, Salazar G, et al. Structure and function of the global ocean microbiome. Science. 2015;348:1261359.

40. Blondel VD, Guillaume J-L, Lambiotte R, Lefebvre E. Fast unfolding of communities in large networks. J Stat Mech Theory Exp. 2008;2008:P10008.

41. Vanni C, Schechter MS, Acinas SG, Barberán A, Buttigieg PL, Casamayor EO, et al. Unifying the known and unknown microbial coding sequence space. eLife. 2022;11.

42. Iyer LM, Leipe DD, Koonin EV, Aravind L. Evolutionary history and higher order classification of AAA+ ATPases. J Struct Biol. 2004;146:11–31.

43. Hedglin M, Kumar R, Benkovic SJ. Replication Clamps and Clamp Loaders. Cold Spring Harb Perspect Biol. 2013;5:a010165.

44. O'Donnell M, Onrust R, Dean FB, chen M, Hurwitz J. Homology in accessory proteins of replicative polymerases—E.coli to humans. Nucleic Acids Res. 1993;21:1–3.

45. Chia N, Cann I, Olsen GJ. Evolution of DNA Replication Protein Complexes in Eukaryotes and Archaea. PLOS ONE. 2010;5:e10866.

46. Yao NY, O'Donnell ME. Evolution of replication machines. Crit Rev Biochem Mol Biol. 2016;51:135–49.

47. Li H, O'Donnell M, Kelch B. Unexpected new insights into DNA clamp loaders. Bioessays. 2022;44:2200154.

48. Barre F-X, Søballe B, Michel B, Aroyo M, Robertson M, Sherratt D. Circles: The replication-recombination-chromosome segregation connection. Proc Natl Acad Sci. 2001;98:8189–95.

49. Romero H, Rösch TC, Hernández-Tamayo R, Lucena D, Ayora S, Alonso JC, et al. Single molecule tracking reveals functions for RarA at replication forks but also independently from replication during DNA repair in Bacillus subtilis. Sci Rep. 2019;9:1997.

50. Frickey T, Lupas AN. Phylogenetic analysis of AAA proteins. J Struct Biol. 2004;146:2–10.

51. Lapointe F-J, Lopez P, Boucher Y, Koenig J, Bapteste E. Clanistics: a multi-level perspective for harvesting unrooted gene trees. Trends Microbiol. 2010;18:341–7.

52. Mirdita M, Schütze K, Moriwaki Y, Heo L, Ovchinnikov S, Steinegger M. ColabFold: making protein folding accessible to all. Nat Methods. 2022;19:679–82.

53. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate protein structure prediction with AlphaFold. Nature. 2021;596:583–9.

54. Cobbe N, Heck MMS. The Evolution of SMC Proteins: Phylogenetic Analysis and Structural Implications. Mol Biol Evol. 2004;21:332–47.

55. Rybenkov VV, Herrera V, Petrushenko ZM, Zhao H. MukBEF, a Chromosomal Organizer. J Mol Microbiol Biotechnol. 2015;24:371–83.

56. Petrushenko ZM, She W, Rybenkov VV. A new family of bacterial condensins. Mol Microbiol. 2011;81:881–96.

57. Kamada K, Barillà D. Combing Chromosomal DNA Mediated by the SMC Complex: Structure and Mechanisms. BioEssays. 2018;40:1700166.

58. Badel C, Bell SD. Chromosome architecture in an archaeal species naturally lacking structural maintenance of chromosomes proteins. Nat Microbiol. 2024;9:263–73.

59. Soppa J. Prokaryotic structural maintenance of chromosomes (SMC) proteins: distribution, phylogeny, and comparison with MukBs and additional prokaryotic and eukaryotic coiled-coil proteins. Gene. 2001;278:253–64.

60. Waldman VM, Stanage TH, Mims A, Norden IS, Oakley MG. Structural mapping of the coiled-coil domain of a bacterial condensin and comparative analyses across all domains of life suggest conserved features of SMC proteins. Proteins Struct Funct Bioinforma. 2015;83:1027–45.

61. Hirano T. The ABCs of SMC proteins: two-armed ATPases for chromosome condensation, cohesion, and repair. Genes Dev. 2002;16:399–414.

62. Gruber S, Arumugam P, Katou Y, Kuglitsch D, Helmhart W, Shirahige K, et al. Evidence that Loading of Cohesin Onto Chromosomes Involves Opening of Its SMC Hinge. Cell. 2006;127:523–37.

63. Connelly JC, Leach DRF. Tethering on the brink: the evolutionarily conserved Mre11–Rad50 complex. Trends Biochem Sci. 2002;27:410–8.

64. Kim E, Barth R, Dekker C. Looping the Genome with SMC Complexes. Annu Rev Biochem. 2023;92 Volume 92, 2023:15–41.

65. Lin Z, Kong H, Nei M, Ma H. Origins and evolution of the recA/RAD51 gene family: Evidence for ancient gene duplication and endosymbiotic gene transfer. Proc Natl Acad Sci. 2006;103:10328–33.

66. Thompson LH, Schild D. Homologous recombinational repair of DNA ensures mammalian chromosome stability. Mutat Res Mol Mech Mutagen. 2001;477:131–53.

67. Cox MM. The RecA protein as a recombinational repair system. Mol Microbiol. 1991;5:1295–9.

68. Seitz EM, Brockman JP, Sandler SJ, Clark AJ, Kowalczykowski SC. RadA protein is an archaeal RecA protein homolog that catalyzes DNA strand exchange. Genes Dev. 1998;12:1248–53.

69. Chintapalli SV, Bhardwaj G, Babu J, Hadjiyianni L, Hong Y, Todd GK, et al. Reevaluation of the evolutionary events within recA/RAD51 phylogeny. BMC Genomics. 2013;14:240.

70. Banerjee A, Neiner T, Tripp P, Albers S-V. Insights into subunit interactions in the Sulfolobus acidocaldarius archaellum cytoplasmic complex. FEBS J. 2013;280:6141–9.

71. Chaudhury P, Does C van der, Albers S-V. Characterization of the ATPase FlaI of the motor complex of the Pyrococcus furiosus archaellum and its interactions between the ATP-binding protein FlaH. PeerJ. 2018;6:e4984.

72. Castelle CJ, Brown CT, Anantharaman K, Probst AJ, Huang RH, Banfield JF. Biosynthetic capacity, metabolic variety and unusual biology in the CPR and DPANN radiations. Nat Rev Microbiol. 2018;16:629–45.

73. Schulz F, Roux S, Paez-Espino D, Jungbluth S, Walsh DA, Denef VJ, et al. Giant virus diversity and host interactions through global metagenomics. Nature. 2020;578:432–6.

74. Wu D, Wu M, Halpern A, Rusch DB, Yooseph S, Frazier M, et al. Stalking the Fourth Domain in Metagenomic Data: Searching for, Discovering, and Interpreting Novel, Deep Branches in Marker Gene Phylogenetic Trees. PLOS One. 2011;6.

75. Makarova KS, Krupovic M, Koonin EV. Evolution of replicative DNA polymerases in archaea and their contributions to the eukaryotic replication machinery. Front Microbiol. 2014;5.

76. Lennon CW, Stanger M, Banavali NK, Belfort M. Conditional Protein Splicing Switch in Hyperthermophiles through an Intein-Extein Partnership. mBio. 2018;9:10.1128/mbio.02304-17.

77. Sayers EW, Bolton EE, Brister JR, Canese K, Chan J, Comeau DC, et al. Database resources of the national center for biotechnology information. Nucleic Acids Res. 2022;50:D20–6.

78. Buchfink B, Reuter K, Drost H-G. Sensitive protein alignments at tree-of-life scale using DIAMOND. Nat Methods. 2021;18:366–8.

79. Hagberg AA, Schult DA, Swart PJ. Exploring Network Structure, Dynamics, and Function using NetworkX. In: Varoquaux G, Vaught T, Millman J, editors. Proceedings of the 7th Python in Science Conference. Pasadena, CA USA; 2008. p. 11–5.

80. Huerta-Cepas J, Szklarczyk D, Heller D, Hernández-Plaza A, Forslund SK, Cook H, et al. eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. Nucleic Acids Res. 2019;47:D309–14.

81. Spielman SJ, Wilke CO. Pyvolve: A Flexible Python Module for Simulating Sequences along Phylogenies. PLOS ONE. 2015;10:e0139047.

82. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. 2003;13:2498–504.

83. Li W, Godzik A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinforma Oxf Engl. 2006;22:1658–9.

84. Katoh K, Standley DM. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. Mol Biol Evol. 2013;30:772–80.

85. Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics. 2009;25:1972–3.

86. Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, von Haeseler A, et al. IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. Mol Biol Evol. 2020;37:1530–4.

87. Hoang DT, Chernomor O, von Haeseler A, Minh BQ, Vinh LS. UFBoot2: Improving the Ultrafast Bootstrap Approximation. Mol Biol Evol. 2018;35:518–22.

88. Kalyaanamoorthy S, Minh BQ, Wong TKF, von Haeseler A, Jermiin LS. ModelFinder: fast model selection for accurate phylogenetic estimates. Nat Methods. 2017;14:587–9.

89. Zielezinski A, Vinga S, Almeida J, Karlowski WM. Alignment-free sequence comparison: benefits, applications, and tools. Genome Biol. 2017;18:1–17.

90. Ren J, Bai X, Lu YY, Tang K, Wang Y, Reinert G, et al. Alignment-Free Sequence Analysis and Applications. Annu Rev Biomed Data Sci. 2018;1:93–114.

91. Chan CX, Bernard G, Poirion O, Hogan JM, Ragan MA. Inferring phylogenies of evolving sequences without multiple sequence alignment. Sci Rep. 2014;4:6504.

92. Simonsen M, Mailund T, Pedersen CNS. Rapid Neighbour-Joining. In: Crandall KA, Lagergren J, editors. Algorithms in Bioinformatics. Berlin, Heidelberg: Springer; 2008. p. 113–22.

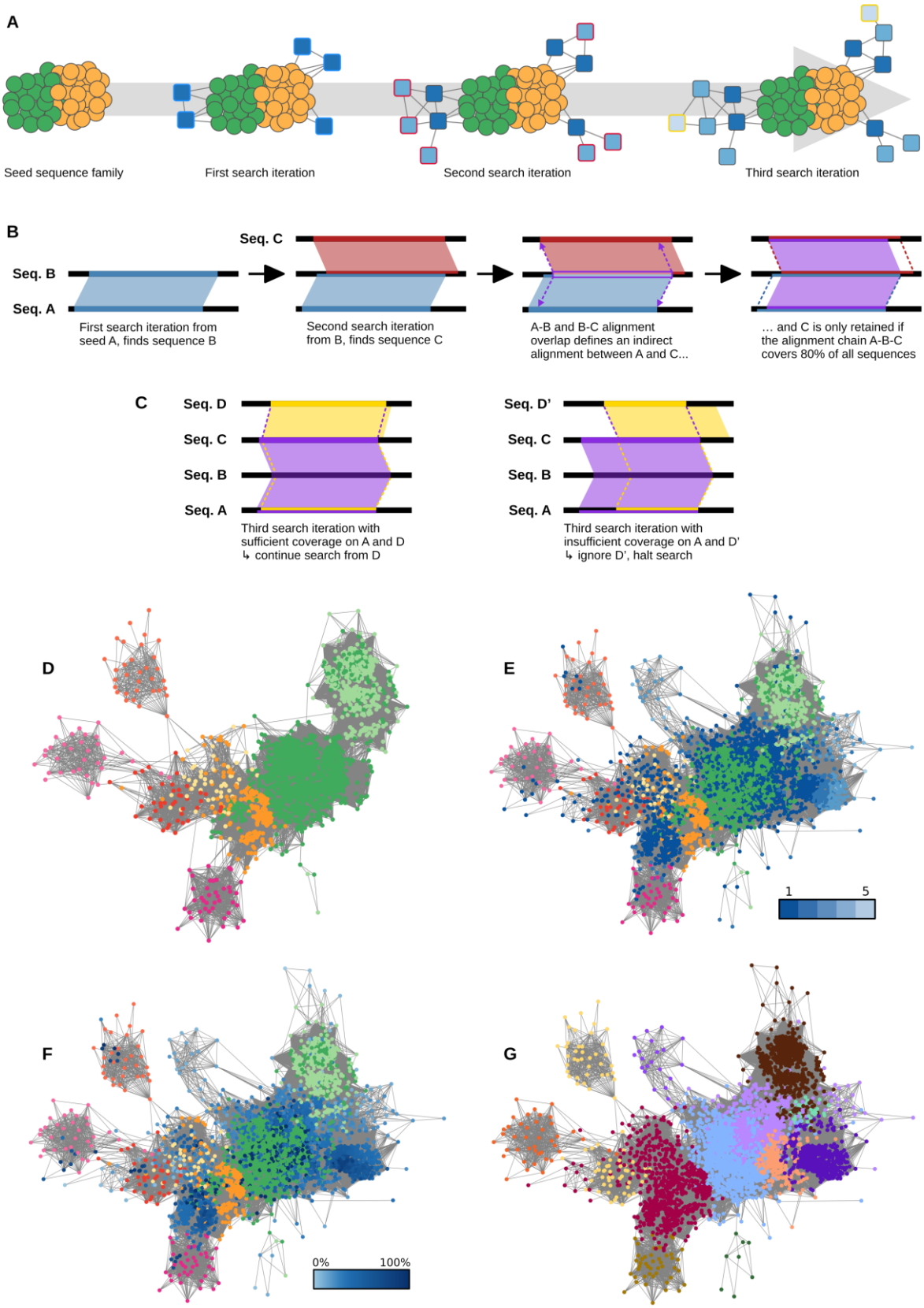
93. Letunic I, Bork P. Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. Nucleic Acids Res. 2021;49:W293–6.

94. Varadi M, Anyango S, Deshpande M, Nair S, Natassia C, Yordanova G, et al. AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. Nucleic Acids Res. 2022;50:D439–44.

95. The UniProt Consortium. UniProt: the Universal Protein Knowledgebase in 2023. Nucleic Acids Res. 2023;51:D523–31.

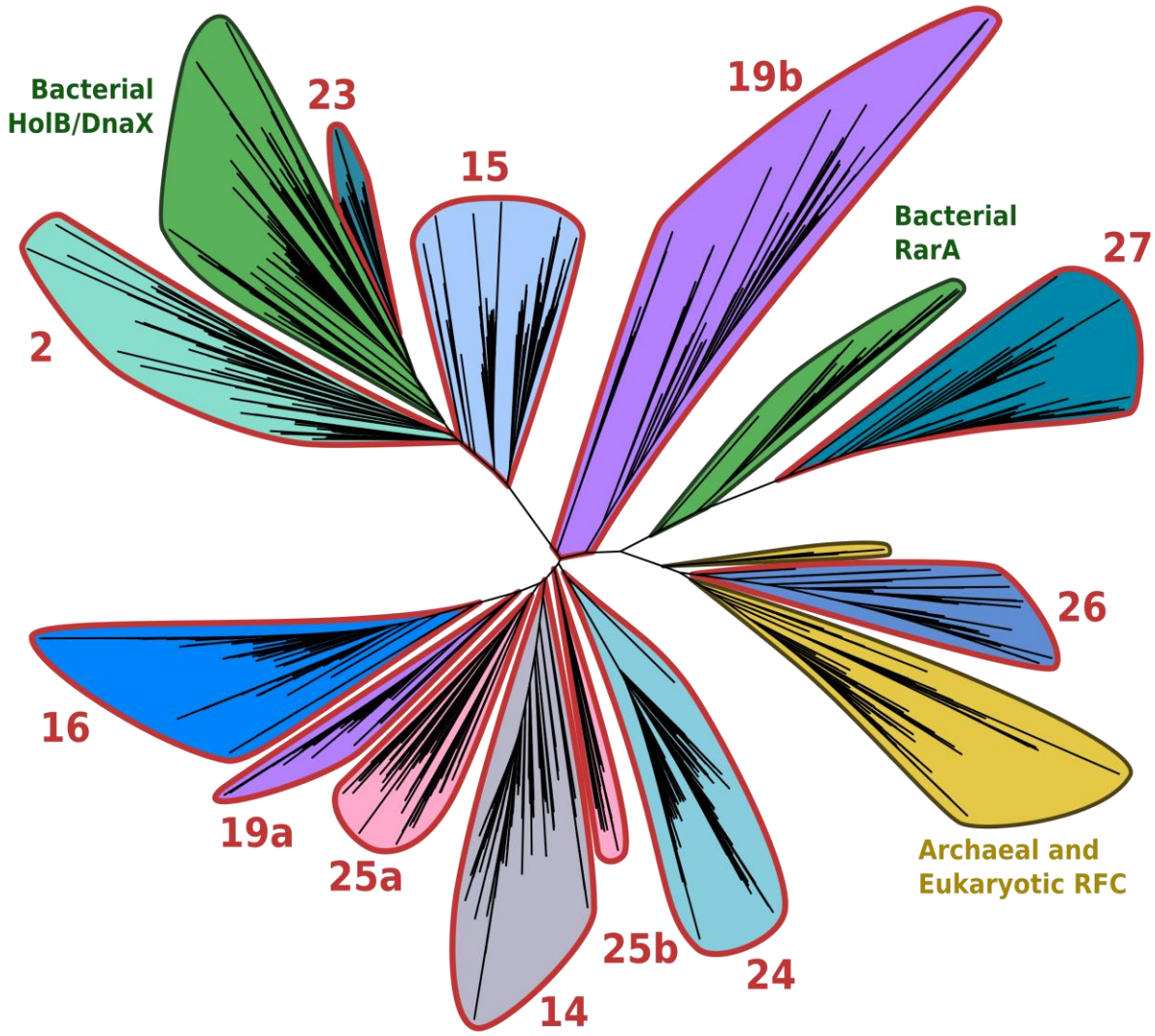
96. van Kempen M, Kim SS, Tumescheit C, Mirdita M, Lee J, Gilchrist CLM, et al. Fast and accurate protein structure search with Foldseek. Nat Biotechnol. 2023;:1–4.

559 **Figures**



560

Figure 1: Iterative homologue search procedure. (**A**) Iterative aggregation of environmental homologues around seed sequences in a similarity network. From a set of seed sequences belonging to a given protein family (green and orange nodes), a first search iteration finds environmental homologues (dark blue nodes) for some of the seeds. A second search iteration then uses these environmental sequences as queries to find more homologues (medium blue nodes, red frame), which are themselves used as queries for a third search iteration finding further environmental homologues (light blue nodes, yellow frame). (**B**) At each iteration of the search, newly found homologues are only retained if their aligned region can be mapped back onto a seed sequence in a way that ensures >80% coverage on all sequences along the chain of aligned sequences. (**C**) **Left**: sequence D is found after three search iterations from seed A, and its alignment with sequence C can be mapped back to A in a way that preserves 80% coverage on all sequences along the "alignment chain". Sequence D is therefore retained and will be used as query for the next iteration of the search. **Right**: sequence D' is found after three search iterations from seed A, but its aligned region cannot be mapped back to A without breaking the 80% coverage requirement. D' is thus not retained as a distant homologue of A in this round of search. (**D-G**) Sequence similarity networks for SMC proteins. (D) shows seed sequences only, (E-G) show seed and environmental sequences. In (D-F), nodes representing seed sequences are coloured according to their taxonomic origin (yellow: non-DPANN archaea; orange: DPANN archaea; light green: CPR bacteria; dark green: non-CPR bacteria; shades of red: four eukaryotic SMC paralogues). In (E), environmental nodes are coloured in blue, with darker shades for sequences retrieved in earlier iterations of the search, and lighter shades for sequences retrieved later. In (F), environmental nodes are coloured in blue, with darker shades for sequences with higher similarity to the known cultured diversity, and lighter shades for sequences with less similarity. In (G), all nodes are coloured according to Louvain clusters inferred in the SSN (one arbitrary colour per cluster).



585

586 **Figure 2: Alignment-free phylogeny of the DNA clamp loader subunits: HolB/DnaX/RarA/RFC**
587 **sequences and environmental homologues from significantly divergent clusters.** Seed sequences
588 are coloured according to the Domain of life of their host organism (green: Bacteria, yellow: Archaea
589 and Eukaryotes). Groups of environmental sequences are coloured according to the network cluster
590 they belong to in the family SSN, and outlined in red. Numerical cluster labels are inherited from Fig.
591 SI-4 and shared with Fig. 3. Note: environmental network clusters 19 and 25 are both split into two
592 groups in this phylogenetic tree.

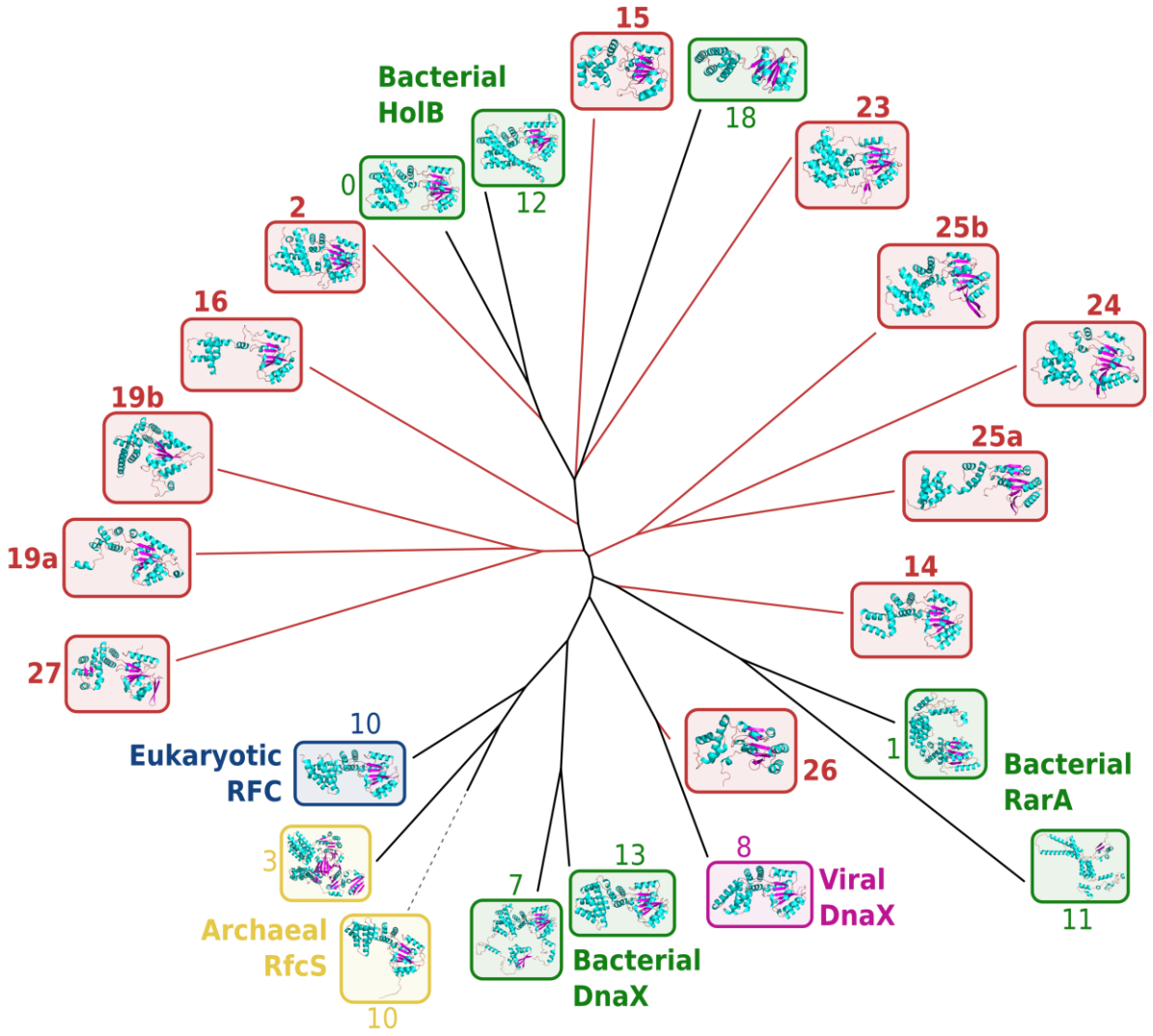
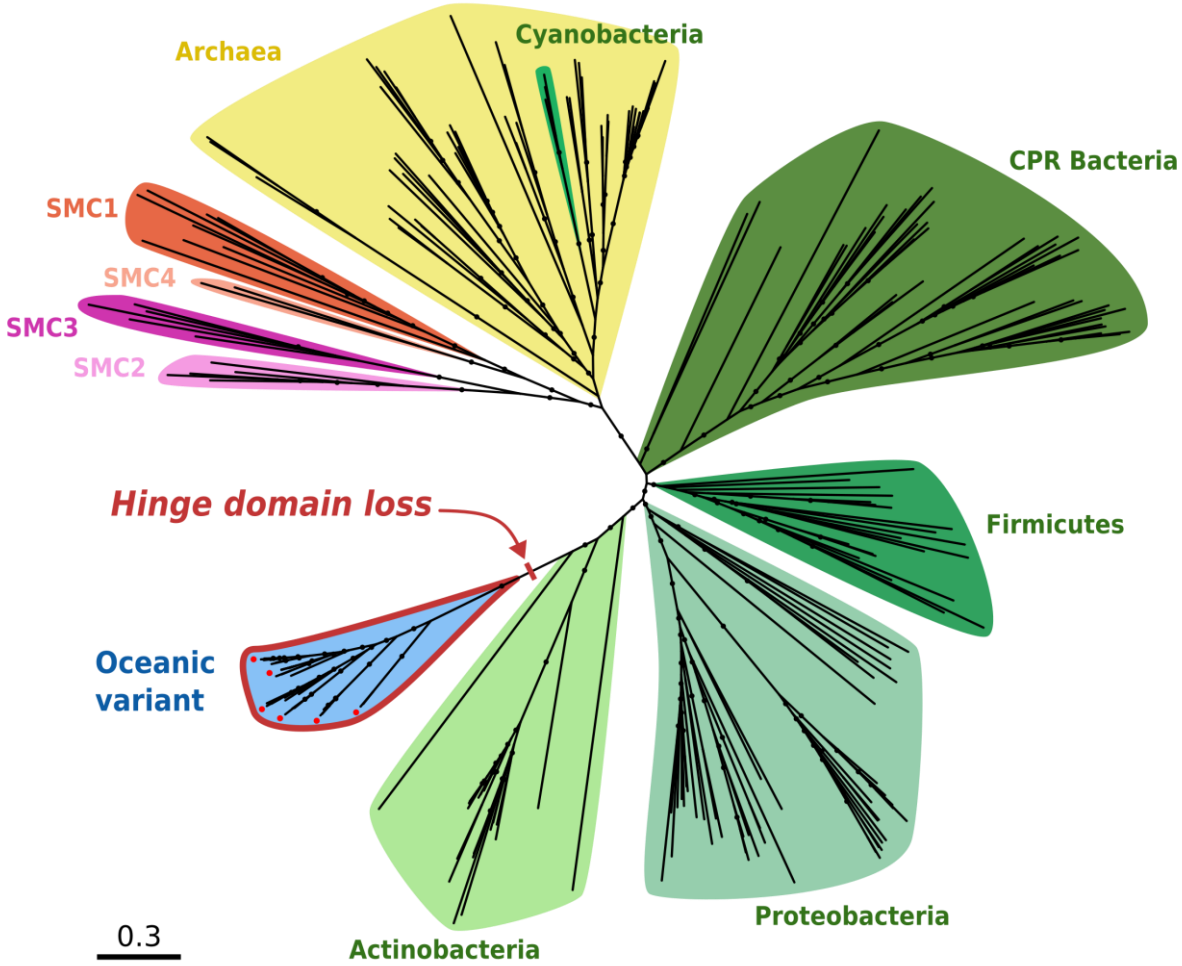


Figure 3: Dendrogram of tertiary structures of DNA clamp loader subunits: HolB/DnaX/RarA/RFC sequences and environmental homologues from significantly divergent clusters. Protein structures were inferred with AlphaFold and compared (all against all) using Foldseek. Leaves and structures are boxed according to the Domain of life of their host organism (green: Bacteria, yellow: Archaea, magenta: Viruses). Environmental leaves and structures are boxed in red, with numerical labels corresponding to the SSN cluster they belong to, in accordance with Fig. 2 and Fig. SI-4.



600

601 **Figure 4: Maximum likelihood phylogenetic tree of SMC sequences and environmental homologues**
602 **from significantly divergent clusters.** Seed sequences are coloured according to the Domain of life of
603 their host organism (green tones: Bacteria, yellow: Archaea, orange and purple tones: Eukaryotes).
604 Environmental sequences are coloured in blue and outlined in red. Red dots indicate environmental
605 sequences for which 3D structures were inferred. Black dots indicate branches with >85% bootstrap
606 support.

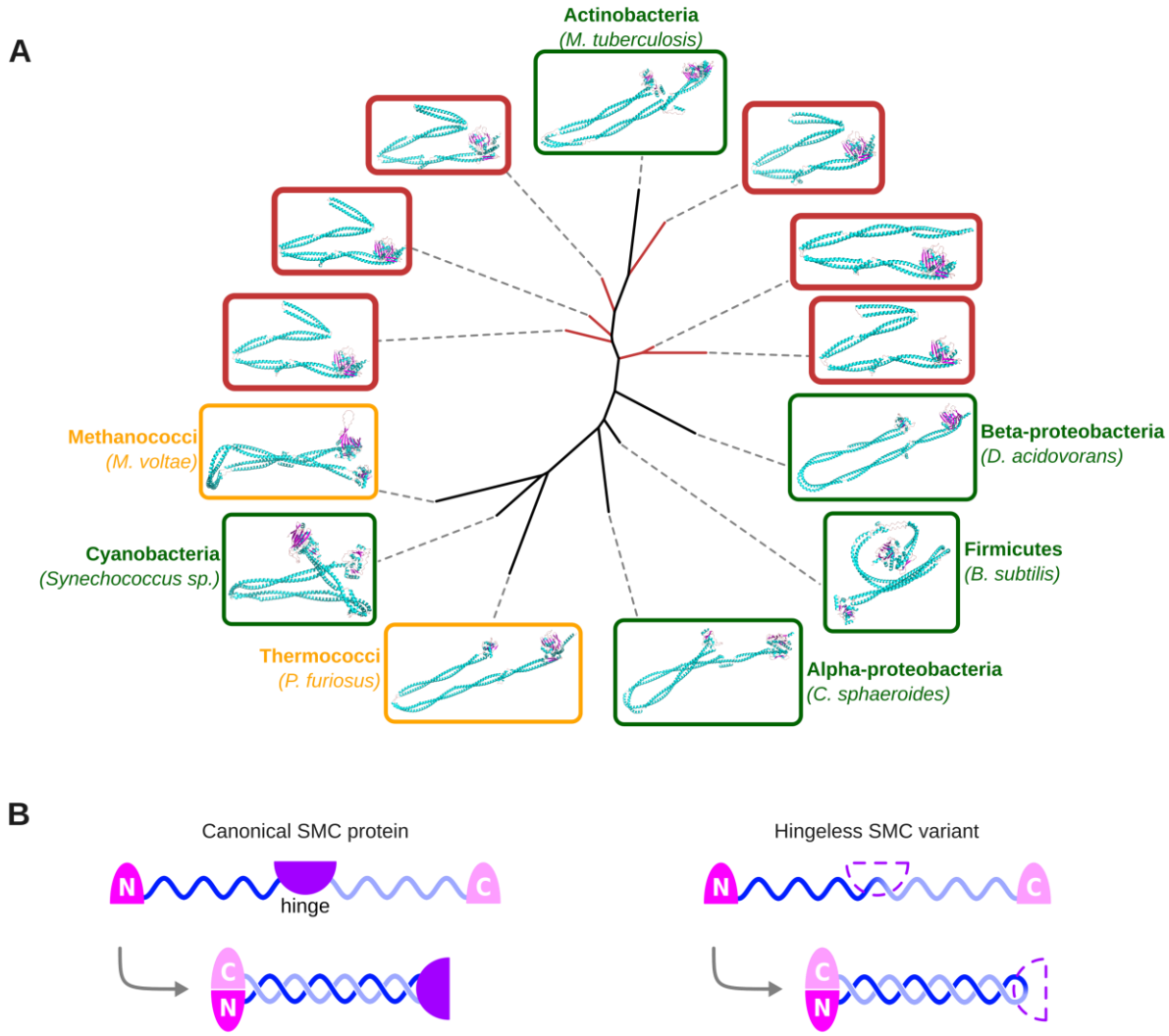


Figure 5: Environmental SMC homologues with divergent tertiary structure. (**A**) Dendrogram of tertiary structures of SMC sequences and selected environmental homologues from significantly divergent clusters. Protein structures were inferred with AlphaFold and compared (all against all) using Foldseek. Leaves and structures are boxed according to the Domain of life of their host organism (green: Bacteria, yellow: Archaea). Environmental leaves and structures are highlighted in red. (**B**) Schematic structure of SMC monomers. Left: canonical SMC protein with N- and C-terminal ATP-binding motifs, linked to a central hinge domain by two coiled-coil regions. This linear structure folds (grey arrow) by joining the two terminal motifs into an ATPase domain, forming a helical coiled-coil with the arm regions between the ATPase and hinge domains. Right: “hinge-less” environmental

SMC homologue lacking a hinge domain. The folded protein still features the ATPase domain at one end of the coiled-coil helix, without the hinge at the opposite end.

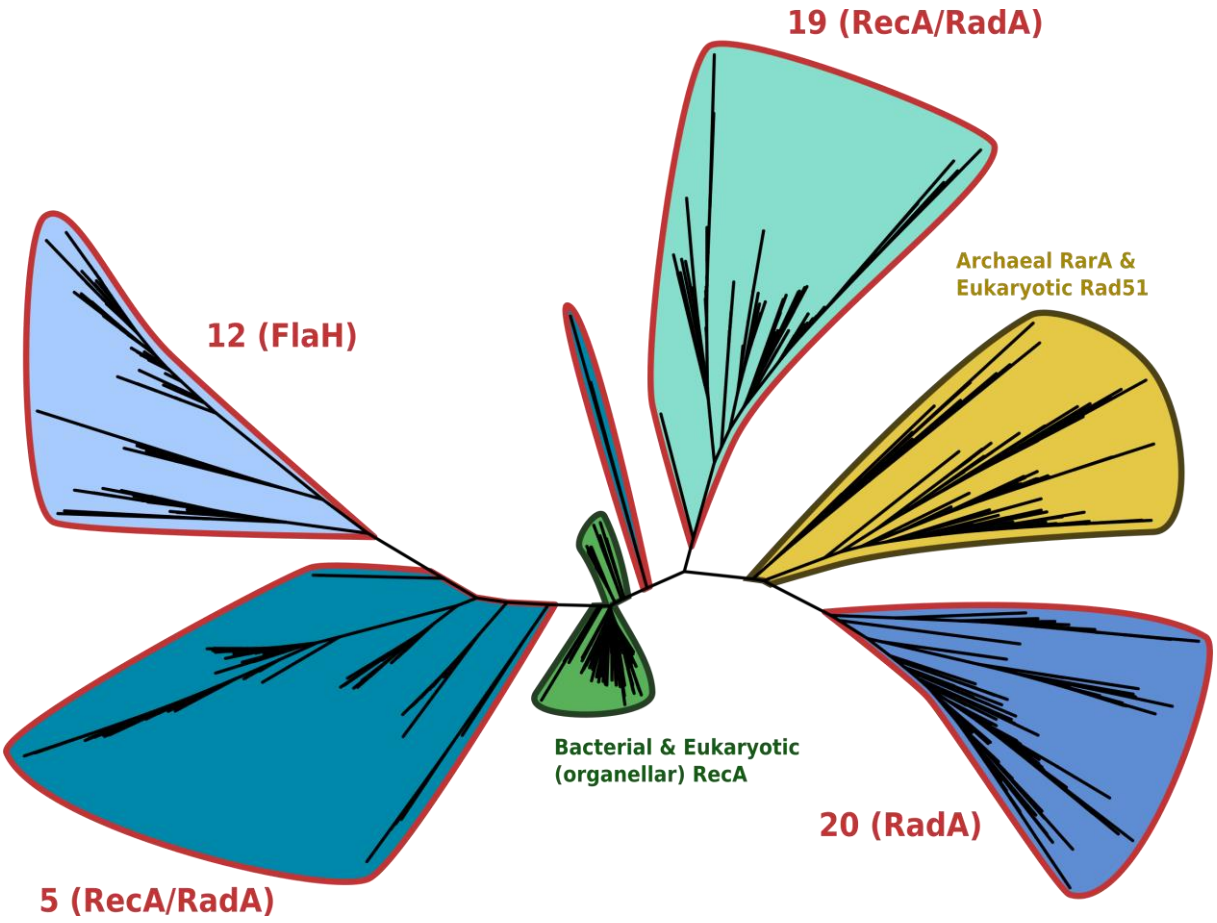


Figure 6: Alignment-free phylogeny of RecA/RadA sequences and environmental homologues from significantly divergent clusters. Seed sequences are coloured according to the Domain of life of their host organism (green: Bacteria and eukaryotic organelles, yellow: Archaea and eukaryotic nuclei). Groups of environmental sequences are coloured according to the network cluster they belong to in the family SSN, and outlined in red. Numerical cluster labels are inherited from Fig. SI-7.

Chapter III. Partial homology and gene remodelling in two multicellular lineages

1. Combinatorics of genes and gene parts

1.1 – Homologous and non-homologous genetic recombination

Gene duplication, sequence divergence, and recombination are generally understood as the three main forces of gene evolution. Among these, genetic recombination stands out conceptually because, unlike duplication and divergence, it goes against the historical model of evolution as a predominantly tree-like process based on clonal replication. The term “recombination” actually encompasses a variety of evolutionary mechanisms that all share a common characteristic: they depict the gene space not as a collection of isolated, atomic units, but rather as a mosaic of compartments that can be arranged and rearranged with some flexibility. Combinatorial processes of gene evolution that involve this mosaicism of genes are fundamentally not tree-like, and therefore (tautologically) gene families that evolved from such processes cannot be accurately described with canonical phylogeny methods. Consequently, remodelled genes can sometimes be overlooked by evolutionary analyses that focus mainly on phylogenetic trees. This is not because of a lack of awareness or consensus on the prevalence of recombination in gene evolution; it is not disputed, for instance, that the majority of protein-coding genes can be decomposed into distinct domains that can be arranged

in a multitude of different architectures [Forslund, Kaduk, and Sonnhammer 2019]. Domains are even the basis for several popular methods of protein function annotation, but investigations into the evolutionary dynamics around domain rearrangements, and more generally around gene remodelling as a whole, are less frequent. In this chapter, we show that network-based approaches can help to detect and analyse these remodelling events, with a particular focus on gene fusions and gene fissions. We studied how these mechanisms have affected the evolution of two different lineages, both characterised by the emergence of complex multicellularity from unicellular ancestors: brown algae and animals.

Processes of genetic recombination are further split in two main categories. The first one is called homologous recombination, which is when DNA is exchanged between two corresponding loci on homologous chromosomes. This is a common source of genetic diversity that occurs in all Domains of life and is facilitated by a variety of mechanisms, including double-stranded DNA breaks repair (Figure 22), and chromosomal crossover between non-sister chromatids during meiosis in eukaryotes [Zickler and Kleckner 2015]. Prokaryotic organisms, which reproduce asexually, can particularly benefit from recombinations with homologous genes acquired horizontally to avoid the deleterious effects of Muller's ratchet [Vos 2009]. Homologous recombination thus contributes to shuffling genetic polymorphisms within populations but does not contribute to the creation of new gene forms. In this way, it is defined in opposition to non-homologous recombination, which encapsulates all other types of combinatorial processes of gene evolution.

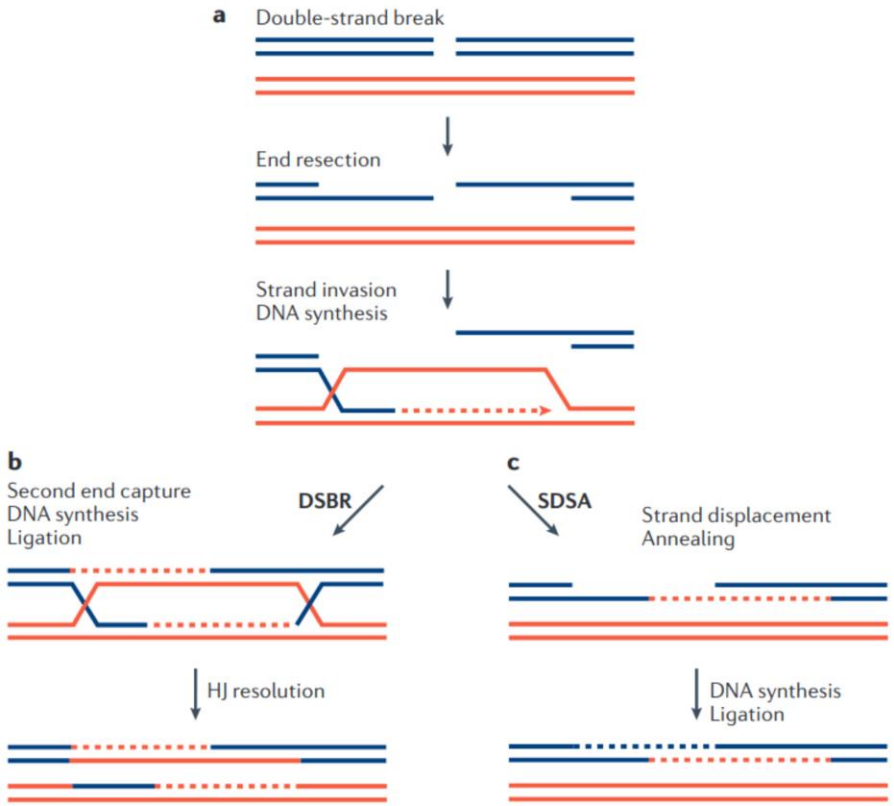


Figure 22: Mechanisms of homologous recombination following a double-stranded break.
A double-stranded break in DNA can be repaired with a homologous sequence (here in orange), for instance a homologous chromosome. Several mechanisms can fix these breaks, including double-stranded break repair (DSBR) by the formation and resolution of Holliday junctions (HJ), and synthesis-dependent strand annealing (SDSA). Once the double-stranded break is fixed, the repaired chromosome contains a portion of the homologous template.
Adapted from: [Sung and Klein 2006].

Non-homologous mechanisms of genetic recombination involve the movement of genetic material between non-homologous (in the stricter sense of homology) gene families. They are sometimes referred to as “illegitimate recombination”, a rather pejorative term that depicts these processes as undesirable and deleterious. While it is true that, locally, most recombination events have adverse effects on their hosts (as is the case for most mutations), many gene families have been created through combinatorial processes, contributing to important adaptations and transitions in the evolution of all organisms [Apic, Gough, and Teichmann 2001, Ekman et al. 2005]. Although various types of non-homologous recombination have been described, each facilitated by specific molecular mechanisms, we focused in particular on gene fusions and fissions, through which organisms can develop new proteins with innovative functions [Pasek, Risler, and Brézellec 2006, Dohmen et al. 2020, Padalko, Nair, and Sousa 2024].

1.2 – Gene fusions and gene fissions

Gene fusion is the process in which a novel gene sequence is created by the merging of two genes (or parts of genes) that previously existed as separate entities. The word "merging" here is to be understood in the sense of concatenation: the resulting fused gene[9] is composed of distinct regions that correspond to each "donor" gene sequence. A variety of mechanisms can give rise to gene fusions. The first one, which is common to all Domains of life, consists in the loss of the intergenic region between two adjacent open reading frames (ORFs), such that the transcription goes on uninterrupted between the two genes. This is the main way for prokaryotes, in particular, to obtain fused genes, whereas the intron-exon gene structure of eukaryotes allows for more modularity: gene fusions can for instance happen by the gain of an exon from a different gene [Marsh and Teichmann 2010]. Larger-scale rearrangements of chromosomes can also produce fusion genes, this time regardless of the relative distance between two genes along the genome[10].

The inverse process, wherein a single gene sequence becomes split in two new distinct genes, is coherently called gene fission. Likewise, gene fissions can arise from various different processes, both locally (e.g. the emergence of a new stop codon and intergenic region) and globally (e.g. genome rearrangements). Gene fissions are usually less frequent than fusions, although relative rates of fusion/fission events vary between different organisms [Kummerfeld and Teichmann 2005, Leonard and Richards 2012]. This can be explained, at least in part, by the relative complexity of evolving a novel intergenic sequence within an existing ORF (which requires the concurrent emergence of a new stop codon, promoter and start codon), compared to the disappearance of the intergenic space between two ORFs (feasible in a single event of stop codon loss or interstitial deletion). Split genes may also be more restricted than fused genes in their function, e.g. if both split genes resulting from a fission must be coexpressed and interact to fulfil the same role as the unit gene pre-fission, which may lead to a counter-selection of gene fissions relative to fusions.

[9] Some authors make a distinction between the fusion of whole genes, for which they reserve the term "fused gene", and the merging of only some parts of genes, called "chimeric genes". For our purposes, we choose to overlook this distinction, and we use the terms interchangeably.

[10] Chromosome rearrangements are often deleterious, and in humans are associated with many types of tumours and cancer [Mitelman, Johansson, and Mertens 2004]. The first detected instance of a fusion gene was actually found in cancer cells, and oncogenic gene fusions are often targeted during diagnostics of these pathologies. This can perhaps explain the use of "illegitimate recombination" in a human health, rather than evolution, context.

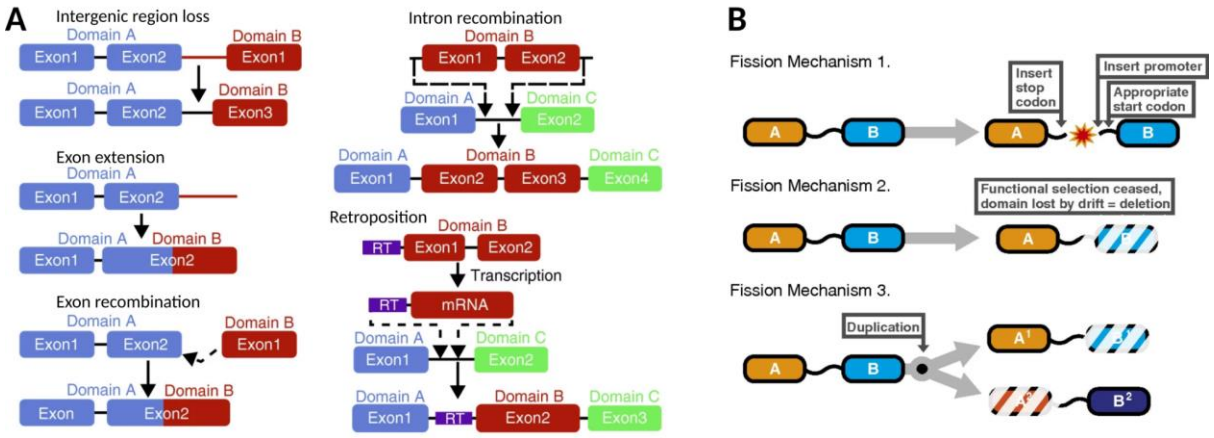


Figure 23: Gene fusions and fissions can occur through diverse mechanisms.
(**A**) The exon-intron structure of eukaryotic genes enables gene fusions via a number of processes. Adapted from [Marsh and Teichmann 2010].
(**B**) Three mutational processes that may result in the fission of a gene with two conserved regions (e.g. domains) A and B. From [Leonard and Richards 2012].

The genetic novelty created by gene fusions and fissions can generate new protein functions. In both cases, however, there exists a general trend to favour recombination between genes performing similar, or associated, functions [Yanai, Derti, and DeLisi 2001]. A number of reasons for this fact can be put forward, both upstream and downstream from natural selection. First, co-functioning genes have a higher chance of also being co-localised within genomes. This is especially true of prokaryotes, in which proteins that function together are frequently encoded by ORFs belonging to the same operon. In this context, local recombination events are more likely to involve genes with similar functions when they occur within the boundaries of an operon. From a structural perspective, furthermore, 'homofunctional' remodelling events may happen with less deleterious consequences. If two genes that code for proteins involved in a same complex are fused, for instance, the resulting fusion protein has a chance of adopting a structure that reflects the two ancestral proteins, which can reduce the risk of destabilising the complex and impeding its normal function. Such events can thus occur in a relatively transparent manner with respect to natural selection. Still, even remodelling events that involve functional partners can lead to variations in function.

1.3 – Protein domains: the Swiss Army knife of protein annotation

Protein domains indubitably constitute the best acknowledged framework for discussions of gene modularity and combinatorial evolution processes [Buljan and Bateman 2009]. Several databases coexist that map the diversity of these domains, all differing in how they recognise, classify and annotate these domains. Despite their differences, all of these databases provide an important ontological basis for the study of non-homologous recombination: an explicit collection of basic

building blocks for protein-coding genes that connect structural, functional and evolutionary aspects of the protein space. Protein domains are thus highly versatile descriptors of protein composition, because they represent good approximations of structural units, functional units and evolutionary units all at once, and they consequently lend themselves to a wide array of studies. Looking at proteins through the lens of domain architecture has produced numerous insights into the evolution of organisms throughout the tree of life. First and foremost is the prevalence of multi-domain proteins in all genomes, representing up to 80% and 60% of all eukaryotic and prokaryotic proteins respectively [Apic, Gough, and Teichmann 2001] – in other words, a majority of proteins derives, at least partly, from gene remodelling events. Domain rearrangements have been associated with some important evolutionary changes, including environmental adaptation in plants, multicellularity in animals, and eusociality in insects [Kersting et al. 2012, Cromar et al. 2014, Dohmen et al. 2020].

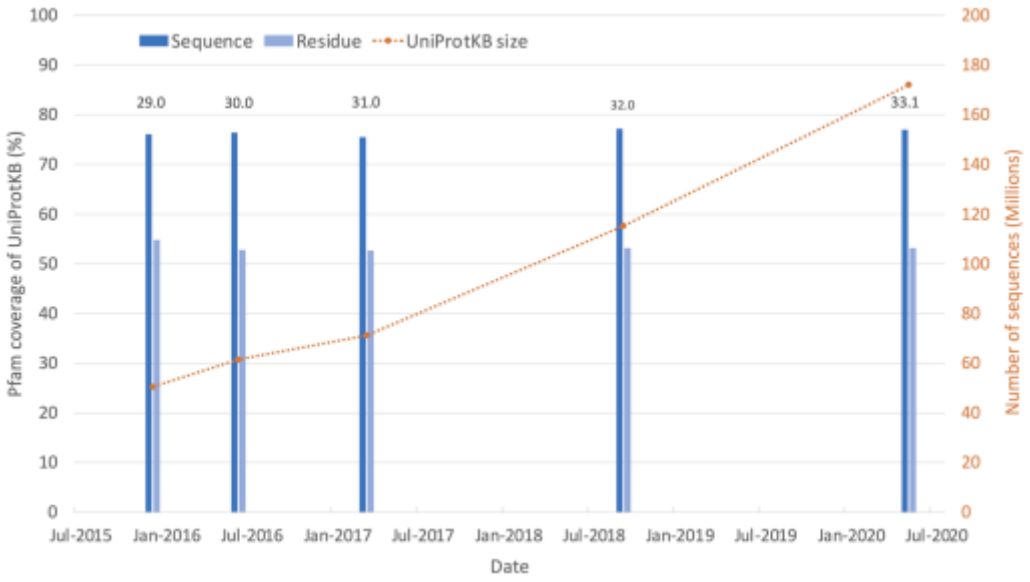


Figure 24: Coverage of UniProtKB by Pfam domains over the last five Pfam releases.
The number of sequences in UniProtKB has more than tripled between 2016 and 2021. In the same time, successive releases of Pfam have maintained approximately the same coverage of UniProtKB: ~77% of sequences have at least one Pfam annotation, and Pfam domains cover on average ~53% of a protein's length.
From: [Mistry et al. 2021].

Despite their undeniable value, however, protein domains fall short of representing the entire protein universe with perfect exhaustivity (Figure 24). Out of all the protein sequences in the comprehensive UniProtKB database, for instance, roughly 23% of them do not contain a single Pfam domain[11] [Mistry et al. 2021]. Furthermore, Pfam domains only cover 53% of a protein sequence's

[11] When also taking into account the 12 other domain databases hosted by InterPro, this figure only goes down to 18% [Paysan-Lafosse et al. 2023].

length on average, meaning that nearly half of the total known proteome cannot be currently described on the sole basis of protein domains. These figures have been remarkably stable over time, in line with the successive updates of both protein sequence and protein domain databases. This suggests that the non-domain-like portions of proteins are not unannotated because our sampling of domains is incomplete, but rather because domains are defined towards a specific archetype of protein sequence, which is in no way meant to be exhaustive. Intrinsically disordered regions, for instance, have no fixed 3D conformation, and although some annotations do exist for disordered domains, most fall outside the scope of what domains are meant to represent. Speaking more generally, domains describe the organisation of proteins rather than genes, and consequently, gene remodelling events do not necessarily operate strictly along the same lines. As an example, many protein domains roughly correspond to exons or groups of exons [Liu and Grigoriev 2004], but this view ignores that intronic sequences may be affected by remodelling too. A domain-centric approach to characterise gene fusions and gene fissions thus has some limitations, and cannot account for all the combinatorial processes that contribute to gene evolution.

2. Using similarity networks to identify remodelled genes

Although gene fusion and gene fission are mechanistically different, they still cannot be fully dissociated because in terms of gene evolution, they are structurally opposite processes: a fission event splitting a gene X into two genes X_1 and X_2 could hypothetically be 'reverted' by a fusion of X_1 and X_2 into X, and vice versa (Figure 25A). Gene fusions and gene fissions both involve a 'long form' gene which contains two distinct (i.e. non-overlapping) regions that also occur independently in separate 'split form' genes. Identifying these patterns can point to putative gene fusions or fissions, but it does not allow us to distinguish between the two, precisely because both can result in this same motif. An analysis of gene fusion and fission thus requires two distinct steps, to first identify putative events of fusion/fission, and then to classify each event into one of these two categories, for which other types of information about these genes must be leveraged. Before the fusion/fission decision is made, we will therefore adopt the terminology of composite and components [Enright et al. 1999]: the central 'long form' gene, which has partial homology to both the 'split form' genes, is called a composite gene, and the others are called components. This allows for a more neutral description of sets of genes that are involved in remodelling events, reflecting the *a priori* ignorance of which are fusions and which are fissions. We will also prefer "gene remodelling" over "recombination", as the latter has many different uses beyond our specific scope of fusion and fission events.

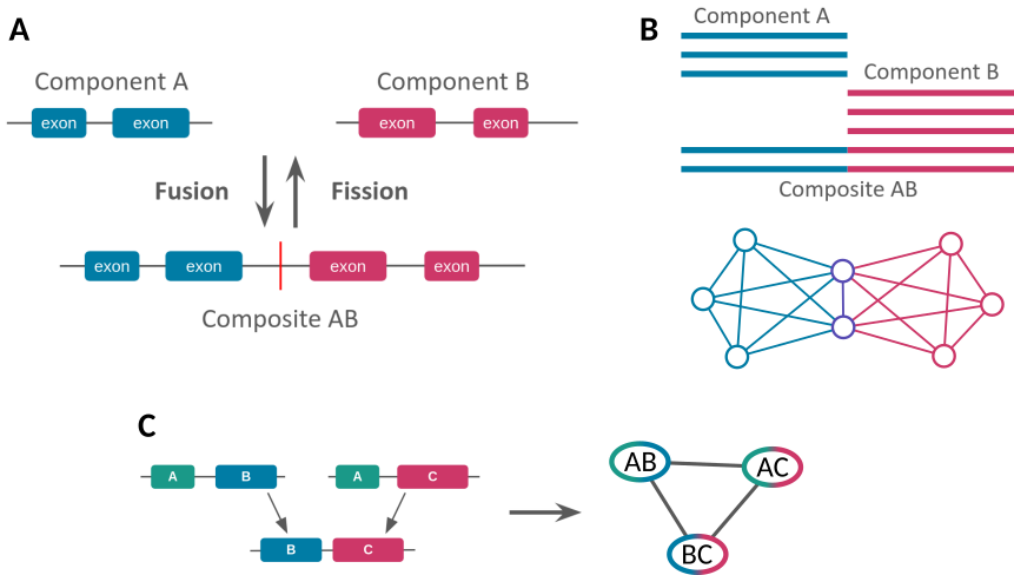


Figure 25: Gene fusions and fissions can result in the same sequence similarity patterns.
(**A**) The 'composite/component' terminology reflects the fact that gene fusion and gene fission are structurally inverse processes.
(**B**) Either of gene fusion and gene fission can produce an intransitive similarity pattern where the composite is similar to both components, which themselves are not similar to one another.
(**C**) In a fusion between two components that are already partly homologous, the resulting similarity pattern may not be an intransitive triplet.

2.1 – A first approach: intransitive homology triplets

Early computational methods for detecting fusion and fission events relied on non-transitive relationships of homology between genes [Jachiet et al. 2013]. The base assumption for this type of approach is that while a composite gene is homologous to each of its components, in general there should not be any homology between different components[12]. The composite gene therefore sits at the centre of an intransitive homology triplet, with components at each end (Figure 25B). This is an easy enough pattern to check for, giving a conceptually simple recipe for identifying putative remodelling events within a given pool of genes: enumerate all possible trios of genes, and retain those forming intransitive triplets. This approach has proven fruitful in past investigations of gene fusions, but it does suffer from a few shortcomings. First, the number of possible triplets to enumerate grows cubically in relation with the number of genes considered, which complicates the analysis of large datasets. Second, while many remodelling events do produce intransitive homology triplets, these patterns can also reflect other sequence relationships, including distant homologies – it is, after all, what we based SHIFT upon (see previous chapter). In practice, however, these anomalistic cases

[12] This is perhaps influenced by an understanding of gene fusion in the stricter sense, i.e. two whole genes being fused into one, as opposed to gene chimerism that merges subparts of genes instead (see section 1.2 of this chapter).

can be eliminated by checking whether components correspond to non-overlapping regions of the composite, which would not correspond to distant homologues. But conversely, intransitive triplets may not represent all fusion or fission events, especially if some genes have undergone several remodelling events in succession (Figure 25C). To detect remodelling events comprehensively, therefore, it might be necessary to take more general approaches.

2.2 – Detecting a broader spectrum of remodelled genes

As we discussed in the Introduction, a sequence similarity network can be viewed as a proxy for the homology relationships between a set of genes. To infer putative fusion/fission events from relationships of homology, therefore, one can build and analyse SSNs that reflect these homologies. However, unlike in the previous chapter where gene similarities were only considered when covering large parts of each sequence (typically >80%), here we must account for partial-length similarities as well. In a way, this amounts to creating SSNs with two different types of edges, in order to distinguish between full-length and partial alignments. Several methods based on SSNs for gene remodelling have been developed in our lab, including FusedTriplets, which implements a gene-based intransitive triplet search [Jachiet et al. 2013]; MosaicFinder, which transposes this idea at the level of gene families (using clique minimal separators to represent composite families) [Jachiet et al. 2013]; and CompositeSearch, which takes a slightly more general approach that does not rely explicitly on intransitive homology relationships [Pathmanathan et al. 2018]. This last programme is the one we used in particular to analyse the effects of gene remodelling events on the evolution of animals and of brown algae.

CompositeSearch relies on the *ab initio* constitution of gene families in the SSN, by clustering the network into modules using the Louvain community detection algorithm [Blondel et al. 2008]. At this step, only edges corresponding to full-length alignments are considered, so that the resulting clusters reflect coherent groups of full homology. Partial-length alignments are then factored in to detect putative composite genes, which is any gene with partial homologues in two different families that align on distinct regions of the composite sequence (in practice, small overlaps between those regions can be tolerated, to compensate for overextensions in BLAST alignments). Finally, once composite genes are called, the families that contain them are themselves reported as composite families. From a conceptual standpoint, this approach differs from those relying on intransitive triplets in that it centres directly around the intrinsic definition of a composite gene, rather than identifying connectivity patterns that are associated with this definition. In that regard, CompositeSearch allows for a more global description of gene remodelling dynamics than previous methods. The distinction between composites and components is also blurred, as gene families can now be simultaneously

composite and components, e.g. when a gene created by fusion is then involved in another fusion event.

2.3 – Fusion, fission, other? Polarisation of gene remodelling events

Despite its improvements over previous methods, CompositeSearch still cannot address the issue of classifying different types of remodelling events: composite families are reported as composites, but other steps are required to understand what this reflects from their evolutionary trajectory. To tackle this issue, we have developed a post-treatment method that works from the output of CompositeSearch to infer cases of gene fusion and fission within the set of reported composite families. This process of investigating remodelling events to decide in which direction the remodelling occurred (from composite to components, i.e. fission, or from components to composites, i.e. fusion) is sometimes called "polarising" the events, as it amounts to picking an orientation for the arrow of time.

The polarisation approach that we adopted uses evolutionary relationships between the host organisms of composite genes. Based on the phylogenetic tree of species that are represented in the dataset, the presence/absence of each gene family in extant species is used to infer the moment of emergence of that gene (Figure 26A), in accordance with the Dollo parsimony method [Rogozin et al. 2006]. In other words, a gene family present in the genomes of a number of species is considered to have originated no later than in the last common ancestor of those species. We can then apply a simple heuristic to label composite families as fused (i.e. originated in a fusion event) or split (underwent a fission event). If the components of a composite family existed prior to the composite's origin, then it is classified as a fusion event (Figure 26B); conversely, when a composite predates the emergence of its components, then it is considered as a gene fission (Figure 26C). Many intermediate cases also arise in practice, where the relative order of evolution between composite and components is not as clear-cut. A particularly frequent pattern consists of the composite point of origin being 'sandwiched' between a component that emerged earlier, and a second component that only evolved in a branch below the composite origin. In such cases, we reasoned that the composite must have originated by gene fusion, because at least one building block of its sequence was already present in its ancestral lineage; however, a single fusion event is insufficient to explain the seemingly later emergence of the other component, and subsequent events of gene fission or loss could have occurred to produce this phylogenetic distribution.

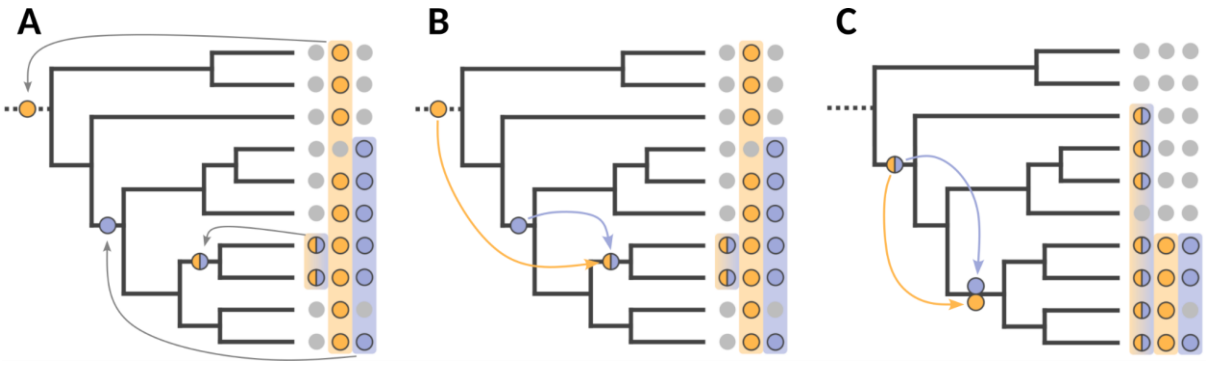


Figure 26: Polarisation of remodelling events from composite and component presence/absence.
(**A**) Using Dollo parsimony, the points of emergence of a composite family and its components are inferred on the species tree, according to the phyletic distribution of each family in extant genomes. Circles (plain yellow and purple for two components, and dual-tone for the corresponding composite) indicate the extant distribution (at the tips of the tree) and the points of origin (on internal branches) of each family.
(**B**) The order of emergence between components and composite is used to classify each remodelling event as fusion or fission. Here, the components appeared before the composite gene, suggesting a gene fusion event.
(**C**) In this other case, the composite gene predates the component forms, which is suggestive of gene fission.

This is a relatively simple model to infer fusion and fission events, and it relies on a few assumptions. First of all, it is highly reliant on the phylogenetic tree of species represented in the dataset and is therefore only well-suited for studying gene remodelling in lineages that evolve in a tree-like fashion. The lineages of eukaryotes that we applied this method to largely conform to this expectation, but alternative techniques may be preferred when working with bacterial genomes for instance, given their extensive use of horizontal gene transfer [Bapteste et al. 2009, Soucy, Huang, and Gogarten 2015]. Second, the Dollo parsimony model that is used to infer points of origin for each gene family is also based on a fairly restrictive set of hypotheses: each gene can only emerge once, and once lost it cannot be acquired again. This is only really appropriate when the gene families considered are orthogroups. This is another complication for using this method on prokaryotic genomes, again due to HGT which will make patterns of homoplasy emerge in presence/absence data. Even in the absence of HGT, inaccurate detection of orthologous gene families in the dataset can lead to erroneous results. Moreover, there is no reason why some remodelling events could not happen convergently in unrelated lineages, as indeed we observed in metazoans. In both the research projects for which we conducted analyses of remodelling events, we tried to account for this limitation, albeit in different ways. In the brown algae study, another team had already performed a detection of orthologous gene families in the genome dataset, and we therefore used their results to define the gene families in the SSN, instead of performing Louvain clustering. In the case of metazoans, on the other hand, the families defined by CompositeSearch (based on SSN topology) were used for the composite

annotation, but then those families were checked to correspond to mono- or polyphyletic groups in the species phylogeny, prior to the polarisation step which treated sub-families of polyphyletic clusters independently. With those caveats in mind, we found that our method was able to label most of the composite families as either fusions or fissions, although some problematic cases persisted, e.g. in cases where the origins of a composite and its components were inferred in unrelated branches of the species phylogeny. Such cases could not be assigned to gene fusion or fission with our methods, and were considered undecided.

3. Important role of remodelled genes in the early evolution of brown algae

We conducted a gene remodelling analysis as part of our contribution to a broad study on the biology and evolution of brown algae. This study was the fruit of the Phaeoexplorer project, a research consortium involving more than 100 collaborators that produced and analysed a large resource of novel genomes for this lineage. An article presenting the main outcomes of this project has been submitted and accepted for publication in Cell, and is reproduced below. Our specific analysis of remodelled genes is featured in this article, but it only represents a small part of all the work done with these new genomes, and thus only a small part of the article is centred specifically on our results. We therefore describe and discuss our contribution in more detail in this present section, before the reproduction of the article as a whole.

Along with some green and red algae, brown seaweeds are one of the three types of macroscopic algae populating coastal seawaters on Earth. They constitute an abundant and central component of those ecosystems, as exemplified by the forests of kelp that serve as habitats and food sources for many marine species. Brown algae (*Phaeophyceae*) form a class within the larger group of photosynthetic stramenopiles that acquired a chloroplast following a secondary endosymbiosis with a red alga [Keeling 2009], and they emerged around 450 million years ago during the Great Ordovician Biodiversification Event (GOBE) [Choi et al. 2024]. They are notable for their acquisition of complex multicellularity, in contrast with sister clades that are either unicellular or form simple multicellular filaments[13]. Over the course of their evolution, brown algae have developed a broad diversity of cell

[13] "Simple" multicellularity, or pluricellularity, typically consists of intercellular aggregations in "one dimension" (filaments) or two (biofilms), where most cells keep a direct interface with the environment, have limited exchanges with their neighbours and do not differentiate into specialised cell types. Complex multicellular organisms, on the other hand, develop different tissue types with high levels of cell-cell communication and gene regulation, and display three-dimensional organisations requiring complex systems of biomolecule transportation [Knoll 2011].

cycles and morphological plans, in adaptation to diverse coastal ecosystems, including fully submerged or intertidal seawater, brackish waters, and on a few occasions freshwater [Dittami et al. 2017]. Interestingly, the GOBE has previously been cited as having created favourable conditions for the evolution of multicellularity in algae, due to the emergence of marine herbivores that grazed on algae [LoDuca et al. 2017]: in that context, growing in size and developing different tissues could help reducing the detrimental effect of herbivores on the alga's survival, for instance by directing the grazing action toward leaf-like structures[14] that are relatively easy to regrow.

The Phaeoexplorer group produced 60 genomes of brown algae and closely related species, covering all the major orders of brown algae. Among these genomes, 17 were acquired from long-read sequencing, and were part of a high-quality subset of 21 genomes (with the inclusion of four quality genomes already published) on which most analyses were focused. The general trend that was observed from analysing those genomes consisted in a marked gain of new gene families and functions early in the evolution of brown algae, contributing to the development of novel metabolic pathways central to the transition to complex multicellularity. Gene loss and gene family amplification (the increase in copy number of a gene within a genome), on the other hand, were much more prevalent later in the diversification of the lineage, and supposedly drove the emergence of a diverse range of morphological and physiological phenotypes.

[14] Called *blades*, *laminae* or *fronds* based on their morphology.

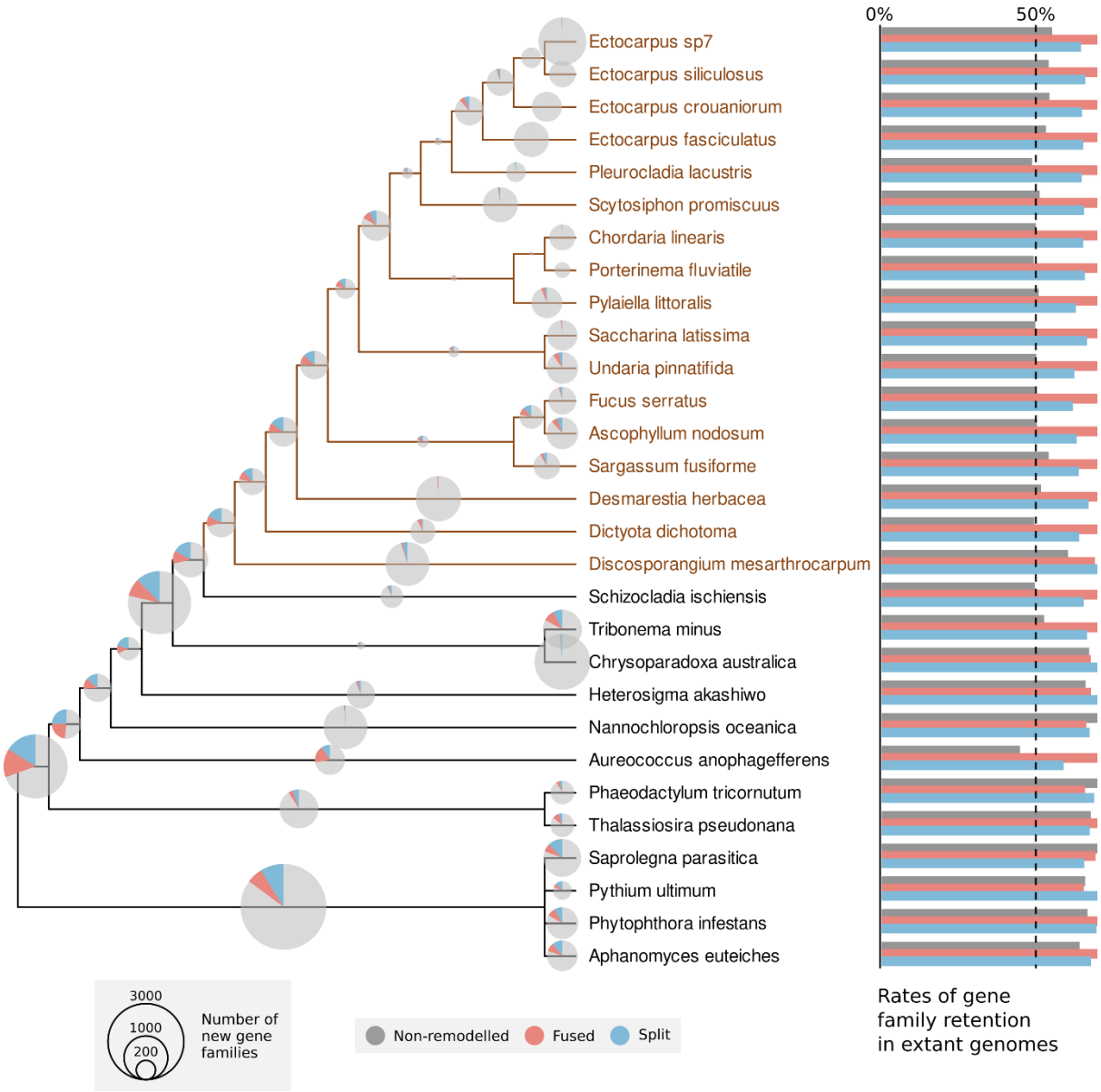


Figure 27: Emergence and retention of fused and split genes in the evolution of brown algae. Phylogenetic distribution of fused (red), split (blue) and non-remodelled (grey) gene family originations across the evolution of brown algae. Pie charts on each branch of the phylogeny indicate the relative contribution of gene fusion and fission to the overall emergence of novel gene families, quantified by the area of the circle. Bars on the right indicate the percentage of gene families retained in extant genomes among all gene families that emerged during the evolution of our species set. Brown algae species are indicated in brown, and other stramenopiles in black. Note that only the topology of the species tree is displayed here, without specific branch lengths.

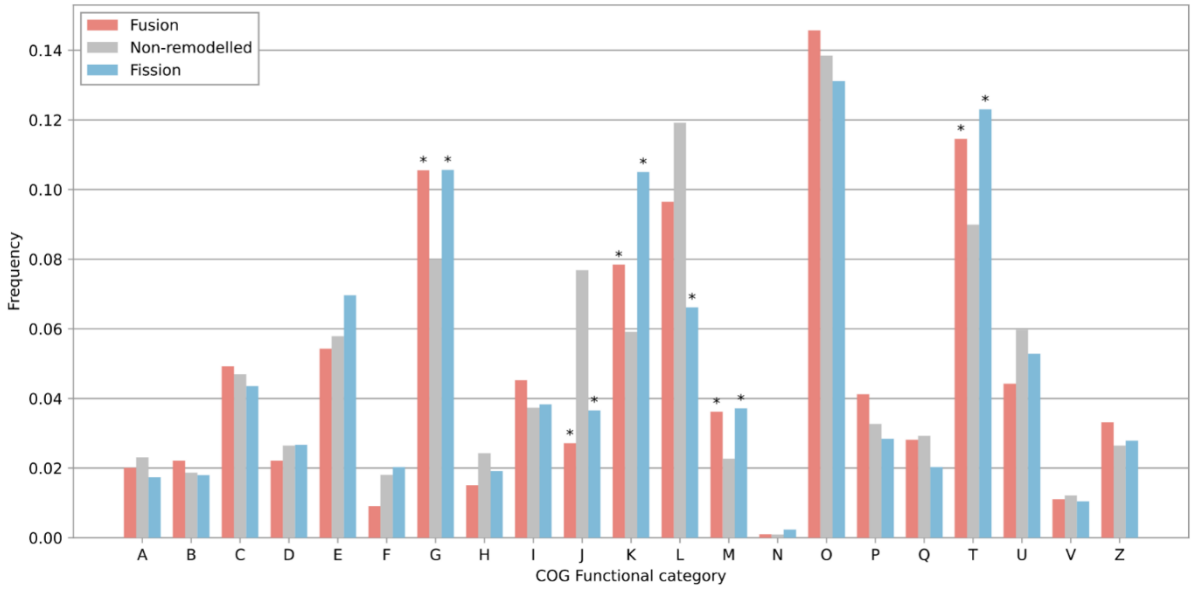


Figure 28: Functional enrichment of certain COG categories in Phaeophyceae remodelled genes. Distribution of gene families in COG functional categories for fused, split, and non-remodelled orthologous groups. Orthogroups with no annotation or annotated as 'S (unknown function)' were discarded, so that only orthogroups with known functions were taken into account. Asterisks above bars indicate a statistically significant divergence from non-remodelled gene families (p-value < 0.05, two-sided Chi-squared test with Yates correction).

Our gene remodelling analysis also focused on the high-quality subset of 21 genomes, because misassemblies of reads in low quality genomes can result in artefacts such as gene chimeras that could be picked up by CompositeSearch as composite genes. To produce results that were compatible and interoperable with those of other working groups, we based our search on orthologous gene families (orthogroups) that had previously been defined from those genomes. We found that 12.6% of all orthogroups (excluding singletons) were potential composites. In particular, 6.7% of orthogroups were formed in gene fusion events, whereas 4.8% had undergone fission events[15]. We thus observed more gene fusions than fissions, in line with most of the previous literature on fusion/fission events in other lineages, but the disparity between the two was markedly less than the four-to-one ratio that is generally documented [Kummerfeld and Teichmann 2005]. As with other gene family gains, the majority of fusion and fission events occurred in the early stages of Phaeophyceae evolution, and were less frequent in more recent branches of the phylogeny (Figure 27). An analysis of retention rates showed that the gene families created in remodelling events were less frequently lost than non-remodelled genes in extant genomes and, importantly, this preferential retention was much more

[15] Of the remaining 1.1% of orthogroups that were detected as composites, 0.7% could not be called because composites and components all originated before the emergence of brown algae, and 0.4% had an unresolved polarisation due to incongruent phyletic patterns between composites and components.

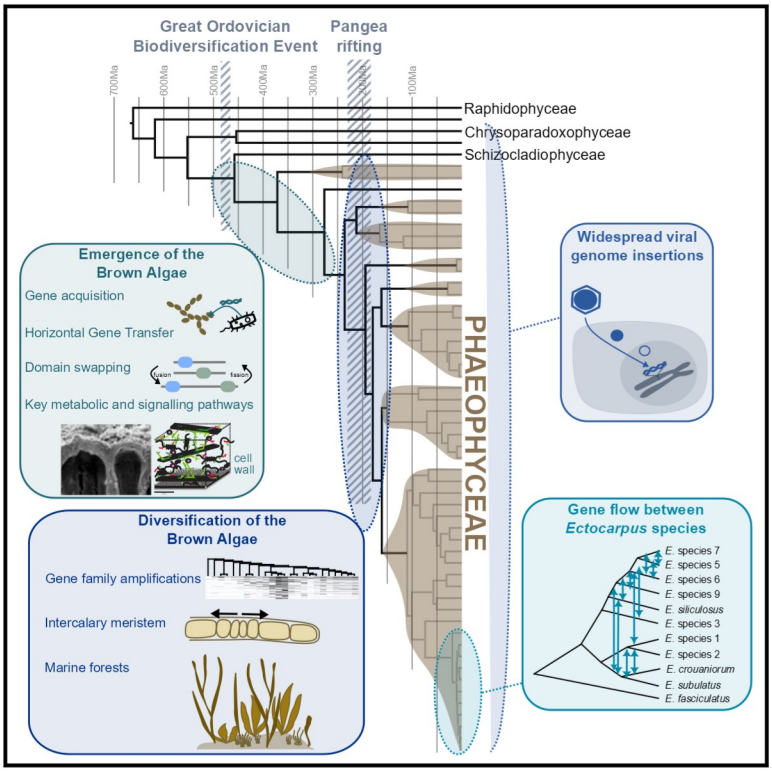
pronounced in brown algae than in other species of stramenopiles (Figure 27). These results suggest that gene fusions and gene fissions played an important role in the initial emergence of brown algae, and that they mainly contributed to processes that were critical for this lineage, such that their loss was markedly selected against. Indeed, orthogroups that resulted from remodelling events were enriched in a few specific functional categories (Figure 28), particularly related to cell wall and signalling (COG categories G - Carbohydrate transport and metabolism, M - cell wall biogenesis, T - signal transduction) and transcription (COG category K). These functional classes may have contributed to the development of complex phenotypes, including the brown algal cell wall and extracellular matrix (ECM) that are both composed of algin, a polysaccharide specific to members of the Phaeophyceae class [Mazéas et al. 2023]. Remodelled genes in those functional categories may have also facilitated the emergence of complex multicellularity, thanks to innovations in signalling pathways and cell-cell communication.

In summary, our results indicate that gene remodelling events played a substantial role, along with other routes of novel gene family foundation, in the emergence and the early evolution of brown algae from unicellular stramenopiles. Gene fusion and fission may have contributed, in particular, to functions that helped the onset of complex multicellularity, as well as metabolic adaptations to intertidal and subtidal ecosystems (e.g. thanks to the flexibility provided by alginate-based cell walls and ECMs, which helps resist the push-pull forces of seawater movement). In the later stages of Phaeophyceae diversification, remodelling events seemingly became less frequent (or less frequently fixed), but early remodelled genes were preferentially conserved, which suggests that they may be of particular importance for the success of their hosts. Other collaborators in the Phaeoexplorer consortium have noted that brown algal genes are also more intron-rich than genes of other stramenopiles, due to a rapid period of intron acquisition just before Phaeophyceae diverged from their closest sister class, Schizocladiophyceae. Interestingly, this increase of intron content in the genes of the common ancestor of brown algae may have set the stage for the subsequent wave of gene remodelling events that contributed to the emergence and the diversification of Phaeophyceae, acting as a sort of precursor event to the development of increasing biological complexity.

Cell

Evolutionary genomics of the emergence of brown algae as key components of coastal ecosystems

Graphical abstract



Highlights

- An intense period of genome evolution during early emergence of the brown algae
- Gene family amplifications linked to diversification of the brown algae
- Extensive gene flow between species at the genus level in *Ectocarpus*
- Insertions of diverse *Phaeovirus* genomes are widespread in brown algae

Authors

France Denoeud, Olivier Godfroy, Corinne Cruaud, ..., Patrick Wincker, Jean-Marc Aury, J. Mark Cock

Correspondence

kawai@kobe-u.ac.jp (H.K.), akirapeters@gmail.com (A.F.P.), hsyoon2011@skku.edu (H.S.Y.), cherve@sb-roscoff.fr (C.H.), yenh@ysfri.ac.cn (N.Y.), epbapteste@gmail.com (E.B.), valero@sb-roscoff.fr (M.V.), gabriel.markov@sb-roscoff.fr (G.V.M.), corre@sb-roscoff.fr (E.C.), susana.coelho@tuebingen.mpg.de (S.M.C.), pwincker@genoscope.cns.fr (P.W.), jmaury@genoscope.cns.fr (J.-M.A.), cock@sb-roscoff.fr (J.M.C.)

In brief

Comparative genomics charts the evolutionary history of the brown algal lineage, identifying an early period of accelerated genome evolution followed by diversification of the major orders, and a major impact of continuous, widespread viral genome integration.

Denoeud et al., 2024, Cell *187*, 6943–6965
November 27, 2024
https://doi.org/10.1016/j.cell.2024.10.049

CellPress

Cell

CellPress
OPEN ACCESS

Article

Evolutionary genomics of the emergence of brown algae as key components of coastal ecosystems

France Denoeud,[1,60] Olivier Godfroy,[2,60] Corinne Cruaud,[3,61] Svenja Heesch,[4,51,61] Zofia Nehr,[4,61] Nachida Tadrent,[1,52,61] Arnaud Couloux,[1,61] Loraine Brillet-Guéguen,[5,6,61] Ludovic Delage,[7,61] Dean Mckeown,[6,61] Taizo Motomura,[8,62] Duncan Sussfeld,[1,9,62] Xiao Fan,[10,11,62] Lisa Mazéas,[2,62] Nicolas Terrapon,[12,13,62] Josué Barrera-Redondo,[14,62] Romy Petroll,[14,62] Lauric Reynes,[15,62] Seok-Wan Choi,[16,62] Jihoon Jo,[16,62] Kavitha Uthanumallian,[17,62] Kenny Bogaert,[18,53,62] Céline Duc,[19,62] Pélagie Ratchinski,[4,62] Agnieszka Lipinska,[4,14,62] Benjamin Noel,[1,62] Eleanor A. Murphy,[20,21,62] Martin Lohr,[22,62] Ananya Khatei,[23,54,62] Pauline Hamon-Giraud,[24,62] Christophe Vieira,[25,62] Komlan Avia,[26,62] Svea Sanja Akerfors,[22] Shingo Akita,[27] Yacine Badis,[4] Tristan Barbeyron,[2] Arnaud Belcour,[24,55] Wahiba Berrabah,[1] Samuel Blanquart,[24] Ahlem Bouguerba-Collin,[2] Trevor Bringloe,[28] Rose Ann Cattolico,[29] Alexandre Cormier,[30] Helena Cruz de Carvalho,[31,32] Romain Dallet,[6] Olivier De Clerck,[18]

(Author list continued on next page)

[1]Génomique Métabolique, Genoscope, Institut François Jacob, CEA, CNRS, Université Evry, Université Paris-Saclay, Evry 91057, France
[2]Sorbonne Université, CNRS, Integrative Biology of Marine Models Laboratory, Station Biologique de Roscoff, Roscoff, France
[3]Genoscope, Institut François Jacob, CEA, CNRS, Univ Evry, Université Paris-Saclay, Evry 91057, France
[4]Sorbonne Université, CNRS, Algal Genetics Group, Integrative Biology of Marine Models Laboratory, Station Biologique de Roscoff, Roscoff, France
[5]CNRS, UMR 8227, Laboratory of Integrative Biology of Marine Models, Sorbonne Université, Station Biologique de Roscoff, Roscoff, France
[6]CNRS, Sorbonne Université, FR2424, ABiMS-IFB, Station Biologique, Roscoff, France
[7]Sorbonne Université, CNRS, UMR 8227, ABIE Team, Integrative Biology of Marine Models Laboratory, Station Biologique de Roscoff, Roscoff, France
[8]Muroran Marine Station, Hokkaido University, Muroran, Japan
[9]Institut de Systématique, Evolution, Biodiversité (ISYEB), UMR 7205, Sorbonne Université, CNRS, Museum, Paris, France
[10]State Key Laboratory of Mariculture Biobreeding and Sustainable Goods, Yellow Sea Fisheries Research Institute, Chinese Academy of Fishery Sciences, Qingdao, Shandong 266071, China
[11]Laboratory for Marine Fisheries Science and Food Production Processes, Laoshan Laboratory, Qingdao, Shandong 266237, China
[12]Aix Marseille University, CNRS, UMR 7257 AFMB, Marseille, France
[13]INRAE, USC 1408 AFMB, Marseille, France
[14]Department of Algal Development and Evolution, Max Planck Institute for Biology, Max-Planck-Ring 5, 72076 Tübingen, Germany
[15]IRL 3614, UMR 7144, DISEEM, CNRS, Sorbonne Université, Station Biologique de Roscoff, Roscoff 29688, France
[16]Department of Biological Sciences, Sungkyunkwan University, Suwon 16419, Republic of Korea
[17]University of Melbourne, Parkville, VIC, Australia
[18]Phycology Research Group, Ghent University, Krijgslaan 281 S8, 9000 Ghent, Belgium
[19]Nantes Université, CNRS, US2B, UMR 6286, 44000 Nantes, France
[20]University of Bristol, Bristol, UK
[21]Marine Biological Association, Plymouth, UK
[22]Johannes Gutenberg University, Mainz, Germany
[23]Algal and Microbial Biotechnology Division, Nord University, Bodø, Norway

(Affiliations continued on next page)

SUMMARY

Brown seaweeds are keystone species of coastal ecosystems, often forming extensive underwater forests, and are under considerable threat from climate change. In this study, analysis of multiple genomes has provided insights across the entire evolutionary history of this lineage, from initial emergence, through later diversification of the brown algal orders, down to microevolutionary events at the genus level. Emergence of the brown algal lineage was associated with a marked gain of new orthologous gene families, enhanced protein domain rearrangement, increased horizontal gene transfer events, and the acquisition of novel signaling molecules and key metabolic pathways, the latter notably related to biosynthesis of the alginate-based extracellular matrix, and halogen and phlorotannin biosynthesis. We show that brown algal genome diversification is tightly linked to phenotypic divergence, including changes in life cycle strategy and zoid flagellar structure. The study also showed that integration of large viral genomes has had a significant impact on brown algal genome content throughout the emergence of the lineage.

Ahmed Debit,[31] Erwan Denis,[1] Christophe Destombe,[15] Erica Dinatale,[14] Simon Dittami,[7] Elodie Drula,[12,13] Sylvain Faugeron,[33] Jeanne Got,[24] Louis Graf,[16] Agnès Groisillier,[19] Marie-Laure Guillemin,[15,34,35] Lars Harms,[36] William John Hatchett,[37] Bernard Henrissat,[38] Galice Hoarau,[37] Chloé Jollivet,[2] Alexander Jueterbock,[23] Ehsan Kayal,[6,56] Andrew H. Knoll,[39] Kazuhiro Kogame,[40] Arthur Le Bars,[6,41] Catherine Leblanc,[7] Line Le Gall,[9] Ronja Ley,[22] Xi Liu,[6] Steven T. LoDuca,[42] Pascal Jean Lopez,[43] Philippe Lopez,[9] Eric Manirakiza,[19] Karine Massau,[6] Stéphane Mauger,[15,57] Laetitia Mest,[4,58] Gurvan Michel,[2] Catia Monteiro,[7] Chikako Nagasato,[8] Delphine Nègre,[6,59] Eric Pelletier,[1] Naomi Phillips,[44] Philippe Potin,[7] Stefan A. Rensing,[45] Ellyn Rousselot,[19] Sylvie Rousvoal,[7] Declan Schroeder,[46] Delphine Scornet,[4] Anne Siegel,[24] Leila Tirichine,[19] Thierry Tonon,[47] Klaus Valentin,[36] Heroen Verbruggen,[28] Florian Weinberger,[48] Glen Wheeler,[21] Hiroshi Kawai,[49,63,*] Akira F. Peters,[50,63,*] Hwan Su Yoon,[16,63,*] Cécile Hervé,[2,63,*] Naihao Ye,[10,11,63,*] Eric Bapteste,[9,63,*] Myriam Valero,[15,63,*] Gabriel V. Markov,[7,63,*] Erwan Corre,[6,63,*] Susana M. Coelho,[14,63,*] Patrick Wincker,[1,63,*] Jean-Marc Aury,[1,63,*] and J. Mark Cock[4,64,*]

[24]University of Rennes, Inria, CNRS, IRISA, Equipe Dyliss, Rennes, France
[25]Research Institute for Basic Sciences, Jeju National University, Jeju 63243, Republic of Korea
[26]INRAE, Université de Strasbourg, UMR SVQV, 68000 Colmar, France
[27]Faculty of Fisheries Sciences, Hokkaido University, Minato-cho 3-1-1, Hakodate, Hokkaido 041-8611, Japan
[28]University of Melbourne, Parkville, VIC, Australia
[29]University of Washington, Seattle, WA, USA
[30]Ifremer, IRSI, SeBiMER Service de Bioinformatique de l'Ifremer, 29280 Plouzané, France
[31]Institut de Biologie de l'ENS (IBENS), Département de Biologie, École normale supérieure, CNRS, INSERM, Université PSL, 75005 Paris, France
[32]Université Paris Est-Créteil (UPEC), Faculté des Sciences et Technologie, 61, Avenue du Général De Gaulle, 94000 Créteil, France
[33]Facultad de Ciencias Biológicas, Pontificia Universidad Católica de Chile, Santiago, Chile
[34]Núcleo Milenio MASH, Instituto de Ciencias Ambientales y Evolutivas, Facultad de Ciencias, Universidad Austral de Chile, Valdivia, Chile
[35]Centro FONDAP de Investigación en Dinámica de Ecosistemas Marinos de Altas Latitudes (IDEAL), Valdivia, Chile
[36]Alfred Wegener Institute (AWI), Bremenhaven, Germany
[37]Nord University, Bodø, Norway
[38]Department of Biotechnology and Biomedicine, Technical University of Denmark, Kgs Lyngby, Denmark
[39]Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, MA 02138, USA
[40]Biological Sciences, Faculty of Science, Hokkaido University, Sapporo 060-0810, Japan
[41]CNRS, Institut Français de Bioinformatique, IFB-core, Évry, France
[42]Department of Geography and Geology, Eastern Michigan University, Ypsilanti, MI 48197, USA
[43]Centre National de la Recherche Scientifique, UMR BOREA MNHN/CNRS-8067/SU/IRD/Université de Caen Normandie/Université des Antilles, Plouzané, France
[44]Biology Department, Arcadia University, Glenside, PA, USA
[45]University of Freiburg, Freiburg im Breisgau, Germany
[46]University of Minnesota, St. Paul, MN, USA
[47]Centre for Novel Agricultural Products (CNAP), Department of Biology, University of York, Heslington, York YO10 5DD, UK
[48]GEOMAR Helmholtz Centre for Ocean Research, Kiel, Germany
[49]Kobe University Research Center for Inland Seas, Kobe, Japan
[50]Bezhin Rosko, 29250 Santec, France
[51]Present address: Applied Ecology & Phycology, Institute for Biosciences, University of Rostock, Albert-Einstein-Strasse 3, 18059 Rostock, Germany
[52]Present address: Institut de Recherche sur la Biologie de l'Insecte, UMR 7261, CNRS, Université de Tours, Tours 37200, France
[53]Present address: Department of Algal Development and Evolution, Max Planck Institute for Biology, Tübingen 72076, Germany
[54]Present address: ICAR-Directorate of Coldwater Fisheries Research, Bhimtal, India
[55]Present address: Univ. Grenoble Alpes, Inria, 38000 Grenoble, France
[56]Present address: Department of Ecology, Evolution, and Organismal Biology, Iowa State University, Ames, IA, USA
[57]Present address: CNRS, La Rochelle Université, UMR7266, Littoral Environnement et Sociétés, La Rochelle, France

(Affiliations continued on next page)

INTRODUCTION

The brown algae (Phaeophyceae) are a lineage of complex multicellular organisms that emerged about 450 mya[1] from within a group of photosynthetic stramenopile taxa (derived from a secondary endosymbiosis involving a red alga[2]) that are either unicellular or have very simple filamentous multicellular thalli (Figure 1). The emerging brown algae acquired a number of characteristic features that are thought to have contributed to the evolutionary success of this lineage, including complex polysaccharide-based cell walls that confer protection and flexibility in the highly dynamic intertidal environment,[3] complex halogen[4] and phlorotannin[5] metabolisms that are thought to play important roles in multiple processes including defense, adhesion and cell-wall modification, and a remarkable diversity of life cycles and developmental body architectures adapted to diverse marine environments.[6] As a result of these attributes, many brown algae have become

[58]Present address: Vegenov, Saint Pol de Léon, France
[59]Present address: Nantes Université, Institut des Substances et Organismes de la Mer, ISOMer, UR 2160, Nantes, France
[60]These authors contributed equally
[61]These authors contributed equally
[62]These authors contributed equally
[63]Senior author
[64]Lead contact
*Correspondence: kawai@kobe-u.ac.jp (H.K.), akirapeters@gmail.com (A.F.P.), hsyoon2011@skku.edu (H.S.Y.), cherve@sb-roscoff.fr (C.H.), yenh@ysfri.ac.cn (N.Y.), epbapteste@gmail.com (E.B.), valero@sb-roscoff.fr (M.V.), gabriel.markov@sb-roscoff.fr (G.V.M.), corre@sb-roscoff.fr (E.C.), susana.coelho@tuebingen.mpg.de (S.M.C.), pwincker@genoscope.cns.fr (P.W.), jmaury@genoscope.cns.fr (J.-M.A.), cock@sb-roscoff.fr (J.M.C.)
https://doi.org/10.1016/j.cell.2024.10.049

established as key components of extensive coastal ecosystems. These seaweed-based ecosystems provide high value Earth-system-scale services, including the sequestration of several megatons of carbon per year globally, comparable to values reported for terrestrial forests,[7] but this important role of seaweed ecosystems is threatened by climate-related declines in seaweed populations worldwide.[8] However, appropriate conservation measures, coupled with the development of seaweed mariculture as a highly sustainable and low impact approach to food and biomass production, could potentially reverse this trend, allowing seaweeds to play a significant role in mitigating the effects of climate change.[9] To attain this objective, it will be necessary to address important gaps in our knowledge of the biology and evolutionary history of the brown algal lineage. For example, these seaweeds remain poorly described in terms of genome sequencing due, in part, to difficulties with extracting nucleic acids. The Phaeoexplorer project (https://phaeoexplorer.sb-roscoff.fr/) has generated a large dataset of genome sequences, spanning all the major orders of the Phaeophyceae.[10] This extensive genomic dataset has been analyzed here to study the origin and evolution of key genomic features during the emergence and diversification of this important group of marine organisms.

RESULTS

In-depth sequencing of brown algal genomes

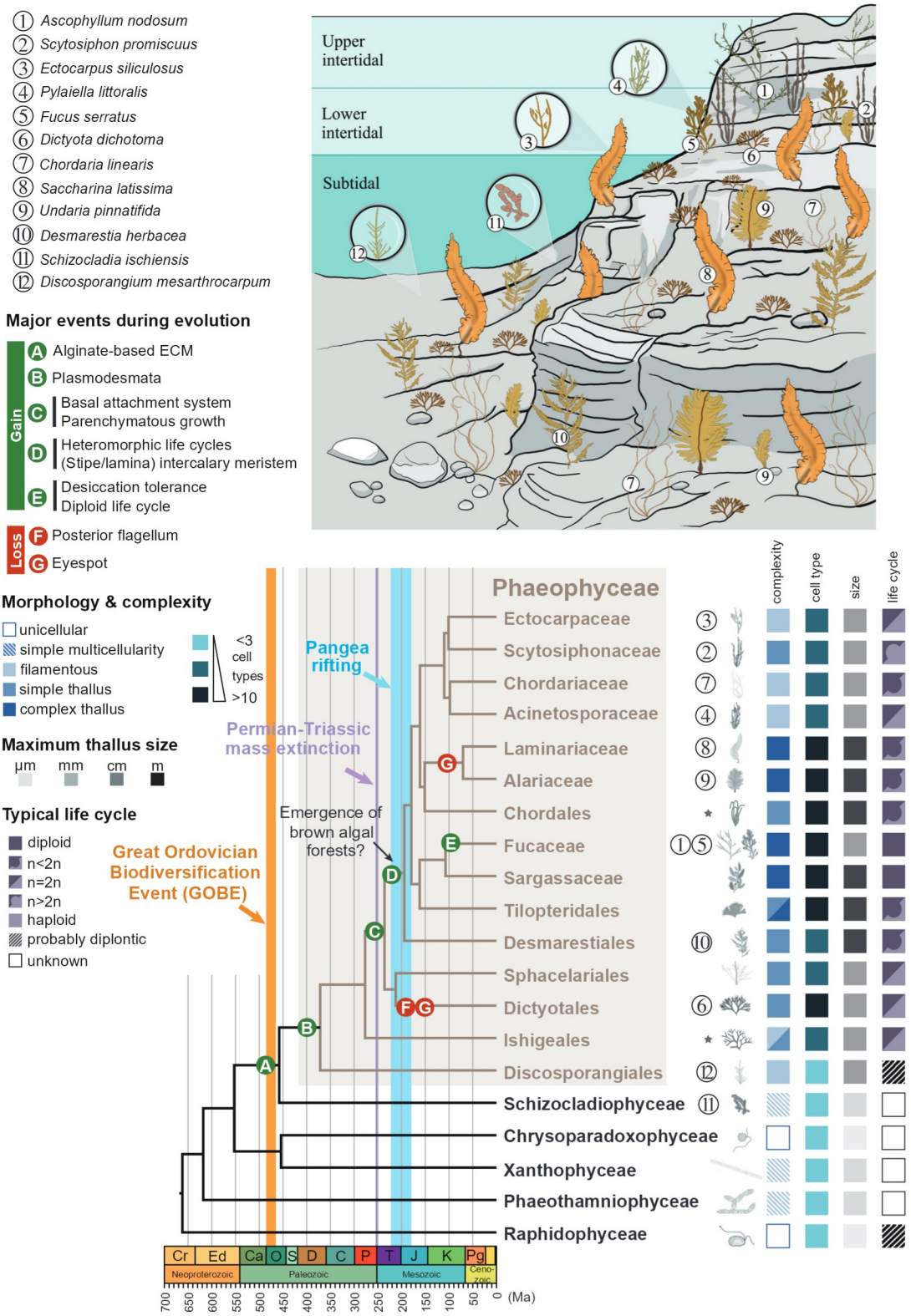
Until now, good quality genome assemblies have been obtained for only five brown algal species,[11–15] together with about 46 draft genome assemblies.[16–20] Here, we report work that has significantly expanded the genomic data available by sequencing and assembling 17 good quality genomes using long-read technology (Table S1), plus an additional 43 draft genome assemblies. These 60 genomes correspond to 40 brown algae and four closely related species, covering 16 Phaeophyceae families providing a dense coverage of this lineage (Figures S1A and S1B; Table S1A). The sequenced species include brown algae that occur at different levels of the intertidal and subtidal and are representative of the broad diversity of this group of seaweeds in terms of size, levels of multicellular complexity, biogeography and life cycle structure (Figures 1, S1C, and S1D). The analyses carried out in this study have focused principally on a set of 21 good quality reference genomes, which include four previously published genomes (Table S1B).

Marked changes in genome content and gene structure during the emergence of the Phaeophyceae lineage

Recent evidence indicates that the brown algae emerged about 450 mya during the Great Ordovician Biodiversification Event (GOBE),[1] a conclusion that is supported by a fossil-calibrated tree built with a nuclear-gene-based phylogeny constructed using the Phaeoexplorer genome data (Figures 1 and S2A). An increase in atmospheric oxygen at the time of the GOBE, which coincided with the emergence of herbivorous marine invertebrates,[21] is likely to have created conditions conducive to the observed transition toward increased multicellular complexity during early brown algal evolution.

To investigate genomic modifications associated with the emergence and diversification of the brown algae, we first carried out a series of genome-wide analyses aimed at identifying broad trends in genome evolution over evolutionary time (Figure 2). Dollo analysis of gain and loss of orthogroups (i.e., gene families) indicated marked gains during early brown algal evolution followed by a broad tendency to lose orthogroups later as the different brown algal orders diversified (Figures 2B and S2). Similarly, a phylostratigraphy analysis indicated that 29.6% of brown algal genes cannot be traced back to ancestors outside the Phaeophyceae, with the majority of gene founder events occurring early during the emergence of the brown algae (Figures 2E and S3A; Table S2), again indicating a burst of gene birth during the emergence of this lineage. Both the Dollo analysis and the phylostratigraphy approach indicated that the gene families acquired during early brown algal evolution were significantly enriched in genes that could not be assigned to a cluster of orthologous genes (COG) category, suggesting a burst in the acquisition of genetic novelty (Figure 2G).

One of the factors underlying the marked burst of gene gain during the emergence of the brown algae was an increase in the rate of acquisition of new genes via horizontal gene transfer (HGT). A phylogeny-based search for genes potentially derived from HGT events indicated that they constitute about 1% of brown algal gene catalogs and that the novel genes were principally acquired from bacterial genomes (Figures 2F and S3B). The proportion of class-specific HGT events compared with more ancient HGT events was greater for the brown algae (33.5% of HGT events) than for the closely related taxa Xanthophyceae (*Tribonema minus*) and Raphidophyceae (*Heterosigma akashiwo*; mean of 17.1% for the two taxa, Wilcoxon $p = 0.021$), indicating that higher levels of HGT occurred during the emergence of the brown algae than in closely related taxa (Figure 2F).



(legend on next page)

The marked increase in the rate of gene gain appears to have been a key factor in the emergence of the brown algal lineage but this was not the only process that enriched brown algal genomes during this period. Domain fusions and fissions (composite genes) were prevalent during the early stages of brown algal emergence (Figures 2D and S3C), affecting about 7% of brown algal gene complements. In contrast, gene family amplifications were most prevalent at a later stage of brown algal evolution, corresponding to the diversification of the major brown algal orders during the Mesozoic (Figures 2C, S3D, and S3E; Table S3). However, the amplified gene families were significantly enriched in genes that had been gained during the emergence of the brown algae (χ^2 p = 1.04e−15; Table S1C), indicating that gene gain during the early evolution of the lineage nonetheless played a crucial role by establishing the majority of the gene families that would later undergo amplifications.

Analysis of the predicted functions of the three sets of gene families identified as having been amplified, derived from domain fusions/fissions or derived from an HGT event (Figure 2G) indicated that they were enriched in several functional categories, notably carbohydrate metabolism, signal transduction, and transcription. These functional categories may have been important in the emergence of the complex brown algal cell wall or correspond to a complexification of signaling pathways as multicellular complexity increased. Interestingly, many of the genes acquired at, or shortly after, the origin of the Phaeophyceae encode secreted or membrane proteins (Figure S3A), suggesting roles in cell-cell communication that may have been important for the emergence of complex multicellularity or as components of defense mechanisms. The acquisition of plasmodesmata by brown algae directly after their divergence from their sister taxon *Schizocladia ischiensis*[22] (Figure 1) underlines the importance of cell-cell communication from the outset of brown algal evolution.

The emergence of the brown algae also corresponded with changes in gene structure. On average, brown algal genes tend to be more intron-rich than those of the other stramenopile groups,[23] including closely related taxa (Figure S4A), with the notable exception of *Chrysoparadoxa australica* (Figure S4A). A comparison of orthologous genes indicated a phase of rapid intron acquisition just before the divergence of the Phaeophyceae and the Schizocladiophyceae, followed by a period of relative intron stability up to the present day (Figure S4B). This phase of accelerated intron acquisition coincided approximately with the periods of marked gene gain and domain reorganization discussed above and may have been an indirect consequence of increased multicellular complexity (Figure 1) due to a concomitant decrease in effective population size.[24] Once established, increased intron density may have facilitated some of the genome-wide tendencies described above, such as increased reorganization of composite genes, for example, and thereby played an important role in a context of increasing developmental complexity.[25–28]

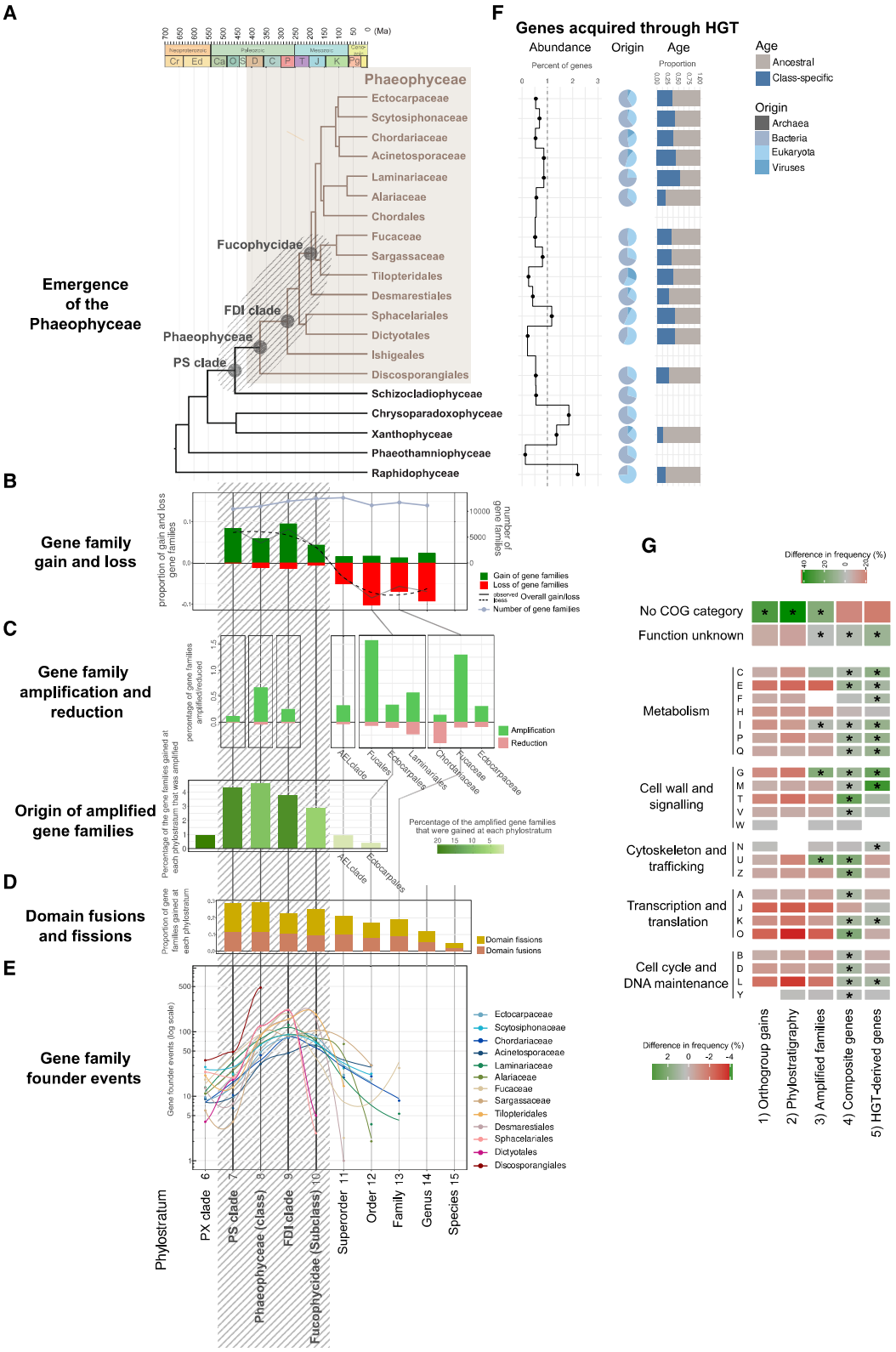
Acquisition of key metabolic and signaling pathways during the emergence of the Phaeophyceae

The success of the brown algae as an evolutionary lineage has been attributed, at least in part, to the acquisition of several key metabolic pathways, particularly those associated with cell-wall biosynthesis, and both halogen and phlorotannin metabolism.[3–5] Large complements of carbohydrate-active enzyme (CAZYme) genes (237 genes on average) were found in all brown algal orders and in their sister taxon *S. ischiensis*, but this class of gene was less abundant in the more distantly related unicellular alga *H. akashiwo* (Figures 3A, S5A, and S5B; Tables S4A and S4B). The evolutionary history of carbohydrate metabolism gene families was investigated by combining information from the genome-wide analyses of gene gain/loss, HGT and gene family amplification (Figure 3B). This analysis indicated that several key genes and gene families (mannuronan C5 epimerase [ManC5-E] and polysaccharide lyase 41 [PL41]) were acquired by the common ancestor of brown algae and *S. ischiensis*, with strong evidence in some cases that this occurred via HGT (PL41). Moreover, marked amplifications were detected for several families (AA15, ManC5-E, GH114, GT23, and PL41), indicating that both gain and amplification of gene families played important roles in the emergence of the brown algal carbohydrate metabolism gene set. Alginate is a major component of brown algal cell walls, and it plays an important role in conferring resistance to the biomechanical effect of wave action.[3] It is therefore interesting that ManC5-E, an enzyme whose action modulates the rigidity of the alginate polymer, appears to have been acquired very early (Figures 3B and 3C). The acquisition of ManC5-E, together with other alginate pathway enzymes such as PL41 (Figures 3A, 3B, and 3D), was probably an important evolutionary step, enabling the emergence of large, resilient substrate-anchored multicellular organisms in the highly dynamic and stressful coastal environment (Figure 1).

Vanadium-dependent haloperoxidases (vHPOs) are a central component of brown algal halogen metabolism, which has been implicated in multiple biological processes including defense, adhesion, chemical signaling, and the oxidative stress response. All three classes of brown algal vHPO (algal types I and II and bacterial-type[29–31]) appear to have been acquired early during the emergence of the Phaeophyceae (Figures 3A, S5C, and S5D; Tables S4C and S4D). Closely related stramenopile species do not possess any of these three types of haloperoxidase, with the exception of the sister taxon, *S. ischiensis*, which possesses three intermediate algal type (i.e., equidistant phylogenetically from class I and class II algal types) haloperoxidase genes (Figures 3A, S5C, and S5D). Algal type I and II vHPO

Figure 1. Ecology, diversity, and evolutionary features of the brown algae

The upper panel indicates approximate positions in the intertidal of key species whose genomes have been sequenced by the Phaeoexplorer project. The lower panel illustrates the diversity of brown algae (maximal values for each taxa) and indicates a number of key evolutionary events that occurred during the emergence of the Phaeophyceae. Some lineages may have secondarily lost a characteristic after its acquisition. Note that members of the genus *Ishige* (Ishigeaceae) also exhibit desiccation tolerance (not shown). ECM, extracellular matrix; asterisk (*), these orders were not analyzed in this study; Cr, Cryogenian; Ed, Ediacaran; Ca, Cambrian; O, Ordovician; S, Silurian; D, Devonian; C, Carboniferous; P, Permian; T, Triassic; J, Jurassic; K, Cretaceous; Pg, Paleogene.
See also Figures S1, S2, S4, and S5.



(legend on next page)

genes probably diverged from an intermediate-type ancestral gene similar to the *S. ischiensis* genes early during Phaeophyceae evolution. It is likely that the initial acquisitions of algal- and bacterial-type vHPOs represented independent events although the presence of probable vestiges of bacterial-type vHPO genes in *S. ischiensis* means that it is not possible to rule out acquisition of both types of vHPO through a single event.

Gene gain may not, however, have been the proximal factor responsible for all the key metabolic innovations that occurred in the emerging brown algal lineage. Phlorotannins are characteristic brown algal polyphenolic compounds that occur in all Phaeophyceae species, with the exception of some members of the Sargassaceae. Phlorotannins are derived from phloroglucinol and brown algae possess three classes of type III polyketide synthase, two of which (PKS1 and PKS2) were acquired prior to the emergence of the Phaeophyceae and the third (PKS3) evolving much later within the Ectocarpales (Figures 3A and 4A; Table S4E). Interestingly, PKS1 proteins from different brown algal species have been shown to have different activities leading to the production of distinct metabolites,[32–34] indicating that the acquisition of novel functions by this class of enzymes may have played an important role in the emergence of the brown algal capacity to produce phlorotannins. Moreover, many stramenopile PKS type III genes encode proteins with signal peptides or signal anchors (Table S4E). For the brown algae, this feature is consistent with the cellular production site of phlorotannins and the observed transport of these compounds by physodes, secretory vesicles characteristic of brown algae.[35] Cross-linking of phlorotannins, embedded within other brown algal cell-wall compounds such as alginates, has been demonstrated *in vitro* through the action of vHPOs[36–38] and indirectly suggested by *in vivo* observations colocalizing vHPOs with physode fusions at the cell periphery.[39,40] Consequently, vHPOs are good candidates for the enzymes that cross-link phlorotannins and other compounds, perhaps even for the formation of covalent bonds between phloroglucinol monomers and oligomers, which could occur via activation of aromatic rings through halogenation. These observations suggest that the acquisition of vHPOs by the common ancestor of brown algae and *S. ischiensis*, together perhaps with modifications of the existing PKS enzymes, triggered the emergence of new metabolic pathways leading to the production of the phlorotannin molecules characteristic of the Phaeophyceae lineage.

The acquisition of increased multicellular complexity and adaptation to new ecological niches during the early stages of brown algal evolution (i.e., during and immediately following the GOBE) is expected to have required modification and elaboration of signaling pathways. Membrane-localized signaling proteins (Figure 3A) are of particular interest in this context not only as potential mediators of intercellular signaling in a multicellular organism but also because of potential interactions with the elaborate brown algal extracellular matrices (cell walls).[3,42] A detailed analysis of the brown algal receptor kinase (RK) gene family, revealed that it actually includes two types of receptor, the previously reported leucine-rich repeat (LRR) RKs[11] and a newly discovered class of receptors with a beta-propeller extracellular domain (Figure 3A; Table S4F).

Major changes in epigenetic regulation also appear to have occurred during the emergence of the brown algae (see also supplemental information). *DNA METHYLTRANSFERASE 1* (*DNMT1*) genes were identified in *Discosporangium mesarthrocarpum* and two closely related outgroup species (*S. ischiensis* and *C. australica*) but not in other brown algal genomes, indicating that the common ancestor of brown algae probably possessed *DNMT1* but that this gene was lost after divergence of the Discosporangiales from other brown algal taxa (Figure 3A; Table S4G). This is consistent with the reported absence of DNA methylation in the filamentous brown alga *Ectocarpus*[11] and a very low level of DNA methylation in the kelp *Saccharina japonica*[43] (which is thought to be mediated by DNMT2). Our analysis indicates that most brown algae either lack DNA methylation or exhibit very low levels of methylation and that this feature was acquired early during brown algal diversification.

Impact of morphological, life cycle, and reproductive diversification during the Mesozoic on brown algal genome evolution

A second major step in the evolutionary history of the Phaeophyceae was the rapid diversification of the major brown algal orders, which began after the origin of the Fucophycidae/Dictyotales/Ishigeales (FDI) clade, here estimated at 235.97 Ma (95% highest posterior density region [HPD]: 158.88–312.48

Figure 2. Genome-wide analyses of brown algal genome and gene content evolution

(A) Time-calibrated cladogram based on Figure S2A. The gray hatched area, which indicates key nodes corresponding to the origin and early emergence of the brown algae, is mirrored in (B)–(F).

(B) Gene family (orthogroup) gain (green) and loss (red) during the emergence and diversification of the brown algae based on a Dollo parsimony reconstruction (Figure S2B).

(C) Upper: timing of gene family amplification and reduction during the evolutionary history of the Phaeophyceae (CAFE5 analysis). Lower: time of origin (orthogroup gain, based on the Dollo parsimony reconstruction) of the 180 most strongly amplified gene families.

(D) Composite gene analysis. Proportions of gene families showing domain fusion (orange) or domain fission (yellow) at different age strata.

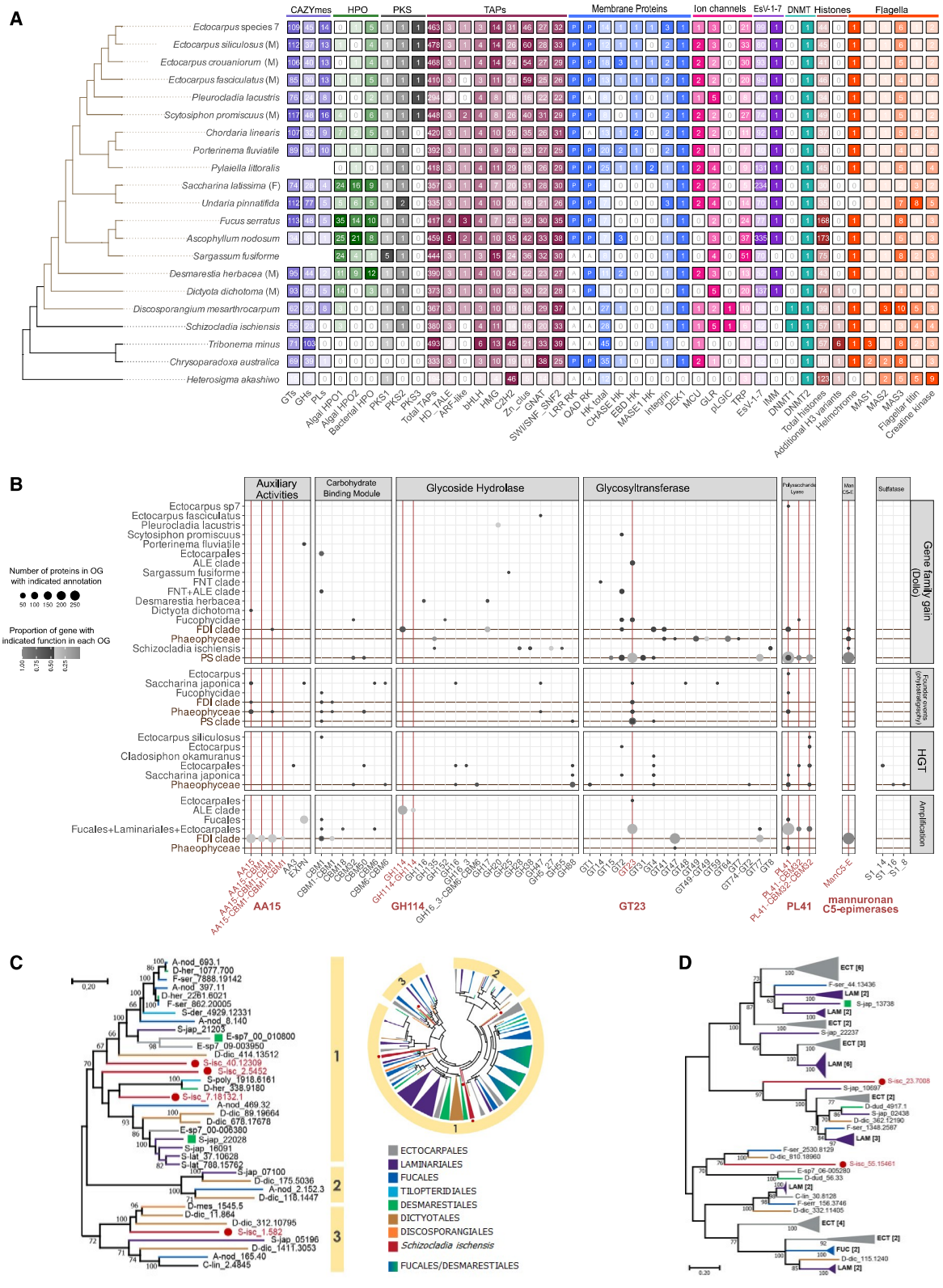
(E) Inferred gene family founder events after accounting for homology detection failure.

(F) Horizontal-gene-transfer-derived genes in orthologous groups and across species. The black trace represents the percentage of genes resulting from HGT events per species. Pie charts summarize the predicted origins (donor taxa) of the HGT genes. The right-hand bar graph indicates the proportions of ancestral (i.e., acquired before the root of the phylogenetic class, in gray) and class-specific (i.e., acquired within the phylogenetic class, in blue) HGT genes.

(G) Enrichment of COG categories in sets of gene families identified as being (1) gained at the four indicated early nodes by the Dollo analysis, (2) gene founder events at the four indicated phylostrata, (3) amplified in the Phaeophyceae (180 most strongly amplified families), (4) domain fusions or fissions, and (5) HGT derived. Asterisks indicate significantly enriched categories.

FDI clade, Fucophycideae/Dictyotales/Ishigeales; PS clade, Phaeophyceae plus Schizocladiophyceae; PX clade, Phaeophyceae plus Xanthophyceae.

See also Figure S3.



(legend on next page)

mya, broadly consistent with previous work[1]; Figures 1 and S2A). This diversification closely followed the Permian-Triassic mass extinction event (which dramatically impacted marine ecosystems in which red and green algae played dominant roles) and was facilitated by Triassic marine environments that favored chlorophyll-c containing algae (e.g., high phosphate and low iron concentration), along with the appearance of new coastal niches created by Pangea rifting (Figure 1). This context would have facilitated the diversification of the brown algal lineage,[44,45] resulting in organisms that now exhibit a broad range of morphological complexity (ranging from filamentous to complex parenchymatous thalli), different types of life cycle and diverse reproductive strategies and metabolic capacities[3,6,46,47] (Figure 1). The Phaeoexplorer dataset was analyzed to identify genomic features associated with this diversification of phenotypic characteristics and to evaluate the impact on genome evolution and function.

We found indications that the diversification of life cycles, in some cases linked with the emergence of large, complex body architectures, impacted genome evolution through population genetic effects. Most brown algae have haploid-diploid life cycles involving alternation between sporophyte and gametophyte generations, the only exception being the Fucales, which have diploid life cycles. The theoretical advantages of different types of life cycle have been discussed in detail,[48] and one proposed advantage of a life cycle with a haploid phase is that this allows effective purifying counter-selection of deleterious alleles. When the brown algae with haploid-diploid life cycles were compared with species from the Fucales, increased rates of both synonymous and non-synonymous mutation rates were detected in the latter, consistent with the hypothesis that deleterious alleles are phenotypically masked in species where most genes function in a diploid context (Figure S5E). Comparison of non-synonymous substitution rates (dN) for genes in brown algae with different levels of morphological complexity, ranging from simple filamentous thalli though parenchymatous to morphologically complex, indicated significantly lower values of dN for filamentous species (Figure S5E). This observation suggests that the emergence of larger, more complex brown algae may have resulted in reduced effective population sizes and consequently weaker counter-selection of non-synonymous substitutions.[24]

The diversification of the brown algae in terms of developmental complexity and life cycle structure was associated with modifications to reproductive systems, including, for example, partial or complete loss of flagella from female gametes in oogamous species and more subtle modifications such as loss of the eyespot in several kelps or of the entire posterior flagellum in *Dictyota dichotoma*.[49,50] Interestingly, these latter modifications are correlated with loss of the *HELMCHROME* gene, which is thought to be involved in light reception and zoid phototaxis,[41] from these species (Figures 3A and 4B). In addition, an analysis of the presence of genes for 70 high-confidence flagellar proteins[41] across eight species with different flagellar characteristics identified proteins that correlate with presence or absence of the eyespot or of the posterior flagellum (Figure 4B; Table S4H).

Brown algal diversification and the emergence of marine forests was also associated with genomic changes affecting metabolic and signaling pathways

Forests of brown algae (i.e., Laminariales, Desmarestiales, Tilopteridales, and Fucales[51]) are a key aspect of the modern marine biosphere. One of the pivotal innovations related to their emergence was a new developmental tissue, an intercalary meristem situated in the zone between the stipe and the lamina. The presence of this tissue is an ancestral state of the brown algal crown radiation (BACR) clade, and this study indicates that the intercalary meristem was acquired as early as 190 mya (Figure 1). This type of intercalary meristem would have facilitated the transition from annual to perennial life history and would, therefore, have been important for the establishment and maintenance of marine forests, particularly when upper parts of thalli are subjected to heavy grazing pressure.[10] Our results indicate that the Desmarestiales, Tilopteridales, and Fucales were all present by the early Cretaceous (Figure 1). Thus, it is possible that brown algal forests, at least at a small scale, provided both nutrients and shelter

Figure 3. Gene family evolution during the emergence of the brown algal lineage and a focus on carbohydrate metabolism

(A) Variations in size for a broad range of key gene families in the brown algae and closely related taxa. Numbers indicate the size of the gene family. Note that the *S. ischiensis* algal-type HPOs appear to be intermediate between classes I and II. Brown tree branches, Phaeophyceae.

(B) Overview of information from the orthogroup Dollo analysis, the phylostratigraphy analysis, the horizontal gene transfer analysis and the gene family amplification analysis for a selection of cell-wall active protein (CWAP) families. Dots represents functional family/orthogroup couples, with the size being proportional to the number of proteins annotated in the orthogroup (OG), and the color representing the proportion of the functional annotation that falls into this OG. Phaeophyceae plus Schizocladiophyceae (PS) and FDI clade, identified as gene innovation stages, are highlighted in brown. Functional categories with interesting evolutionary histories are highlighted in red.

(C) Phylogenetic tree of mannuronan C5-epimerases (ManC5-E). The phylogeny on the left, with three clusters indicated, is representative of the global view on the right.

(D) Phylogenetic tree of the polysaccharide lyase 41 (PL41) family. Green squares, biochemically characterized proteins. Brown algal sequences are color-coded in relation to their taxonomy, as indicated in (C). Schizocladiaphyceae sequences are shown in red and with a red circle.

P, present; A, absent; CAZYmes, carbohydrate-active enzymes; HPO, vanadium haloperoxidase; PKS, type III polyketide synthase; TAPs, transcription-associated proteins; EsV-1-7, EsV-1-7 domain proteins; DNMT, DNA methyltransferase; GTs, glycosyltransferases; GHs, glycoside hydrolases; ARF, auxin response factor-related; bHLH, basic-helix-loop-helix; HMG, high mobility group; Zn-clus, zinc cluster; C2H2, C2H2 zinc finger; GNAT, Gcn5-related N-acetyltransferase; SNF2, sucrose nonfermenting 2; LRR, leucine-rich repeat; QAD, β-propeller domain; RK, membrane-localized receptor kinase; HK, histidine kinase; CHASE, cyclases/histidine kinases associated sensory extracellular domain; EBD, ethylene-binding-domain-like; MASE1, membrane-associated sensor 1 domain; DEK1, defective kernel1; MCU, mitochondrial calcium uniporter; GLR, glutamate receptor; pLGIC, pentameric ligand-gated ion channel; TRP, transient receptor potential channel; IMM, IMMEDIATE UPRIGHT; H3, histone H3; MAS, mastigoneme proteins; AA, auxiliary activity; ECT, Ectocarpales; LAM, Laminariales; FUC, Fucales; DES, Desmarestiales.

See also Figure S5.

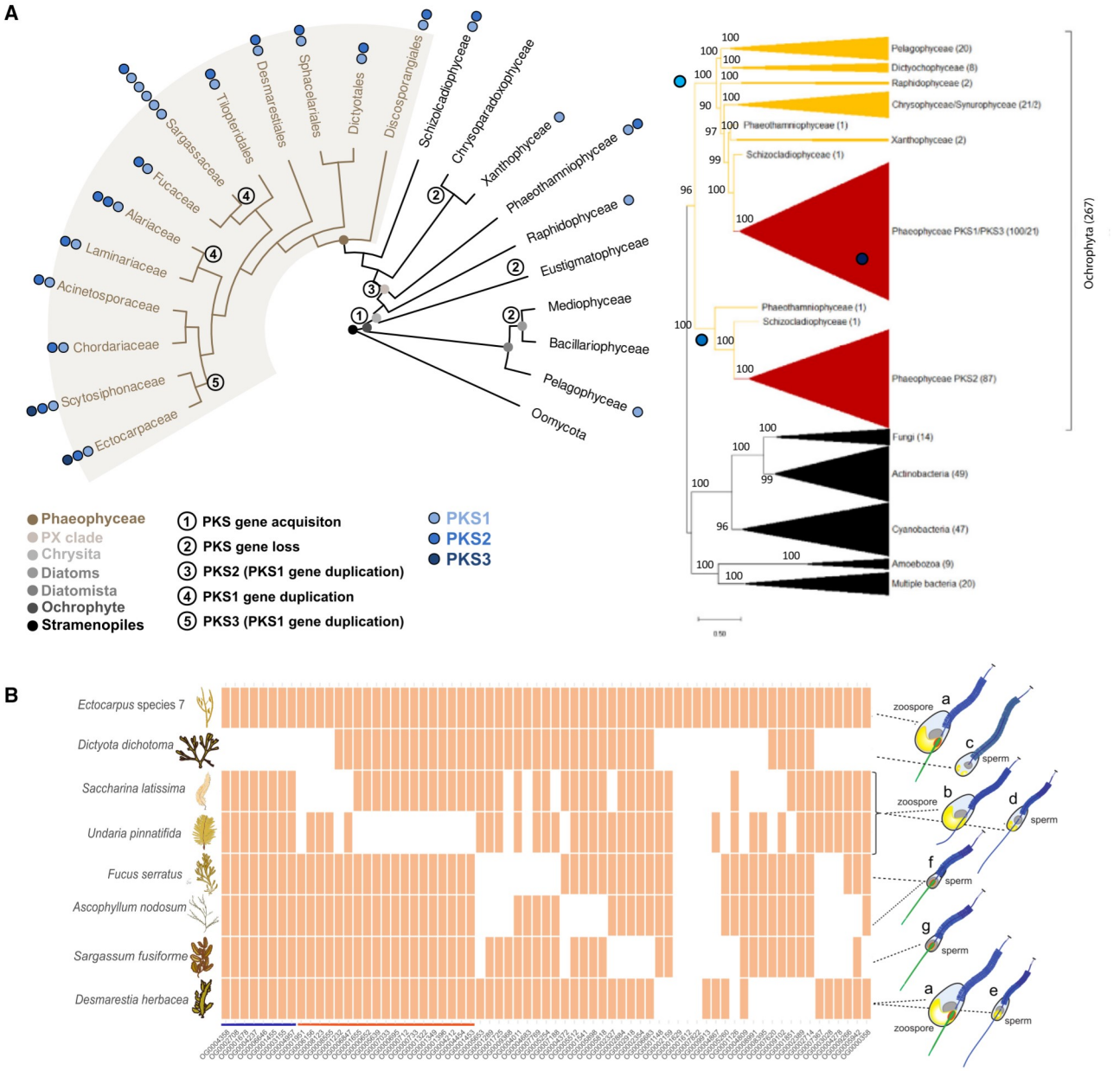


Figure 4. Evolution of key gene families during the emergence of the brown algal lineage

(A) Evolution of type III polyketide synthase (PKS) genes in the stramenopiles (left). Right: condensed view of a phylogenetic reconstruction tree of stramenopile PKS III and closely related sequences. In brackets: number of sequences identified in each phylogenetic group. Bootstrap values are indicated.

(B) Loss of orthogroups corresponding to flagellar proteome components[41] in eight brown algal species from five orders. For the zoid drawings: gray, nucleus; yellow, chloroplast; blue, anterior flagella with mastigonemes; red, eyespot. The posterior flagellum is shown either in green to indicate the presence of green autofluorescence correlated with the presence of the eyespot or in blue in species without an eyespot. Bars below the heatmap indicate gene losses associated with loss of just the eyespot (orange) or of the entire posterior flagellum (blue).

for the marine herbivorous animals that became common during the Cretaceous Period (e.g., algae-eating echinoids, sea turtles, and euteleostean fish[52,53]).

While our estimates of kelp antiquity are earlier than those of Starko et al.,[54] they are consistent with their suggestion that Cenozoic cooling facilitated the geographic expansion of the kelp forest ecosystem. Indeed, many of the animals found today in kelp forest ecosystems originated toward the end of the Cretaceous Period, or later.[55] Currently, our understanding of Mesozoic marine noncalcified macroalgae on the basis of fossils[56–58] is too poor to provide much guidance in this regard, but documentation by Kiel et al.[55] of fossil holdfasts indicates that kelp forests were present by the late Paleogene period (~32 mya). The highly complex, multi-layered, and canopy-forming kelp

forests of today, however, seem to have emerged only relatively recently, during the mid-Neogene, following the expansion of cooler water shelf environments.[54,55]

Comparative analysis of the Phaeoexplorer genome dataset identified a number of gene family expansions that potentially played important roles in the adaptation of the brown algae to their diverse niches and, more particularly, in the emergence of large, forest-forming species such as the kelps. For example, the ManC5-E family expanded markedly in the Laminariales and Fucales (Figure 3C), the two main orders that constitute extant phaeophycean forests. The capacity of ManC5-E to modify organ flexibility[3] may therefore have been an important factor for large organisms coping with the harsh hydrodynamic conditions of coastal environments.[59] In addition, five different orthogroups containing proteins with the mechanosensor wall stress-responsive component (WSC) domain were identified as having increased in size during the diversification of the brown algal lineage (Table S3), indicating that metabolic innovations affecting cell walls may have been concomitant with a complexification of associated signaling pathways.

Haloperoxidase gene families expanded independently in several brown algal orders, again with expansions being particularly marked in the Fucales and the Laminariales (Figures 3A and S5C). In the Laminariales, the algal type I family are specialized for iodine rather than bromine,[60] and this may have been an innovation that occurred specifically within the Laminariales, resulting in a halogen metabolism with an additional layer of complexity.

One of the proposed roles of halogenated molecules in brown algae is in biotic defense[4] and, clearly, an effective defense system would have been an important prerequisite for the emergence of the large, perennial organisms that constitute marine forests. Additional immunity-related families[61] that expanded during the diversification of the brown algae include five orthogroups that contain either GTPases with a central Ras of complex proteins/C-terminal of Roc domain tandem (ROCO GTPases) or nucleotide-binding adaptor shared by apoptotic protease-activating factor 1, R proteins, and CED-4 tetratricopeptide repeat (NB-ARC-TPR) genes (Table S3).

Finally, one of the most remarkable gene family amplifications detected in this study was for proteins containing the EsV-1-7 domain, a short, cysteine-rich motif that may represent a novel class of zinc finger.[62] EsV-1-7 domain proteins are completely absent from animal and land plant genomes and most stramenopiles either have just one member (oomycetes and eustigmatophytes) or entirely lack this gene family.[62] Analysis of the Phaeoexplorer data (Figure 3A; Table S4I) indicated that the EsV-1-7 gene family started to expand in the common ancestor of the brown algae and the raphidophyte *H. akashiwo*, with 31–54 members in the non-Phaeophyceae taxa that share this ancestor. Further expansion of the family then occurred in most brown algal orders, particularly in some members of the Laminariales (234 genes in *Saccharina latissima*) and the Fucales (335 genes in *Ascophyllum nodosum*), with the genes tending to be clustered in tandem arrays (Tables S3 and S4I). These observations are consistent with the previous description of a large EsV-1-7 domain family (95 genes) in *Ectocarpus* species[7,62] and with recent observations by Nelson et al.[20] One member of this family, IMMEDIATE UPRIGHT (IMM), has been shown to play a key role in the establishment of the elaborate basal filament system of *Ectocarpus* sporophytes,[62] suggesting that EsV-1-7 domain proteins may be novel developmental regulators in brown algae. Orthologs of the *IMM* gene were found in brown algal crown group taxa and in *D. dichotoma* but not in *D. mesarthrocarpum* (Figure 3A; Table S4I), indicating that this gene originated within the EsV-1-7 gene family as the first brown algal orders started to diverge.

Recent evolutionary events within the genus *Ectocarpus*

The above analyses focused on deep-time evolutionary events related to the emergence of the Phaeophyceae and the later diversification of the brown algal orders during the Mesozoic. To complement these analyses an evaluation of relatively recent and ongoing evolutionary events in the brown algae was conducted by sequencing 22 new strains from the genus *Ectocarpus*, which originated about 19 mya (Figure S2C).

A phylogenetic tree was constructed for 11 selected *Ectocarpus* species based on 261 high-quality alignments of 1:1 orthologs (Figure 5A). The tree indicates substantial divergence between *E. fasciculatus* and two well-supported clades, designated clade 1 and clade 2. Incongruencies between the species tree and trees for individual genes indicated introgression events and/or incomplete lineage sorting across the *Ectocarpus* genus. D-statistic analysis, specifically ABBA-BABA tests, detected incongruities among species quartets, indicating potential gene flow at various times during the evolution of the *Ectocarpus* genus. Evidence for gene flow was particularly strong for clade 2 and there was also evidence for marked exchanges between the two clades (Figure 5B), suggesting that gene flow has not been limited to recently diverged species pairs. These findings suggest a complex evolutionary history involving rapid divergence, hybridization, and introgression among species within the *Ectocarpus* genus, with evidence for hybridization occurring between 10.5 (for clades 1 and 2) and 3.3 mya (for *Ectocarpus* species 5 and 7) based on the fossil-calibrated tree (Figure S2C). A similar scenario has been reported for the genus *Drosophila*,[63] suggesting that recurrent hybridization and introgression among species may be a common feature associated with rapid species radiations. Major environmental changes such as the expansion of cold-water coastal areas following the green-house/cold-house Eocene-Oligocene transition (~30 mya[64]), and particularly the rapid climate destabilization and temperature drop associated with the end of the mid-Miocene thermal maximum (~15 mya[64]), may have created many new opportunities for the rapid expansion and diversification of the *Ectocarpus* genus.

Brown algal genomes contain large amounts of inserted viral sequences

A particularly striking result of this study was the identification of extensive amounts of integrated DNA sequence corresponding to large DNA viruses of the *Phaeovirus* family (Figure 6A; Table S5), which integrate into brown algal genomes as part of their lysogenic life cycles.[65] Analysis of 72 genomes in the Phaeoexplorer and associated public genome dataset identified a total of 792 viral regions (VRs) of *Nucleocytoviricota* (NCV) origin in 743 contigs, with a combined length of 32.3 Mbp.

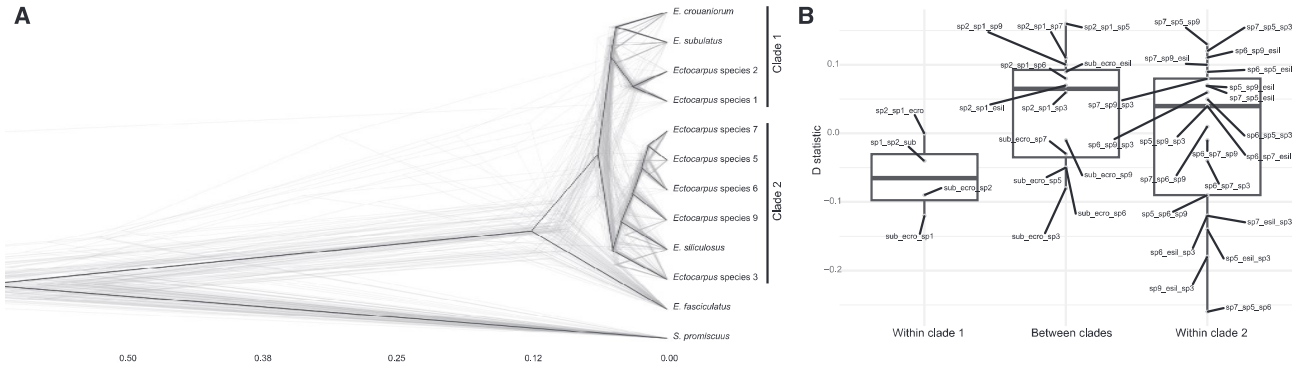


Figure 5. Evidence for gene flow within the genus *Ectocarpus*

(A) DensiTree visualisation of gene trees (gray lines) for 261 orthologs shared by 11 *Ectocarpus* species and the outgroup species *S. promiscuus*, together with the consensus species tree (black lines). All nodes of the species tree have posterior probabilities greater than 0.99.

(B) Boxplot reporting D-statistic (Patterson's D) values between P2 and P3 species. Within-lineage comparisons (i.e., within clades 1 and 2) and between-lineage comparisons are distinguished on the x axis. The annotation of each dot indicates species that were designated as P2 and P3. *Ectocarpus fasciculatus* was defined as the outgroup.

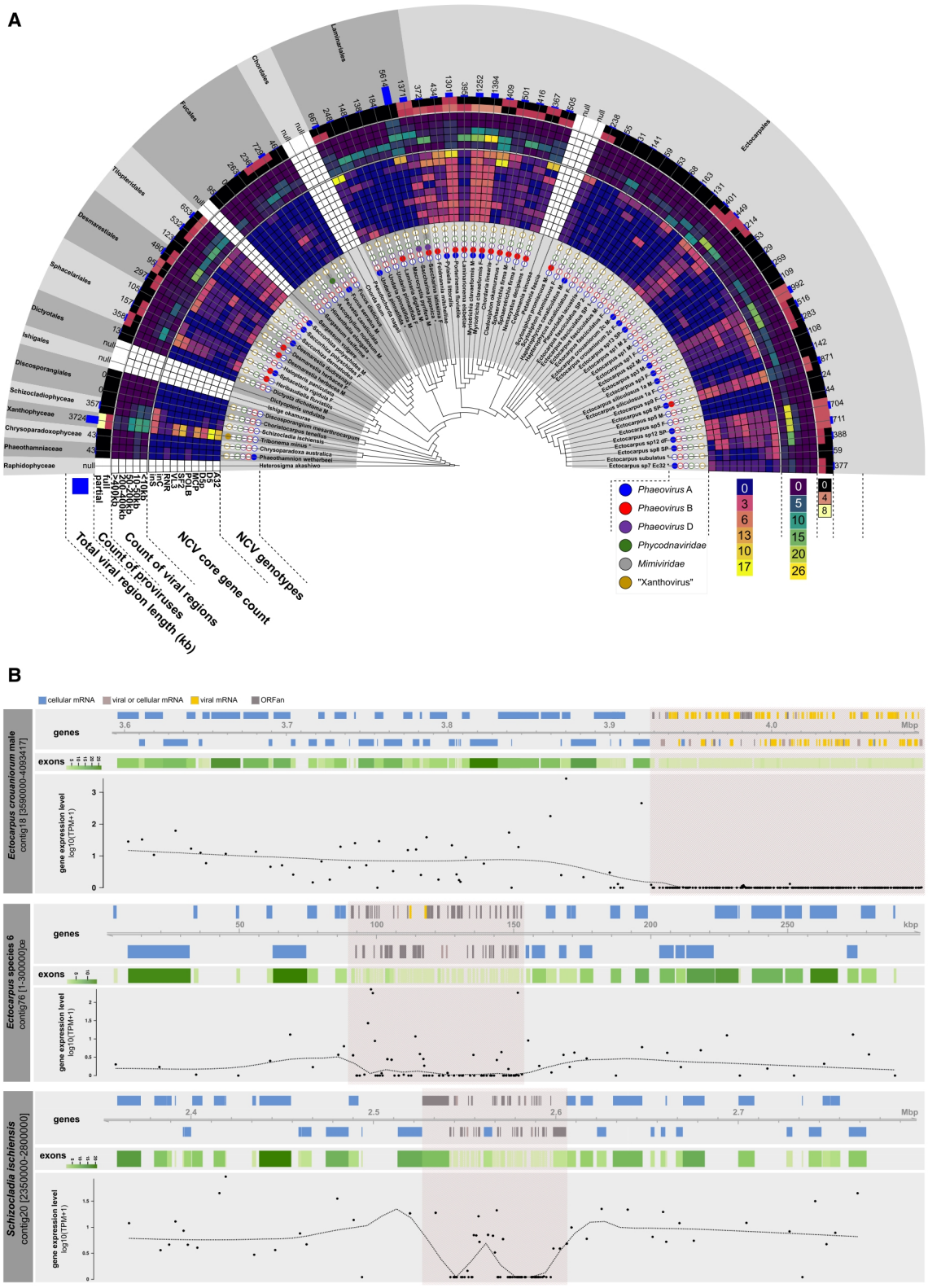
See also Figure S2.

Individual VRs ranged in size from two to 705 kbp, but the majority (81.3%) were between two and 50 kbp, while only 9% were longer than the expected minimum size (100 kbp) for an NCV genome. At least one flanking region could be identified for 40.8% of the VRs, providing direct evidence for insertion of the sequence in the algal genome (Table S5C). Figure 6B shows three examples of long VRs. Most genes in VRs are monoexonic and transcriptionally silent, as previously observed for the 310 kbp VR in the *Ectocarpus* species 7 strain Ec32 genome.[11]

On average, each of the 72 analyzed genomes contained 469 kbp of VR (with a maximum of 5,614 kbp) and only two genomes contained no VRs (both from the Discosporangiales). There were a number of outlier genomes that contained more than 1 Mbp of VRs (*T. minus*, *S. latissima*, *S. japonica*, *P. fluviatile*, and *Myriotrichia clavaeformis* male and female). At least one partial provirus (a VR possessing several key NCV marker genes) was present in 39 genomes, 29 of which had at least one full provirus with a complete set of seven key NCV marker genes (Figure 6A; Table S5). In addition to the previously known infections in Ectocarpales[65] and Laminariales,[66] integrated NCV proviruses were found in all Phaeophyceae orders screened, except the Discosporangiales and Dictyotales, and were also detected in *T. minus* (Xanthophyceae). Moreover, NCV marker gene composition indicated that multiple integrated proviruses were present in 16 genomes from multiple Phaeophyceae orders (Ectocarpales, Desmarestiales, Sphacelariales, Tilopteridales, and Laminariales), and the Xanthophyceae (Figure 6A; Table S5). Phylogenetic analysis of the major capsid protein (MCP) and DNA polymerase genes indicated that the majority of the integrated NCVs belonged to the genus *Phaeovirus*, the sole viral group known to infect brown algae (Figures S6A and S6B). However, this analysis also revealed integrated sequences corresponding to other viral groups. Viral sequences in *T. minus* belonged to a putative novel genus closely related to *Phaeovirus*, for which we propose the name *Xanthovirus*. Finally, mimiviridae-related VRs were identified in *S. latissima* and *Pelvetia canaliculata*, but since they are partial proviruses and do not appear to possess integrase genes, they may have originated from ancient endogenization events, similar to those described in chlorophytes.[67]

The identification of integrated NCVs across almost all brown algal orders and in closely related outgroup taxa suggests that the lysogenic life cycle strategy of phaeoviruses is ancient and that giant viral genomes have been integrating into the genomes of brown algae throughout the latters' evolutionary history. This conclusion was supported by the phylostratigraphic analysis, which detected the appearance of many novel virus-related genes dating back to the origin of the Phaeophyceae (Figure S3A). Marked differences were detected in total VR size and NCV marker gene presence across the brown algal genome set, and large differences were even detected between strains from the same genus (between 24 and 992 kbp of VR in different *Ectocarpus* spp. for example; Figure 6A; Table S5). These differences indicate dynamic changes in VR content over evolutionary time, presumably due, at least in part, to differences in rates of viral genome integration, a process that can involve multiple, separate insertion events,[68] and rates of VR loss due to meiotic segregation.[69] In addition, the abundant presence of partial proviruses and NCV fragments in brown algal genomes indicates that inserted VRs can degenerate and fragment, probably also leading to VR loss over time. The identification of large-scale viral genome insertion events over such a long timescale (at least 450 mya[1]) suggests that NCVs may have had a major impact on the evolution of brown algal genomes throughout the emergence of the lineage.

The widespread presence of large quantities of viral genes in brown algal genomes creates a favorable situation for recruitment of this genetic information by the algal host via HGT (provided the acquired genes confer a selective advantage[70]), but clear evidence of this type of HGT event can be difficult to obtain. However, phylogenetic evidence indicates that several *Ectocarpus* species 7 histidine kinases (HKs) were derived by HGT from viral insertions[71] and analysis of the Phaeoexplorer genomes supported this hypothesis. HKs are widespread in the



(legend on next page)

stramenopiles but several classes of membrane-localized HK were either only found in brown algae (cyclases/histidine kinases associated sensory extracellular [CHASE] domain HKs and HKs with an extracellular domain resembling an ethylene binding motif[71]) or only in brown algae and closely related taxa (membrane-associated sensor 1 [MASE1] domain HKs[71]) and appear to be absent from other stramenopile lineages (Figure 3A; Table S4J). These classes of HK all exhibit a patchy pattern of distribution across the brown algae and are often monoexonic suggesting possible multiple acquisitions from viruses via HGT following integration of viral genomes into algal genomes (Figures 3A and S6C). Phylogenetic analysis provided further support for a HGT origin for these classes of HK (Figure S6C).

DISCUSSION

Comparative analysis of the genome resource presented in this study has provided insights into genome evolution across the entire evolutionary history of the brown algae. A period of marked genome evolution concomitant with the emergence of the brown algal lineage during the GOBE was correlated with an increase in multicellular complexity, possibly driven, at least in part, by increases in atmospheric oxygen and herbivory. During this period, the brown algae acquired key components of several metabolic pathways, notably cell-wall polysaccharide, phlorotannin, and halogen metabolisms, that were essential for their colonization of intertidal and subtidal environments. The capacity to synthesize flexible and resilient alginate-based cell walls[72] allows these organisms to resist the hydrodynamic forces of wave action,[59] whereas phlorotannins and halogen derivatives are thought to play important roles in defense.[73] There is also evidence that cell-wall cross-linking by phlorotannins may be important for strong adhesion to substrata, another important characteristic in the dynamic intertidal and subtidal coastal environments.[74] The capacity to adhere strongly and resist both biotic and abiotic stress factors would prove essential for the success of large, sedentary multicellular organisms in these intertidal niches over evolutionary time.

The period of increased gene gain during the emergence of the brown algae was followed by a period of overall gene loss that extended up until the present day (Figure S2B). Interestingly, similar periods of ancestral gene gain followed by gene loss have also been observed for both the animal and land plant lineages,[75] indicating that this may be a common feature of multicellular eukaryotic lineages.

About 220 mya after the emergence of the brown algae, the aftermath of the Permian-Triassic mass extinction event and the initiation of Pangea rifting appear to have created favorable environments for rapid diversification of the main brown algal orders,[44,45] resulting in the emergence of a diversity of developmental, life cycle, and reproductive strategies, with correlated effects on genome evolution. During this period some orders, such as the Laminariales and Fucales, acquired characteristics such as an intercalary meristems and modified metabolic, defense, and developmental processes that are predicted to have been important prerequisites for the emergence of marine forests.

Analysis of the genomes of multiple *Ectocarpus* species demonstrated that genomic modifications, including gene gain and gene loss have continued to occur up until the present time and indicated that these modifications can potentially be transmitted between species as a result of gene flow occurring within a genus due to incomplete reproductive boundaries and introgression.

Finally, one of the most surprising observations was that brown algal genomes contain many inserted viral sequences corresponding to large DNA viruses of the *Phaeovirus* family. Inserted viral sequences are widespread in eukaryotic genomes[76,77] and insertions corresponding to nucleocytoplasmic large DNA viruses have been found in green algal genomes[67,78] but the brown algal *Phaeovirus* VRs are remarkable because they are nearly ubiquitous in this lineage (being present in 67 of 69 brown algal genomes analyzed) and because individual genomes can contain several phylogenetically diverse *Phaeovirus* insertions and insertions of a broad range of different sizes. The near ubiquitous occurrence of these elements may be attributed to the capacity of phaeoviruses to insert into their hosts' genomes as part of their life cycle.

The above observations illustrate how the Phaeoexplorer genome dataset, along with the various analyses carried out in this study, can be used to link the gene content of brown algal genomes to biological processes and characteristics that have played fundamental roles during the evolution of this lineage. The establishment of this genome resource represents an important step forward for a key lineage that has remained poorly characterized at the genome level. The Phaeoexplorer dataset not only provides good quality genome assemblies for many, previously uncharacterized brown algal species but also represents a tool to explore genome function via comparative genomics approaches, adding an important evolutionary dimension to efforts to understand gene function in this lineage. The identification and analysis of key metabolic and signaling genes implicated

Figure 6. Inserted viral regions in brown algal genomes

(A) Annotated phylogeny summarizing key statistics of the presence of *Nucleocytoviricota* (NCV) sequences in the genomes of brown algae and closely related taxa. Eight genomes sourced from public databases are labeled with an asterisk. Outer layers around the tree are as follows (1) NCV genotypes in each genome, (2) NCV core gene count indicates the number of copies of each viral core gene (A32, A32 packaging ATPase; D5/D5p, D5 helicase/primase; MCP, major capsid protein; POLB, DNA polymerase B; SF2, superfamily 2 helicase; VL3, very late transcription factor 3; RNR, ribonucleotide reductase; inC, integrase recombinase; inS, integrase resolvase), (3) count of viral regions is the number of viral regions within each size range category as indicated, (4) count of proviruses is the estimated number of complete or partial integrated viral genomes in a genome, (5) total viral region length is the sum of the lengths in kbp of all viral regions within a genome. The outermost layer indicates the taxonomic class or order of the host clades.

(B) Three examples of contigs containing large viral insertions (pink shading). Genes (colored boxes) were classified as viral, cellular (i.e., cellular organism), known proteins of unclear origin (viral or cellular) or unknown (ORFan) based on comparisons with viral and cellular protein databases (see STAR Methods). Transcript abundances are shown with a locally estimated scatterplot smoothing (LOESS) plot. Exons, exons per gene.

See also Figure S6.

in a broad range of brown algal biological functions represents an important resource for future research programs aimed at optimizing brown seaweed production in a mariculture context or at preserving and protecting natural seaweed populations in the context of climate change. Both of these approaches could potentially contribute to mitigation of the effects of climate change via multiple positive effects in terms of carbon capture, ecosystem services, and the promotion of highly sustainable cultivation practices.

To facilitate future use of this genome dataset, the annotated genomes have been made available through a website portal (https://phaeoexplorer.sb-roscoff.fr). The existing genome dataset provides very good coverage of the phylogenetic diversity of the Phaeophyceae and reasonably complete gene catalogs for each species, but future work is needed to improve further the quality of the genome assemblies described here and to add genomes for additional species, particularly members of the minor brown algal orders that are not represented in the dataset. The large proportion of genes with no predicted function in brown algal genomes is also a limitation that needs to be addressed. The recent development of CRISPR-Cas9 methodology for brown algae,[79,80] together with the other tools and resources currently available for the model brown alga *Ectocarpus*,[81] provide the means to deploy the functional genomics approaches necessary to address this question.

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, J. Mark Cock (cock@sb-roscoff.fr).

Materials availability

All the laboratory-cultivated strains grown to provide material for genome sequencing can be accessed via the Roscoff Culture Collection (https://www.roscoff-culture-collection.org).

Data and code availability

- All sequence data, including DNA and RNA sequencing data, genome assemblies, and annotations, have been deposited in the European Bioinformatics Institute/European Nucleotide Archive (EBI/ENA) database under the project accession PRJEB76691 and are publicly available. Additional data and results have been deposited in the CNRS Research Data depository (https://doi.org/10.57745/9U1J85) and are publicly available.
- This paper does not report original code.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

ACKNOWLEDGMENTS

This work was supported by the France Génomique National Infrastructure project Phaeoexplorer (ANR-10-INBS-09), the European Research Council project Sexsea (638240), the Investissements d'Avenir project Idealg (ANR-10-BTBR-04-01), the European BG-01 BlueGrowth H2020 project Genialg (727892), Laoshan Laboratory grants (LSKJ202203801, LSKJ202203204), the Taishan Scholars Program and China Agriculture Research System (CARS-50), the CNRS international research network DABMA (00022), the ANR projects Epicycle (ANR-19-CE20-0028-01), BrownSugar (ANR-20-CE44-0011), HaloGene (ANR-22-CE20-0025), Seabioz (ANR-20-CE43-0013) and BrownLincs (ANR-23-CE20-0048-01), the National Research Foundation of Korea (2022R1A2B5B03002312, 2022R1A5A1031361) and Ministry of Oceans and Fisheries (KIMST-20180430) granted to H.S.Y., the projects Connect Talent EpiAlg Région Pays de la Loire-Nantes Métropole and Etoiles Montantes M-EpiCC Région Pays de la Loire, the MITI-funded project Algometabionte, Dr. Karl Feldbausch-Stiftung, the CNRS, and Sorbonne University. We are grateful to the Roscoff Bioinformatics platform ABiMS (http://abims.sb-roscoff.fr), which is part of the Institut Français de Bioinformatique (ANR-11-INBS-0013) and BioGenouest network, for providing both help and computing and storage resources.

AUTHOR CONTRIBUTIONS

Software, L.B.-G., R.D., A.L.B., X.L., D.N., and E.C.; formal analysis, F.D., O.G., L.D., D.M., T.M., D. Sussfeld, X.F., L. Mazéas, N. Terrapon, J.B.-R., R.P., L.R., S.-W.C., J.J., K.U., K.B., C. Duc, P.R., A.L., E.A.M., M.L., A.K., P.H.-G., C.V., S.S.A., S.A., K.A., Y.B., T. Barbeyron, A.B., S.B., A.B.-C., A. Cormier, H.C.d.C., A.D., E. Dinatale, S.D., E. Drula, J.G., L.G., A.G., M.-L.G., L.H., B.H., A.J., E.K., A.H.K., C.L., L.L.G., R.L., S.T.L., P.J.L., E.M., S.M., G.M., C.N., S.A.R., E.R., D. Schroeder, A.S., L.T., T.T., K.V., H.V., G.W., H.K., A.F.P., H.S.Y., C.H., N.Y., E.B., M.V., G.V.M., E.C., S.M.C., J.-M.A., and J.M.C.; investigation, C.C., S.H., Z.N., N. Tadrent, A. Couloux, B.N., W.B., E. Denis, C.J., L. Mest, S.R., and D. Scornet; resources, O.G., A. Couloux, L.B.-G., T. Bringloe, R.A.C., C. Destombe, S.F., W.J.H., G.H., K.K., A.L.B., K.M., C.M., N.P., P.P., S.R., D. Scornet, H.V., F.W., H.K., A.F.P., M.V., E.C., and J.-M.A.; data curation, O.G., C.C., A. Couloux, L.B.-G., J.-M.A., and J.M.C.; writing – original draft, F.D., O.G., J.-M.A., and J.M.C.; writing – review and editing, all authors; visualization, F.D., O.G., L.B.-G., L.D., D.M., T.M., D. Sussfeld, X.F., L. Mazéas, J.B.-R., R.P., L.R., K.U., K.B., P.R., A.L., M.L., P.H.-G., C.V., A.D., A.L.B., H.K., E.C., and J.M.C.; supervision, O.G., N. Terrapon, M.L., R.A.C., H.C.d.C., O.D.C., S.D., G.H., A.J., C.B., E.P., P.L., S.A.R., A.S., L.T., H.V., G.W., H.K., H.S.Y., C.H., N.Y., E.B., M.V., G.V.M., E.C., S.M.C., J.-M.A., and J.M.C.; project administration, F.D., E.B., M.V., G.V.M., E.C., S.M.C., P.W., J.-M.A., and J.M.C.; funding acquisition, H.C.d.C., C.B., P.P., C.H., N.Y., S.M.C., and J.M.C.; conceptualization, J.M.C.

DECLARATION OF INTERESTS

The authors declare no competing interests.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- EXPERIMENTAL MODEL DETAILS
  - *Ascophyllum nodosum*
  - *Chordaria linearis* strain ClinC8C
  - *Choristocarpus tenellus* strain KU-1152
  - *Chrysoparadoxa australica* strain CS-1217
  - *Cladosiphon okamuranus* strain S-strain
  - *Desmarestia dudresnayi* strain DdudBR16
  - *Desmarestia herbacea* strain DmunF
  - *Desmarestia herbacea* strain DmunM
  - *Dictyota dichotoma* strain KB07f IV
  - *Dictyota dichotoma* strain ODC1387m
  - *Dictyota dichotoma* strain KB07m IV
  - *Dictyota dichotoma* strain KB07sp VI
  - *Discosporangium mesarthrocarpum* strain MT17-79
  - *Ectocarpus crouaniorum* strain Ec861
  - *Ectocarpus crouaniorum* strain Ec862
  - *Ectocarpus fasciculatus* strain Ec846
  - *Ectocarpus fasciculatus* strain Ec847
  - *Ectocarpus fasciculatus* strain EfasUO1
  - *Ectocarpus fasciculatus* strain EfasUO2
  - *Ectocarpus siliculosus* strain Ec863
  - *Ectocarpus siliculosus* strain Ec864
  - *Ectocarpus* species 1 strain Ec sil Puy CHCH Z9 G5f

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.cell.2024.10.049.

Received: February 19, 2024
Revised: July 20, 2024
Accepted: October 28, 2024
Published: November 20, 2024

REFERENCES

1. Choi, S.-W., Graf, L., Choi, J.W., Jo, J., Boo, G.H., Kawai, H., Choi, C.G., Xiao, S., Knoll, A.H., Andersen, R.A., et al. (2024). Ordovician origin and subsequent diversification of the brown algae. Curr. Biol. *34*, 740–754.e4. https://doi.org/10.1016/j.cub.2023.12.069.

2. Keeling, P.J. (2009). Chromalveolates and the evolution of plastids by secondary endosymbiosis. J. Eukaryot. Microbiol. *56*, 1–8. https://doi.org/10.1111/j.1550-7408.2008.00371.x.

3. Mazéas, L., Yonamine, R., Barbeyron, T., Henrissat, B., Drula, E., Terrapon, N., Nagasato, C., and Hervé, C. (2023). Assembly and synthesis of the extracellular matrix in brown algae. Semin. Cell Dev. Biol. *134*, 112–124. https://doi.org/10.1016/j.semcdb.2022.03.005.

4. Küpper, F.C., and Carrano, C.J. (2019). Key aspects of the iodine metabolism in brown algae: a brief critical review. Metallomics *11*, 756–764. https://doi.org/10.1039/c8mt00327k.

5. Schoenwaelder, M.E.A. (2008). The biology of phenolic containing vesicles. Algae *23*, 163–175. https://doi.org/10.4490/ALGAE.2008.23.3.163.

6. Cock, J.M., Godfroy, O., Macaisne, N., Peters, A.F., and Coelho, S.M. (2014). Evolution and regulation of complex life cycles: a brown algal perspective. Curr. Opin. Plant Biol. *17*, 1–6. https://doi.org/10.1016/j.pbi.2013.09.004.

7. Eger, A.M., Marzinelli, E.M., Beas-Luna, R., Blain, C.O., Blamey, L.K., Byrnes, J.E.K., Carnell, P.E., Choi, C.G., Hessing-Lewis, M., Kim, K.Y., et al. (2023). The value of ecosystem services in global marine kelp forests. Nat. Commun. *14*, 1894. https://doi.org/10.1038/s41467-023-37385-0.

8. Wernberg, T., Russell, B.D., Thomsen, M.S., Gurgel, C.F.D., Bradshaw, C.J.A., Poloczanska, E.S., and Connell, S.D. (2011). Seaweed communities in retreat from ocean warming. Curr. Biol. *21*, 1828–1832. https://doi.org/10.1016/j.cub.2011.09.028.

9. Ross, F.W.R., Boyd, P.W., Filbee-Dexter, K., Watanabe, K., Ortega, A., Krause-Jensen, D., Lovelock, C., Sondak, C.F.A., Bach, L.T., Duarte, C.M., et al. (2023). Potential role of seaweeds in climate change mitigation. Sci. Total Environ. *885*, 163699. https://doi.org/10.1016/j.scitotenv.2023.163699.

10. Bringloe, T.T., Starko, S., Wade, R.M., Vieira, C., Kawai, H., De Clerck, O.D., Cock, J.M., Coelho, S.M., Destombe, C., Valero, M., et al. (2020). Phylogeny and evolution of the brown algae. Crit. Rev. Plant Sci. *39*, 281–321. https://doi.org/10.1080/07352689.2020.1787679.

11. Cock, J.M., Sterck, L., Rouzé, P., Scornet, D., Allen, A.E., Amoutzias, G., Anthouard, V., Artiguenave, F., Aury, J.M., Badger, J.H., et al. (2010). The *Ectocarpus* genome and the independent evolution of multicellularity in brown algae. Nature *465*, 617–621. https://doi.org/10.1038/nature09016.

12. Ye, N., Zhang, X., Miao, M., Fan, X., Zheng, Y., Xu, D., Wang, J., Zhou, L., Wang, D., Gao, Y., et al. (2015). *Saccharina* genomes provide novel insight into kelp biology. Nat. Commun. *6*, 6986. https://doi.org/10.1038/ncomms7986.

13. Graf, L., Shin, Y., Yang, J.H., Choi, J.W., Hwang, I.K., Nelson, W., Bhattacharya, D., Viard, F., and Yoon, H.S. (2021). A genome-wide investigation of the effect of farming and human-mediated introduction on the ubiquitous seaweed *Undaria pinnatifida*. Nat. Ecol. Evol. *5*, 360–368. https://doi.org/10.1038/s41559-020-01378-9.

14. Wang, S., Lin, L., Shi, Y., Qian, W., Li, N., Yan, X., Zou, H., and Wu, M. (2020). First draft genome assembly of the seaweed *Sargassum fusiforme*. Front. Genet. *11*, 590065. https://doi.org/10.3389/fgene.2020.590065.

15. Diesel, J., Molano, G., Montecinos, G.J., DeWeese, K., Calhoun, S., Kuo, A., Lipzen, A., Salamov, A., Grigoriev, I.V., Reed, D.C., et al. (2023). A scaffolded and annotated reference genome of giant kelp (*Macrocystis pyrifera*). BMC Genomics *24*, 543. https://doi.org/10.1186/s12864-023-09658-x.

16. Dittami, S.M., Scornet, D., Petit, J.L., Ségurens, B., Da Silva, C., Corre, E., Dondrup, M., Glatting, K.H., König, R., Sterck, L., et al. (2009). Global expression analysis of the brown alga *Ectocarpus siliculosus* (Phaeophyceae) reveals large-scale reprogramming of the transcriptome in response to abiotic stress. Genome Biol. *10*, R66. https://doi.org/10.1186/gb-2009-10-6-r66.

17. Wang, S., and Wu, M. (2023). The draft genome of the "golden tide" seaweed, *Sargassum horneri*: characterization and comparative analysis. Genes (Basel) *14*, 1969. https://doi.org/10.3390/genes14101969.

18. Nishitsuji, K., Arimoto, A., Iwai, K., Sudo, Y., Hisata, K., Fujie, M., Arakaki, N., Kushiro, T., Konishi, T., Shinzato, C., et al. (2016). A draft genome of the brown alga, *Cladosiphon okamuranus*, S-strain: a platform for future studies of "mozuku" biology. DNA Res. *23*, 561–570. https://doi.org/10.1093/dnares/dsw039.

19. Nishitsuji, K., Arimoto, A., Higa, Y., Mekaru, M., Kawamitsu, M., Satoh, N., and Shoguchi, E. (2019). Draft genome of the brown alga, *Nemacystus decipiens*, Onna-1 strain: fusion of genes involved in the sulfated fucan biosynthesis pathway. Sci. Rep. *9*, 4607. https://doi.org/10.1038/s41598-019-40955-2.

20. Nelson, D.R., Mystikou, A., Jaiswal, A., Rad-Menendez, C., Preston, M.J., Boever, F.D., Assal, D.C.E., Daakour, S., Lomas, M.W., Twizere, J.-C., et al. (2024). Macroalgal deep genomics illuminate multiple paths to aquatic, photosynthetic multicellularity. Mol. Plant *17*, 747–771. https://doi.org/10.1016/j.molp.2024.03.011.

21. LoDuca, S.T., Bykova, N., Wu, M., Xiao, S., and Zhao, Y. (2017). Seaweed morphology and ecology during the great animal diversification events of the Early Paleozoic: A tale of two floras. Geobiology *15*, 588–616. https://doi.org/10.1111/gbi.12244.

22. Kawai, H., Maeba, S., Sasaki, H., Okuda, K., and Henry, E.C. (2003). Schizocladia ischiensis: a new filamentous marine chromophyte belonging to a new class, Schizocladiophyceae. Protist *154*, 211–228. https://doi.org/10.1078/143446103322166518.

23. Roy, S.W., and Penny, D. (2007). A very high fraction of unique intron positions in the intron-rich diatom *Thalassiosira pseudonana* indicates widespread intron gain. Mol. Biol. Evol. *24*, 1447–1457. https://doi.org/10.1093/molbev/msm048.

24. Lynch, M., and Conery, J.S. (2003). The origins of genome complexity. Science *302*, 1401–1404. https://doi.org/10.1126/science.1089370.

25. Vosseberg, J., Stolker, D., von der Dunk, S.H.A., and Snel, B. (2023). Integrating phylogenetics with intron positions illuminates the origin of the complex spliceosome. Mol. Biol. Evol. *40*, msad011. https://doi.org/10.1093/molbev/msad011.

26. Yang, P., Wang, D., and Kang, L. (2021). Alternative splicing level related to intron size and organism complexity. BMC Genomics *22*, 853. https://doi.org/10.1186/s12864-021-08172-2.

27. Nilsen, T.W., and Graveley, B.R. (2010). Expansion of the eukaryotic proteome by alternative splicing. Nature *463*, 457–463. https://doi.org/10.1038/nature08909.

28. Chen, L., Bush, S.J., Tovar-Corona, J.M., Castillo-Morales, A., and Urrutia, A.O. (2014). Correcting for differential transcript coverage reveals a strong relationship between alternative splicing and organism complexity. Mol. Biol. Evol. *31*, 1402–1413. https://doi.org/10.1093/molbev/msu083.

29. Wischang, D., Radlow, M., Schulz, H., Vilter, H., Viehweger, L., Altmeyer, M.O., Kegler, C., Herrmann, J., Müller, R., Gaillard, F., et al. (2012). Molecular cloning, structure, and reactivity of the second bromoperoxidase from *Ascophyllum nodosum*. Bioorg. Chem. *44*, 25–34. https://doi.org/10.1016/j.bioorg.2012.05.003.

30. Radlow, M., Czjzek, M., Jeudy, A., Dabin, J., Delage, L., Leblanc, C., and Hartung, J. (2018). X-ray diffraction and density functional theory provide insight into vanadate binding to homohexameric bromoperoxidase II and the mechanism of bromide oxidation. ACS Chem. Biol. *13*, 1243–1259. https://doi.org/10.1021/acschembio.8b00041.

31. Fournier, J.-B., Rebuffet, E., Delage, L., Grijol, R., Meslet-Cladière, L., Rzonca, J., Potin, P., Michel, G., Czjzek, M., and Leblanc, C. (2014). The Vanadium Iodoperoxidase from the marine Flavobacteriaceae species *Zobellia galactanivorans* reveals novel molecular and evolutionary features of halide specificity in the vanadium haloperoxidase enzyme family. Appl. Environ. Microbiol. *80*, 7561–7573. https://doi.org/10.1128/AEM.02430-14.

32. Meslet-Cladière, L., Delage, L., Leroux, C.J.J., Goulitquer, S., Leblanc, C., Creis, E., Gall, E.A., Stiger-Pouvreau, V., Czjzek, M., and Potin, P. (2013). Structure/function analysis of a Type III polyketide synthase in the brown alga *Ectocarpus siliculosus* reveals a biochemical pathway in phlorotannin monomer biosynthesis. Plant Cell *25*, 3089–3103. https://doi.org/10.1105/tpc.113.111336.

33. Baharum, H., Morita, H., Tomitsuka, A., Lee, F.C., Ng, K.Y., Rahim, R.A., Abe, I., and Ho, C.L. (2011). Molecular cloning, modeling, and site-directed mutagenesis of type III polyketide synthase from *Sargassum binderi* (Phaeophyta). Mar. Biotechnol. (NY) *13*, 845–856. https://doi.org/10.1007/s10126-010-9344-5.

34. Zhao, D.-S., Hu, Z.-W., Dong, L.-L., Wan, X.-J., Wang, S., Li, N., Wang, Y., Li, S.-M., Zou, H.-X., and Yan, X. (2021). A Type III polyketide synthase (SfuPKS1) isolated from the edible seaweed *Sargassum fusiforme* exhibits broad substrate and catalysis specificity. J. Agric. Food Chem. *69*, 14643–14649. https://doi.org/10.1021/acs.jafc.1c05868.

35. Schoenwaelder, M.E.A., and Wiencke, C. (2000). Phenolic compounds in the embryo development of several Northern Hemisphere fucoids. Plant Biol. *2*, 24–33. https://doi.org/10.1055/s-2000-9178.

36. Salgado, L.T., Cinelli, L.P., Viana, N.B., Tomazetto de Carvalho, R., De Souza Mourão, P.A., Teixeira, V.L., Farina, M., and Filho, A.G.M.A. (2009). A vanadium bromoperoxidase catalyzes the formation of high-molecular-weight complexes between brown algal phenolic substances and alginates(1). J. Phycol. *45*, 193–202. https://doi.org/10.1111/j.1529-8817.2008.00642.x.

37. Berglin, M., Delage, L., Potin, P., Vilter, H., and Elwing, H. (2004). Enzymatic cross-linking of a phenolic polymer extracted from the marine alga *Fucus serratus*. Biomacromolecules *5*, 2376–2383. https://doi.org/10.1021/bm0496864.

38. Bitton, R., Berglin, M., Elwing, H., Colin, C., Delage, L., Potin, P., and Bianco-Peled, H. (2007). The influence of halide-mediated oxidation on algae-born adhesives. Macromol. Biosci. *7*, 1280–1289. https://doi.org/10.1002/mabi.200700099.

39. Arnold, T.M., and Targett, N.M. (2003). To grow and defend: lack of tradeoffs for brown algal phlorotannins. Oikos *100*, 406–408. https://doi.org/10.1034/j.1600-0706.2003.11680.x.

40. Salgado, L.T., Tomazetto, R., Cinelli, L.P., Farina, M., and Amado Filho, G.M. (2007). The influence of brown algae alginates on phenolic compounds capability of ultraviolet radiation absorption in vitro. Braz. J. Oceanogr. *55*, 145–154. https://doi.org/10.1590/S1679-87592007000200007.

41. Fu, G., Nagasato, C., Oka, S., Cock, J.M., and Motomura, T. (2014). Proteomics analysis of heterogeneous flagella in brown algae (stramenopiles). Protist *165*, 662–675. https://doi.org/10.1016/j.protis.2014.07.007.

42. Kloareg, B., Badis, Y., Cock, J.M., and Michel, G. (2021). Role and evolution of the extracellular matrix in the acquisition of complex multicellularity in eukaryotes: A macroalgal perspective. Genes *12*, 1059. https://doi.org/10.3390/genes12071059.

43. Fan, X., Han, W., Teng, L., Jiang, P., Zhang, X., Xu, D., Li, C., Pellegrini, M., Wu, C., Wang, Y., et al. (2020). Single-base methylome profiling of the giant kelp *Saccharina japonica* reveals significant differences in DNA methylation to microalgae and plants. New Phytol. *225*, 234–249. https://doi.org/10.1111/nph.16125.

44. Knoll, A.H., Summons, R.E., Waldbauer, J.R., and Zumberge, J.E. (2007). Chapter 8. The geological succession of primary producers in the oceans. In Evolution of Primary Producers in the Sea, P.G. Falkowski and A.H. Knoll, eds. (Academic Press), pp. 133–163. https://doi.org/10.1016/B978-012370518-1/50009-6.

45. Schettino, A., and Turco, E. (2009). Breakup of Pangaea and plate kinematics of the central Atlantic and Atlas regions. Geophys. J. Int. *178*, 1078–1097. https://doi.org/10.1111/j.1365-246X.2009.04186.x.

46. Belcour, A., Got, J., Aite, M., Delage, L., Collén, J., Frioux, C., Leblanc, C., Dittami, S.M., Blanquart, S., Markov, G.V., et al. (2023). Inferring and comparing metabolism across heterogeneous sets of annotated genomes using AuCoMe. Genome Res. *33*, 972–987. https://doi.org/10.1101/gr.277056.122.

47. Coelho, S.M., Mignerot, L., and Cock, J.M. (2019). Origin and evolution of sex-determination systems in the brown algae. New Phytol. *222*, 1751–1756. https://doi.org/10.1111/nph.15694.

48. Coelho, S.M., Peters, A.F., Charrier, B., Roze, D., Destombe, C., Valero, M., and Cock, J.M. (2007). Complex life cycles of multicellular eukaryotes: new approaches based on the use of model organisms. Gene *406*, 152–170. https://doi.org/10.1016/j.gene.2007.07.025.

49. Kawai, H. (1992). A summary of the Morphology of Chloroplasts and Flagellated Cells in the Phaeophyceae. Algae *7*, 33–43.

50. Kinoshita, N., Nagasato, C., and Motomura, T. (2017). Phototaxis and chemotaxis of brown algal swarmers. J. Plant Res. *130*, 443–453. https://doi.org/10.1007/s10265-017-0914-8.

51. Fragkopoulou, E., Serrão, E.A., De Clerck, O., Costello, M.J., Araújo, M.B., Duarte, C.M., Krause-Jensen, D., and Assis, J. (2022). Global biodiversity patterns of marine forests of brown macroalgae. Glob. Ecol. Biogeogr. *31*, 636–648. https://doi.org/10.1111/geb.13450.

52. Vermeij, G.J., and Lindberg, D.R. (2000). Delayed herbivory and the assembly of marine benthic ecosystems. Paleobiology *26*, 419–430. https://doi.org/10.1666/0094-8373(2000)026<0419:DHATAO>2.0.CO;2.

53. Alfaro, M.E., Faircloth, B.C., Harrington, R.C., Sorenson, L., Friedman, M., Thacker, C.E., Oliveros, C.H., Černý, D., and Near, T.J. (2018). Explosive diversification of marine fishes at the Cretaceous-Palaeogene boundary. Nat. Ecol. Evol. *2*, 688–696. https://doi.org/10.1038/s41559-018-0494-6.

54. Starko, S., Soto Gomez, M., Darby, H., Demes, K.W., Kawai, H., Yotsukura, N., Lindstrom, S.C., Keeling, P.J., Graham, S.W., and Martone, P.T. (2019). A comprehensive kelp phylogeny sheds light on the evolution of an ecosystem. Mol. Phylogenet. Evol. *136*, 138–150. https://doi.org/10.1016/j.ympev.2019.04.012.

55. Kiel, S., Goedert, J.L., Huynh, T.L., Krings, M., Parkinson, D., Romero, R., and Looy, C.V. (2024). Early Oligocene kelp holdfasts and stepwise evolution of the kelp ecosystem in the North Pacific. Proc. Natl. Acad. Sci. USA *121*, e2317054121. https://doi.org/10.1073/pnas.2317054121.

56. Basson, P.W. (1981). Late Cretaceous alga, *Delesserites libanensis* sp. nov. Rev. Palaeobot. Palynol. *33*, 363–370. https://doi.org/10.1016/0034-6667(81)90093-2.

57. Krings, M., and Mayr, H. (2004). *Bassonia hakelensis* (Basson) nov. comb., a rare non-calcareous marine alga from the Cenomanian (Upper Cretaceous) of Lebanon. Zitteliana *44*, 105–111.

58. Barthel, K.W., and Swinburne, N.H.M. (1994). Solnhofen: a Study in Mesozoic Palaeontology (Cambridge University Press).

59. Martone, P.T., Kost, L., and Boller, M. (2012). Drag reduction in wave-swept macroalgae: alternative strategies and new predictions. Am. J. Bot. *99*, 806–815. https://doi.org/10.3732/ajb.1100541.

60. Colin, C., Leblanc, C., Michel, G., Wagner, E., Leize-Wagner, E., Van Dorsselaer, A., and Potin, P. (2005). Vanadium-dependent iodoperoxidases in *Laminaria digitata*, a novel biochemical function diverging from brown algal bromoperoxidases. J. Biol. Inorg. Chem. *10*, 156–166. https://doi.org/10.1007/s00775-005-0626-8.

61. Zambounis, A., Elias, M., Sterck, L., Maumus, F., and Gachon, C.M.M. (2012). Highly dynamic exon shuffling in candidate pathogen receptors... what if brown algae were capable of adaptive immunity? Mol. Biol. Evol. *29*, 1263–1276. https://doi.org/10.1093/molbev/msr296.

62. Macaisne, N., Liu, F., Scornet, D., Peters, A.F., Lipinska, A., Perrineau, M.-M., Henry, A., Strittmatter, M., Coelho, S.M., and Cock, J.M. (2017). The *Ectocarpus IMMEDIATE UPRIGHT* gene encodes a member of a novel family of cysteine-rich proteins with an unusual distribution across the eukaryotes. Development *144*, 409–418. https://doi.org/10.1242/dev.141523.

63. Suvorov, A., Kim, B.Y., Wang, J., Armstrong, E.E., Peede, D., D'Agostino, E.R.R., Price, D.K., Waddell, P.J., Lang, M., Courtier-Orgogozo, V., et al. (2022). Widespread introgression across a phylogeny of 155 Drosophila genomes. Curr. Biol. *32*, 111–123.e5. https://doi.org/10.1016/j.cub.2021.10.052.

64. Rohling, E.J., Yu, J., Heslop, D., Foster, G.L., Opdyke, B., and Roberts, A.P. (2021). Sea level and deep-sea temperature reconstructions suggest quasi-stable states and critical transitions over the past 40 million years. Sci. Adv. *7*, eabf5326. https://doi.org/10.1126/sciadv.abf5326.

65. Müller, D.G., and Knippers, R. (2011). Phaeovirus. In The Springer Index of Viruses, C. Tidona and G. Darai, eds. (Springer), pp. 1259–1263. https://doi.org/10.1007/978-0-387-95919-1_205.

66. McKeown, D.A., Stevens, K., Peters, A.F., Bond, P., Harper, G.M., Brownlee, C., Brown, M.T., and Schroeder, D.C. (2017). Phaeoviruses discovered in kelp (Laminariales). ISME J. *11*, 2869–2873. https://doi.org/10.1038/ismej.2017.130.

67. Moniruzzaman, M., Weinheimer, A.R., Martinez-Gutierrez, C.A., and Aylward, F.O. (2020). Widespread endogenization of giant viruses shapes genomes of green algae. Nature *588*, 141–145. https://doi.org/10.1038/s41586-020-2924-2.

68. Stevens, K., Weynberg, K., Bellas, C., Brown, S., Brownlee, C., Brown, M.T., and Schroeder, D.C. (2014). A novel evolutionary strategy revealed in the phaeoviruses. PLoS One *9*, e86040. https://doi.org/10.1371/journal.pone.0086040.

69. Bräutigam, M., Klein, M., Knippers, R., and Müller, D.G. (1995). Inheritance and meiotic elimination of a virus genome in the host *Ectocarpus siliculosus* (Phaeophyceae). J. Phycol. *31*, 823–827. https://doi.org/10.1111/j.0022-3646.1995.00823.x.

70. Keeling, P.J. (2024). Horizontal gene transfer in eukaryotes: aligning theory with data. Nat. Rev. Genet. *25*, 416–430. https://doi.org/10.1038/s41576-023-00688-5.

71. Kabbara, S., Hérivaux, A., Dugé de Bernonville, T., Courdavault, V., Clastre, M., Gastebois, A., Osman, M., Hamze, M., Cock, J.M., Schaap, P., et al. (2019). Diversity and evolution of sensor histidine kinases in eukaryotes. Genome Biol. Evol. *11*, 86–108. https://doi.org/10.1093/gbe/evy213.

72. Mazéas, L., Bouguerba-Collin, A., Cock, J.M., Denoeud, F., Godfroy, O., Brillet-Guéguen, L., Babbeyron, T., Lipinska, A.P., Delage, L., Corre, E., et al. (2024). Candidate genes involved in biosynthesis and degradation of the main extracellular matrix polysaccharides of brown algae and their probable evolutionary history. BMC Genom. *25*, 950. https://doi.org/10.1186/s12864-024-10811-3.

73. Potin, P., Bouarab, K., Salaün, J.-P., Pohnert, G., and Kloareg, B. (2002). Biotic interactions of marine algae. Curr. Opin. Plant Biol. *5*, 308–317. https://doi.org/10.1016/s1369-5266(02)00273-x.

74. Tarakhovskaya, E.R. (2014). Mechanisms of bioadhesion of macrophytic algae. Russ. J. Plant Physiol. *61*, 19–25. https://doi.org/10.1134/S1021443714010154.

75. Domazet-Lošo, M., Široki, T., Šimičević, K., and Domazet-Lošo, T. (2024). Macroevolutionary dynamics of gene family gain and loss along multicellular eukaryotic lineages. Nat. Commun. *15*, 2663. https://doi.org/10.1038/s41467-024-47017-w.

76. Holmes, E.C. (2011). The evolution of endogenous viral elements. Cell Host Microbe *10*, 368–377. https://doi.org/10.1016/j.chom.2011.09.002.

77. Feschotte, C., and Gilbert, C. (2012). Endogenous viruses: insights into viral evolution and impact on host biology. Nat. Rev. Genet. *13*, 283–296. https://doi.org/10.1038/nrg3199.

78. Moniruzzaman, M., Erazo-Garcia, M.P., and Aylward, F.O. (2022). Endogenous giant viruses contribute to intraspecies genomic variability in the model green alga *Chlamydomonas reinhardtii*. Virus Evol. *8*, veac102. https://doi.org/10.1093/ve/veac102.

79. Badis, Y., Scornet, D., Harada, M., Caillard, C., Godfroy, O., Raphalen, M., Gachon, C.M.M., Coelho, S.M., Motomura, T., Nagasato, C., et al. (2021). Targeted CRISPR-Cas9-based gene knockouts in the model brown alga *Ectocarpus*. New Phytol. *231*, 2077–2091. https://doi.org/10.1111/nph.17525.

80. Shen, Y., Motomura, T., Ichihara, K., Matsuda, Y., Yoshimura, K., Kosugi, C., and Nagasato, C. (2023). Application of CRISPR-Cas9 genome editing by microinjection of gametophytes of *Saccharina japonica* (Laminariales, Phaeophyceae). J. Appl. Phycol. *35*, 1431–1441. https://doi.org/10.1007/s10811-023-02940-1.

81. Cock, J.M. (2023). The model system *Ectocarpus*: integrating functional genomics into brown algal research. J. Phycol. *59*, 4–8. https://doi.org/10.1111/jpy.13310.

82. Li, D., Liu, C.-M., Luo, R., Sadakane, K., and Lam, T.-W. (2015). MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. Bioinformatics *31*, 1674–1676. https://doi.org/10.1093/bioinformatics/btv033.

83. Noguchi, H., Park, J., and Takagi, T. (2006). MetaGene: prokaryotic gene finding from environmental genome shotgun sequences. Nucleic Acids Res. *34*, 5623–5630. https://doi.org/10.1093/nar/gkl723.

84. Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. *25*, 3389–3402. https://doi.org/10.1093/nar/25.17.3389.

85. Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics *25*, 1754–1760. https://doi.org/10.1093/bioinformatics/btp324.

86. Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nat. Methods *9*, 357–359. https://doi.org/10.1038/nmeth.1923.

87. Bankevich, A., Nurk, S., Antipov, D., Gurevich, A.A., Dvorkin, M., Kulikov, A.S., Lesin, V.M., Nikolenko, S.I., Pham, S., Prjibelski, A.D., et al. (2012). SPAdes: A new genome assembly algorithm and its applications to single-cell sequencing. J. Comput. Biol. *19*, 455–477. https://doi.org/10.1089/cmb.2012.0021.

88. Liu, H., Wu, S., Li, A., Ruan, J., Wu, S., Li, A., and Ruan, J. (2021). SMARTdenovo: a de novo assembler using long noisy reads. Gigabyte *2021*, 1–9. https://doi.org/10.46471/gigabyte.15.

89. Ruan, J., and Li, H. (2020). Fast and accurate long-read assembly with wtdbg2. Nat. Methods *17*, 155–158. https://doi.org/10.1038/s41592-019-0669-3.

90. Kolmogorov, M., Yuan, J., Lin, Y., and Pevzner, P.A. (2019). Assembly of long, error-prone reads using repeat graphs. Nat. Biotechnol. *37*, 540–546. https://doi.org/10.1038/s41587-019-0072-8.

91. Chen, Y., Nie, F., Xie, S.-Q., Zheng, Y.-F., Dai, Q., Bray, T., Wang, Y.-X., Xing, J.-F., Huang, Z.-J., Wang, D.-P., et al. (2021). Efficient assembly of nanopore reads via highly accurate and intact error correction. Nat. Commun. *12*, 60. https://doi.org/10.1038/s41467-020-20236-7.

92. Vaser, R., Sović, I., Nagarajan, N., and Šikić, M. (2017). Fast and accurate de novo genome assembly from long uncorrected reads. Genome Res. *27*, 737–746. https://doi.org/10.1101/gr.214270.116.

93. Aury, J.-M., and Istace, B. (2021). Hapo-G, haplotype-aware polishing of genome assemblies with accurate reads. NAR Genom. Bioinform. *3*, lqab034. https://doi.org/10.1093/nargab/lqab034.

94. Kang, D.D., Li, F., Kirton, E., Thomas, A., Egan, R., An, H., and Wang, Z. (2019). MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. PeerJ *7*, e7359. https://doi.org/10.7717/peerj.7359.

95. Kopylova, E., Noé, L., and Touzet, H. (2012). SortMeRNA: fast and accurate filtering of ribosomal RNAs in metatranscriptomic data. Bioinformatics *28*, 3211–3217. https://doi.org/10.1093/bioinformatics/bts611.

96. Zerbino, D.R., and Birney, E. (2008). Velvet: algorithms for *de novo* short read assembly using de Bruijn graphs. Genome Res. *18*, 821–829. https://doi.org/10.1101/gr.074492.107.

97. Schulz, M.H., Zerbino, D.R., Vingron, M., and Birney, E. (2012). Oases: robust *de novo* RNA-seq assembly across the dynamic range of expression levels. Bioinformatics *28*, 1086–1092. https://doi.org/10.1093/bioinformatics/bts094.

98. Marchler-Bauer, A., Derbyshire, M.K., Gonzales, N.R., Lu, S., Chitsaz, F., Geer, L.Y., Geer, R.C., He, J., Gwadz, M., Hurwitz, D.I., et al. (2015). CDD: NCBI's conserved domain database. Nucleic Acids Res. *43*, D222–D226. https://doi.org/10.1093/nar/gku1221.

99. Bolger, A.M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics *30*, 2114–2120. https://doi.org/10.1093/bioinformatics/btu170.

100. Grabherr, M.G., Haas, B.J., Yassour, M., Levin, J.Z., Thompson, D.A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., et al. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat. Biotechnol. *29*, 644–652. https://doi.org/10.1038/nbt.1883.

101. Bushmanova, E., Antipov, D., Lapidus, A., and Prjibelski, A.D. (2019). rnaSPAdes: a *de novo* transcriptome assembler and its application to RNA-Seq data. GigaScience *8*, giz100. https://doi.org/10.1093/gigascience/giz100.

102. Smit, A.F.A., Hubley, R., and Green, P. RepeatMasker. http://repeatmasker.org.

103. Benson, G. (1999). Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res. *27*, 573–580. https://doi.org/10.1093/nar/27.2.573.

104. Flutre, T., Duprat, E., Feuillet, C., and Quesneville, H. (2011). Considering transposable element diversification in *de novo* annotation approaches. PLoS One *6*, e16526. https://doi.org/10.1371/journal.pone.0016526.

105. Kent, W.J. (2002). BLAT–the BLAST-like alignment tool. Genome Res. *12*, 656–664. https://doi.org/10.1101/gr.229202.

106. Birney, E., Clamp, M., and Durbin, R. (2004). GeneWise and Genomewise. Genome Res. *14*, 988–995. https://doi.org/10.1101/gr.1865504.

107. Buchfink, B., Reuter, K., and Drost, H.-G. (2021). Sensitive protein alignments at tree-of-life scale using DIAMOND. Nat. Methods *18*, 366–368. https://doi.org/10.1038/s41592-021-01101-x.

108. Mott, R. (1997). EST_GENOME: a program to align spliced DNA sequences to unspliced genomic DNA. Comput. Appl. Biosci. *13*, 477–478. https://doi.org/10.1093/bioinformatics/13.4.477.

109. Dubarry, M., Noel, B., Rukwavu, T., Farhat, S., Silva, C.D., Seeleuthner, Y., Lebeurrier, M., and Aury, J.-M. (2016). Gmove a Tool for Eukaryotic Gene Predictions Using Various Evidences. F1000Research *5*. https://doi.org/10.7490/f1000research.1111735.1.

110. Larsson, A. (2014). AliView: a fast and lightweight alignment viewer and editor for large datasets. Bioinformatics *30*, 3276–3278. https://doi.org/10.1093/bioinformatics/btu531.

111. Stamatakis, A. (2015). Using RAxML to infer phylogenies. Curr. Protoc. Bioinformatics *51*, 6.14.1–6.14.14. https://doi.org/10.1002/0471250953.bi0614s51.

112. Rambaut, A., Drummond, A.J., Xie, D., Baele, G., and Suchard, M.A. (2018). Posterior summarization in Bayesian phylogenetics using Tracer 1.7. Syst. Biol. *67*, 901–904. https://doi.org/10.1093/sysbio/syy032.

113. Emms, D.M., and Kelly, S. (2019). OrthoFinder: phylogenetic orthology inference for comparative genomics. Genome Biol. *20*, 238. https://doi.org/10.1186/s13059-019-1832-y.

114. Csurös, M. (2010). Count: evolutionary analysis of phylogenetic profiles with parsimony and likelihood. Bioinformatics *26*, 1910–1912. https://doi.org/10.1093/bioinformatics/btq315.

115. Edgar, R.C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. *32*, 1792–1797. https://doi.org/10.1093/nar/gkh340.

116. Jehl, P., Sievers, F., and Higgins, D.G. (2015). OD-seq: outlier detection in multiple sequence alignments. BMC Bioinformatics *16*, 269. https://doi.org/10.1186/s12859-015-0702-1.

117. Mistry, J., Finn, R.D., Eddy, S.R., Bateman, A., and Punta, M. (2013). Challenges in homology search: HMMER3 and convergent evolution of coiled-coil regions. Nucleic Acids Res. *41*, e121. https://doi.org/10.1093/nar/gkt263.

118. Barrera-Redondo, J., Lotharukpong, J.S., Drost, H.-G., and Coelho, S.M. (2023). Uncovering gene-family founder events during major evolutionary transitions in animals, plants and fungi using GenEra. Genome Biol. *24*, 54. https://doi.org/10.1186/s13059-023-02895-z.

119. Enright, A.J., Van Dongen, S., and Ouzounis, C.A. (2002). An efficient algorithm for large-scale detection of protein families. Nucleic Acids Res. *30*, 1575–1584. https://doi.org/10.1093/nar/30.7.1575.

120. van Kempen, M., Kim, S.S., Tumescheit, C., Mirdita, M., Lee, J., Gilchrist, C.L.M., Söding, J., and Steinegger, M. (2023). Fast and accurate protein structure search with Foldseek. Preprint at bioRxiv. https://doi.org/10.1101/2022.02.07.479398.

121. Pathmanathan, J.S., Lopez, P., Lapointe, F.-J., and Bapteste, E. (2018). CompositeSearch: A generalized network approach for composite gene families detection. Mol. Biol. Evol. *35*, 252–255. https://doi.org/10.1093/molbev/msx283.

122. Overbeek, R., Begley, T., Butler, R.M., Choudhuri, J.V., Chuang, H.-Y., Cohoon, M., de Crécy-Lagard, V., Diaz, N., Disz, T., Edwards, R., et al. (2005). The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. Nucleic Acids Res. *33*, 5691–5702. https://doi.org/10.1093/nar/gki866.

123. Paysan-Lafosse, T., Blum, M., Chuguransky, S., Grego, T., Pinto, B.L., Salazar, G.A., Bileschi, M.L., Bork, P., Bridge, A., Colwell, L., et al. (2023). InterPro in 2022. Nucleic Acids Res. *51*, D418–D427. https://doi.org/10.1093/nar/gkac993.

124. Huerta-Cepas, J., Szklarczyk, D., Heller, D., Hernández-Plaza, A., Forslund, S.K., Cook, H., Mende, D.R., Letunic, I., Rattei, T., Jensen, L.J., et al. (2019). eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. Nucleic Acids Res. *47*, D309–D314. https://doi.org/10.1093/nar/gky1085.

125. Cantalapiedra, C.P., Hernández-Plaza, A., Letunic, I., Bork, P., and Huerta-Cepas, J. (2021). eggNOG-mapper v2: functional annotation, orthology assignments, and domain prediction at the metagenomic scale. Mol. Biol. Evol. *38*, 5825–5829. https://doi.org/10.1093/molbev/msab293.

126. Hyatt, D., Chen, G.-L., Locascio, P.F., Land, M.L., Larimer, F.W., and Hauser, L.J. (2010). Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics *11*, 119. https://doi.org/10.1186/1471-2105-11-119.

127. Aylward, F.O., and Moniruzzaman, M. (2021). ViralRecall-A flexible command-line tool for the detection of giant virus signatures in 'Omic data. Viruses *13*, 150. https://doi.org/10.3390/v13020150.

128. Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics *26*, 841–842. https://doi.org/10.1093/bioinformatics/btq033.

129. Hauser, M., Steinegger, M., and Söding, J. (2016). MMseqs software suite for fast and deep clustering and searching of large protein sequence sets. Bioinformatics *32*, 1323–1330. https://doi.org/10.1093/bioinformatics/btw006.

130. Katoh, K., and Standley, D.M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol. Biol. Evol. *30*, 772–780. https://doi.org/10.1093/molbev/mst010.

131. Tamura, K., Peterson, D., Peterson, N., Stecher, G., Nei, M., and Kumar, S. (2011). MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Mol. Biol. Evol. *28*, 2731–2739. https://doi.org/10.1093/molbev/msr121.

132. Lemoine, F., Correia, D., Lefort, V., Doppelt-Azeroual, O., Mareuil, F., Cohen-Boulakia, S., and Gascuel, O. (2019). NGPhylogeny.fr: new

generation phylogenetic services for non-specialists. Nucleic Acids Res. *47*, W260–W265. https://doi.org/10.1093/nar/gkz303.

133. Capella-Gutiérrez, S., Silla-Martínez, J.M., and Gabaldón, T. (2009). trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics *25*, 1972–1973. https://doi.org/10.1093/bioinformatics/btp348.

134. Petroll, R., Schreiber, M., Finke, H., Cock, J.M., Gould, S.B., and Rensing, S.A. (2021). Signatures of transcription factor evolution and the secondary gain of red algae complexity. Genes *12*, 1055. https://doi.org/10.3390/genes12071055.

135. Petroll, R., Varshney, D., Hiltemann, S., Finke, H., Schreiber, M., de Vries, J., and Rensing, S.A. (2024). Enhanced sensitivity of TAPscan v4 enables comprehensive analysis of streptophyte transcription factor evolution. Preprint at bioRxiv. https://doi.org/10.1101/2024.07.13.602682.

136. Duvaud, S., Gabella, C., Lisacek, F., Stockinger, H., Ioannidis, V., and Durinx, C. (2021). Expasy, the Swiss Bioinformatics Resource Portal, as designed by its users. Nucleic Acids Res. *49*, W216–W227. https://doi.org/10.1093/nar/gkab225.

137. Jones, P., Binns, D., Chang, H.-Y., Fraser, M., Li, W., McAnulla, C., McWilliam, H., Maslen, J., Mitchell, A., Nuka, G., et al. (2014). InterProScan 5: genome-scale protein function classification. Bioinformatics *30*, 1236–1240. https://doi.org/10.1093/bioinformatics/btu031.

138. Thompson, J.D., Higgins, D.G., and Gibson, T.J. (1994). Clustal W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. *22*, 4673–4680. https://doi.org/10.1093/nar/22.22.4673.

139. Castresana, J. (2000). Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Mol. Biol. Evol. *17*, 540–552. https://doi.org/10.1093/oxfordjournals.molbev.a026334.

140. Bray, N.L., Pimentel, H., Melsted, P., and Pachter, L. (2016). Near-optimal probabilistic RNA-seq quantification. Nat. Biotechnol. *34*, 525–527. https://doi.org/10.1038/nbt.3519.

141. Love, M.I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. *15*, 550. https://doi.org/10.1186/s13059-014-0550-8.

142. Andrews, S. (2016). FastQC A Quality Control tool for High Throughput Sequence Data. http://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

143. Krueger, F. (2015). Trim Galore!: A Wrapper around Cutadapt and FastQC to Consistently Apply Adapter and Quality Trimming to FastQ Files, with Extra Functionality for RRBS Data (Babraham Institute).

144. Kim, D., Langmead, B., and Salzberg, S.L. (2015). HISAT: a fast spliced aligner with low memory requirements. Nat. Methods *12*, 357–360. https://doi.org/10.1038/nmeth.3317.

145. Liao, Y., Smyth, G.K., and Shi, W. (2014). featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics *30*, 923–930. https://doi.org/10.1093/bioinformatics/btt656.

146. Yang, Z. (2007). PAML 4: phylogenetic analysis by maximum likelihood. Mol. Biol. Evol. *24*, 1586–1591. https://doi.org/10.1093/molbev/msm088.

147. Revell, L.J. (2012). phytools: an R package for phylogenetic comparative biology (and other things). Methods Ecol. Evol. *3*, 217–223. https://doi.org/10.1111/j.2041-210X.2011.00169.x.

148. Wallau, G.L., Capy, P., Loreto, E., Le Rouzic, A., and Hua-Van, A. (2016). VHICA, a new method to discriminate between vertical and horizontal transposon transfer: application to the mariner family within *Drosophila*. Mol. Biol. Evol. *33*, 1094–1109. https://doi.org/10.1093/molbev/msv341.

149. Dierckxsens, N., Mardulyn, P., and Smits, G. (2017). NOVOPlasty: de novo assembly of organelle genomes from whole genome data. Nucleic Acids Res. *45*, e18. https://doi.org/10.1093/nar/gkw955.

150. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R.; 1000 Genome Project Data Processing Subgroup (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078–2079. https://doi.org/10.1093/bioinformatics/btp352.

151. Tillich, M., Lehwark, P., Pellizzer, T., Ulbricht-Jones, E.S., Fischer, A., Bock, R., and Greiner, S. (2017). GeSeq – versatile and accurate annotation of organelle genomes. Nucleic Acids Res. *45*, W6–W11. https://doi.org/10.1093/nar/gkx391.

152. Laslett, D., and Canback, B. (2004). ARAGORN, a program to detect tRNA genes and tmRNA genes in nucleotide sequences. Nucleic Acids Res. *32*, 11–16. https://doi.org/10.1093/nar/gkh152.

153. Kalyaanamoorthy, S., Minh, B.Q., Wong, T.K.F., von Haeseler, A., and Jermiin, L.S. (2017). ModelFinder: fast model selection for accurate phylogenetic estimates. Nat. Methods *14*, 587–589. https://doi.org/10.1038/nmeth.4285.

154. Hoang, D.T., Chernomor, O., von Haeseler, A., Minh, B.Q., and Vinh, L.S. (2018). UFBoot2: improving the ultrafast bootstrap approximation. Mol. Biol. Evol. *35*, 518–522. https://doi.org/10.1093/molbev/msx281.

155. Haug-Baltzell, A., Stephens, S.A., Davey, S., Scheidegger, C.E., and Lyons, E. (2017). SynMap2 and SynMap3D: web-based whole-genome synteny browsers. Bioinformatics *33*, 2197–2198. https://doi.org/10.1093/bioinformatics/btx144.

156. Haas, B.J., Delcher, A.L., Wortman, J.R., and Salzberg, S.L. (2004). DAGchainer: a tool for mining segmental genome duplications and synteny. Bioinformatics *20*, 3643–3646. https://doi.org/10.1093/bioinformatics/bth397.

157. Bouckaert, R., Heled, J., Kühnert, D., Vaughan, T., Wu, C.-H., Xie, D., Suchard, M.A., Rambaut, A., and Drummond, A.J. (2014). BEAST 2: a software platform for Bayesian evolutionary analysis. PLoS Comput. Biol. *10*, e1003537. https://doi.org/10.1371/journal.pcbi.1003537.

158. Douglas, J., Jiménez-Silva, C.L., and Bouckaert, R. (2022). StarBeast3: adaptive parallelized Bayesian inference under the multispecies coalescent. Syst. Biol. *71*, 901–916. https://doi.org/10.1093/sysbio/syac010.

159. Bouckaert, R.R., and Drummond, A.J. (2017). bModelTest: Bayesian phylogenetic site model averaging and model comparison. BMC Evol. Biol. *17*, 42. https://doi.org/10.1186/s12862-017-0890-6.

160. Kloepper, T.H., and Huson, D.H. (2008). Drawing explicit phylogenetic networks and their integration into SplitsTree. BMC Evol. Biol. *8*, 22. https://doi.org/10.1186/1471-2148-8-22.

161. Gschloessl, B., Guermeur, Y., and Cock, J. (2008). HECTAR: a method to predict subcellular targeting in heterokonts. BMC Bioinf. *9*, 393.

162. R Core Team (2018). R: A Language and Environment for Statistical Computing (R Foundation for Statistical Computing).

163. Minh, B.Q., Schmidt, H.A., Chernomor, O., Schrempf, D., Woodhams, M.D., von Haeseler, A., and Lanfear, R. (2020). IQ-TREE 2: New models and efficient methods for phylogenetic inference in the genomic era. Mol. Biol. Evol. *37*, 1530–1534. https://doi.org/10.1093/molbev/msaa015.

164. Mendes, F.K., Vanderpool, D., Fulton, B., and Hahn, M.W. (2021). CAFE 5 models variation in evolutionary rates among gene families. Bioinformatics *36*, 5516–5518. https://doi.org/10.1093/bioinformatics/btaa1022.

165. Yu, G., Wang, L.-G., Han, Y., and He, Q.-Y. (2012). clusterProfiler: an R package for comparing biological themes among gene clusters. Omics *16*, 284–287. https://doi.org/10.1089/omi.2011.0118.

166. Wickham, H., Chang, W., and Wickham, M.H. (2016). Package 'ggplot2.' Create Elegant Data Visualisations Using the Grammar of Graphics, version 2, pp. 1–189.

167. Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L.D., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., et al. (2019). Welcome to the tidyverse. J. Open Source Software *4*, 1686. https://doi.org/10.21105/joss.01686.

168. Manni, M., Berkeley, M.R., Seppey, M., Simão, F.A., and Zdobnov, E.M. (2021). BUSCO update: novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic,

prokaryotic, and viral genomes. Mol. Biol. Evol. *38*, 4647–4654. https://doi.org/10.1093/molbev/msab199.

169. Suzek, B.E., Wang, Y., Huang, H., McGarvey, P.B., and Wu, C.H.; UniProt Consortium (2015). UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. Bioinformatics *31*, 926–932. https://doi.org/10.1093/bioinformatics/btu739.

170. Varadi, M., Anyango, S., Deshpande, M., Nair, S., Natassia, C., Yordanova, G., Yuan, D., Stroe, O., Wood, G., Laydon, A., et al. (2022). AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. Nucleic Acids Res. *50*, D439–D444. https://doi.org/10.1093/nar/gkab1061.

171. Yutin, N., Wolf, Y.I., Raoult, D., and Koonin, E.V. (2009). Eukaryotic large nucleo-cytoplasmic DNA viruses: clusters of orthologous genes and reconstruction of viral genome evolution. Virol. J. *6*, 223. https://doi.org/10.1186/1743-422X-6-223.

172. Trgovec-Greif, L., Hellinger, H.-J., Mainguy, J., Pfundner, A., Frishman, D., Kiening, M., Webster, N.S., Laffy, P.W., Feichtinger, M., and Rattei, T. (2024). VOGDB—database of virus orthologous groups. Viruses *16*, 1191. https://doi.org/10.3390/v16081191.

173. Barbeyron, T., Brillet-Guéguen, L., Carré, W., Carrière, C., Caron, C., Czjzek, M., Hoebeke, M., and Michel, G. (2016). Matching the diversity of sulfated biomolecules: creation of a classification database for sulfatases reflecting their substrate specificity. PLoS One *11*, e0164846. https://doi.org/10.1371/journal.pone.0164846.

174. Stam, M., Lelièvre, P., Hoebeke, M., Corre, E., Barbeyron, T., and Michel, G. (2023). SulfAtlas, the sulfatase database: state of the art and new developments. Nucleic Acids Res. *51*, D647–D653. https://doi.org/10.1093/nar/gkac977.

175. Mistry, J., Chuguransky, S., Williams, L., Qureshi, M., Salazar, G.A., Sonnhammer, E.L.L., Tosatto, S.C.E., Paladin, L., Raj, S., Richardson, L.J., et al. (2021). Pfam: the protein families database in 2021. Nucleic Acids Res. *49*, D412–D419. https://doi.org/10.1093/nar/gkaa913.

176. Thomas, P.D., Campbell, M.J., Kejariwal, A., Mi, H., Karlak, B., Daverman, R., Diemer, K., Muruganujan, A., and Narechania, A. (2003). Panther: a library of protein families and subfamilies indexed by function. Genome Res. *13*, 2129–2141. https://doi.org/10.1101/gr.772403.

177. Letunic, I., Khedkar, S., and Bork, P. (2021). SMART: recent updates, new developments and status in 2020. Nucleic Acids Res. *49*, D458–D460. https://doi.org/10.1093/nar/gkaa937.

178. Schlösser, U.G. (1994). SAG - Sammlung von Algenkulturen at the university of Göttingen catalogue of strains 1994. Bot. Acta *107*, 113–186. https://doi.org/10.1111/j.1438-8677.1994.tb00784.x.

179. Cormier, A., Avia, K., Sterck, L., Derrien, T., Wucher, V., Andres, G., Monsoor, M., Godfroy, O., Lipinska, A., Perrineau, M.-M., et al. (2017). Re-annotation, improved large-scale assembly and establishment of a catalogue of noncoding loci for the genome of the model brown alga *Ectocarpus*. New Phytol. *214*, 219–232. https://doi.org/10.1111/nph.14321.

180. Debit, A., Vincens, P., Bowler, C., and de Carvalho, H.C. (2023). LncPlankton, V1.0: a comprehensive collection of plankton long noncoding RNAs. Preprint at bioRxiv. https://doi.org/10.1101/2023.11.03.565479.

181. Parker, B.C. (1965). Non-calcareous marine algae from California Miocene deposits. Nova Hedwigia *10*, 273.

182. Akita, S., Vieira, C., Hanyuda, T., Rousseau, F., Cruaud, C., Couloux, A., Heesch, S., Cock, J.M., and Kawai, H. (2022). Providing a phylogenetic framework for trait-based analyses in brown algae: phylogenomic tree inferred from 32 nuclear protein-coding sequences. Mol. Phylogenet. Evol. *168*, 107408. https://doi.org/10.1016/j.ympev.2022.107408.

183. Weisman, C.M., Murray, A.W., and Eddy, S.R. (2020). Many, but not all, lineage-specific genes can be explained by homology detection failure. PLoS Biol. *18*, e3000862. https://doi.org/10.1371/journal.pbio.3000862.

184. Mulhair, P.O., Moran, R.J., Pathmanathan, J.S., Sussfeld, D., Creevey, C.J., Siu-Ting, K., Whelan, F.J., Pisani, D., Constantinides, B., Pelletier, E., et al. (2023). Bursts of novel composite gene families at major nodes in animal evolution. Preprint at bioRxiv. https://doi.org/10.1101/2023.07.10.548381.

185. Goodacre, N., Aljanahi, A., Nandakumar, S., Mikailov, M., and Khan, A.S. (2018). A reference viral database (RVDB) to enhance bioinformatics analysis of high-throughput sequencing for novel virus detection. mSphere *3*, e00069-18. https://doi.org/10.1128/mSphereDirect.00069-18.

186. Maumus, F., Epert, A., Nogué, F., and Blanc, G. (2014). Plant genomes enclose footprints of past infections by giant virus relatives. Nat. Commun. *5*, 4268. https://doi.org/10.1038/ncomms5268.

187. Caspi, R., Billington, R., Fulcher, C.A., Keseler, I.M., Kothari, A., Krummenacker, M., Latendresse, M., Midford, P.E., Ong, Q., Ong, W.K., et al. (2018). The MetaCyc database of metabolic pathways and enzymes. Nucleic Acids Res. *46*, D633–D639. https://doi.org/10.1093/nar/gkx935.

188. Aite, M., Chevallier, M., Frioux, C., Trottier, C., Got, J., Cortés, M.P., Mendoza, S.N., Carrier, G., Dameron, O., Guillaudeux, N., et al. (2018). Traceability, reproducibility and Wiki-exploration for "à-la-carte" reconstructions of genome-scale metabolic models. PLoS Comput. Biol. *14*, e1006146. https://doi.org/10.1371/journal.pcbi.1006146.

189. Talbert, P.B., Ahmad, K., Almouzni, G., Ausió, J., Berger, F., Bhalla, P.L., Bonner, W.M., Cande, W.Z., Chadwick, B.P., Chan, S.W.L., et al. (2012). A unified phylogeny-based nomenclature for histone variants. Epigenetics Chromatin *5*, 7. https://doi.org/10.1186/1756-8935-5-7.

190. Starr, R.C., and Zeikus, J.A. (1993). UTEX-The culture collection of algae at the University of Texas at Austin 1993 list of cultures. J. Phycol. *29*, 1–106. https://doi.org/10.1111/j.0022-3646.1993.00001.x.

191. Dittami, S.M., Gravot, A., Goulitquer, S., Rousvoal, S., Peters, A.F., Bouchereau, A., Boyen, C., and Tonon, T. (2012). Towards deciphering dynamic changes and evolutionary mechanisms involved in the adaptation to low salinities in *Ectocarpus* (brown algae). Plant J. *71*, 366–377. https://doi.org/10.1111/j.1365-313X.2012.04982.x.

192. Rahman, S., Kosakovsky Pond, S.L., Webb, A., and Hey, J. (2021). Weak selection on synonymous codons substantially inflates dN/dS estimates in bacteria. Proc. Natl. Acad. Sci. USA *118*, e2023575118. https://doi.org/10.1073/pnas.2023575118.

193. Duchêne, S., Holmes, E.C., and Ho, S.Y.W. (2014). Analyses of evolutionary dynamics in viruses are hindered by a time-dependent bias in rate estimates. Proc. Biol. Sci. *281*, 20140732. https://doi.org/10.1098/rspb.2014.0732.

194. Akashi, H. (1995). Inferring weak selection from patterns of polymorphism and divergence at "silent" sites in *Drosophila* DNA. Genetics *139*, 1067–1076. https://doi.org/10.1093/genetics/139.2.1067.

195. Akashi, H. (1997). Codon bias evolution in *Drosophila*. Population genetics of mutation-selection drift. Gene *205*, 269–278. https://doi.org/10.1016/s0378-1119(97)00400-9.

196. Subramanian, S. (2008). Nearly neutrality and the evolution of codon usage bias in eukaryotic genomes. Genetics *178*, 2429–2432. https://doi.org/10.1534/genetics.107.086405.

197. Wright, F. (1990). The 'effective number of codons' used in a gene. Gene *87*, 23–29. https://doi.org/10.1016/0378-1119(90)90491-9.

198. Forcelloni, S., and Giansanti, A. (2020). Evolutionary forces and codon bias in different flavors of intrinsic disorder in the human proteome. J. Mol. Evol. *88*, 164–178. https://doi.org/10.1007/s00239-019-09921-4.

199. Kiełbasa, S.M., Wan, R., Sato, K., Horton, P., and Frith, M.C. (2011). Adaptive seeds tame genomic sequence comparison. Genome Res. *21*, 487–493. https://doi.org/10.1101/gr.113985.110.

200. Madeira, F., Pearce, M., Tivey, A.R.N., Basutkar, P., Lee, J., Edbali, O., Madhusoodanan, N., Kolesnikov, A., and Lopez, R. (2022). Search and sequence analysis tools services from EMBL-EBI in 2022. Nucleic Acids Res. *50*, W276–W279. https://doi.org/10.1093/nar/gkac240.

201. Steffen, R., Ogoniak, L., Grundmann, N., Pawluchin, A., Soehnlein, O., and Schmitz, J. (2022). paPAML: an improved computational tool to explore selection pressure on protein-coding sequences. Genes *13*, 1090. https://doi.org/10.3390/genes13061090.

202. Criscuolo, A., and Gribaldo, S. (2010). BMGE (Block Mapping and Gathering with Entropy): a new software for selection of phylogenetic informative regions from multiple sequence alignments. BMC Evol. Biol. *10*, 210. https://doi.org/10.1186/1471-2148-10-210.

203. Green, R.E., Krause, J., Briggs, A.W., Maricic, T., Stenzel, U., Kircher, M., Patterson, N., Li, H., Zhai, W., Fritz, M.H.-Y., et al. (2010). A draft sequence of the Neandertal genome. Science *328*, 710–722. https://doi.org/10.1126/science.1188021.

204. Durand, E.Y., Patterson, N., Reich, D., and Slatkin, M. (2011). Testing for ancient admixture between closely related populations. Mol. Biol. Evol. *28*, 2239–2252. https://doi.org/10.1093/molbev/msr048.

205. 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*, 1065–1093. https://doi.org/10.1534/genetics.112.145037.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Biological samples		
Descriptions of all sequenced samples have been deposited in the EBI/ENA database	This study.	EBI/ENA project PRJEB76691
Critical commercial assays		
OmniPrep Genomic DNA Purification Kit	G Biosciences, St. Louis, MO, USA	N/A
Nucleospin Plant II midi DNA Extraction Kit	Macherey-Nagel, Düren, Germany	N/A
NEBNext DNA Modules Products	New England Biolabs, Ipswich, MA, USA	N/A
NEBNext Sample Reagent Set	New England Biolabs, Ipswich, MA, USA	N/A
Ampure XP	Beckmann Coulter Genomics, Danvers, MA, USA	N/A
Kapa Hifi Hotstart NGS library Amplification kit	Roche, Basel, Switzerland	N/A
Short Read Eliminator Kit	Pacific Biosciences, Menlo Park, CA, USA	N/A
1D Genomic DNA by Ligation	Oxford Nanopore Technologies Ltd, Oxford, UK	SQK-LSK109, SQK-LSK108 or SQK-LSK110
Qiagen RNeasy kit or the Macherey Nagel RNAplus kit	Macherey-Nagel, Düren, Germany	N/A
TruSeq Stranded mRNA Sample Prep	Illumina	N/A
NEBNext Ultra II Directional RNA Library Prep for Illumina	New England BioLabs	N/A
Deposited data		
The sequence data generated by this project is described in Table S1.	This study.	EBI/ENA: PRJEB76691
CNRS Research Data dataset "Data for Phaeoexplorer publication: Evolutionary genomics of the emergence of brown algae as key components of coastal ecosystems"	This study.	CNRS Research Data: https://doi.org/10.57745/9U1J85
Experimental models: Organisms/strains		
The strains used for genome and transcriptome sequencing are listed in Table S1A.	Culture collection references are provided where relevant.	See strain names and culture collection accessions for identifiers.
Software and algorithms		
MEGAHIT version 1.1.1	Li et al.[82]	RRID:SCR_018551 https://github.com/voutcn/megahit
MetaGene version 2008.8.19	Noguchi et al.[83]	http://metagene.cb.k.u-tokyo.ac.jp/
BLAST	Altschul et al.[84]	RRID:SCR_004870 http://blast.ncbi.nlm.nih.gov/Blast.cgi
Burrows-Wheeler Aligner	Li and Durbin[85]	RRID:SCR_010910 http://bio-bwa.sourceforge.net/
Bowtie2 version 2.3.5.1	Langmead and Salzberg[86]	RRID:SCR_016368 http://bowtie-bio.sourceforge.net/bowtie2/index.shtml
SPAdes assembler version 3.8.1	Bankevich et al.[87]	RRID:SCR_000131 https://cab.spbu.ru/software/spades/
filtlong	Wick, R.	RRID:SCR_024020 https://github.com/rrwick/Filtlong
Smartdenovo	Liu et al.[88]	RRID:SCR_017622 https://github.com/ruanjue/smartdenovo

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Redbean	Ruan and Li[89]	N/A
Flye	Kolmogorov et al.[90]	RRID:SCR_017016 https://github.com/fenderglass/Flye
Necat	Chen et al.[91]	https://github.com/xiaochuanle/necat
Racon	Vaser et al.[92]	RRID:SCR_017642 https://github.com/isovic/racon
Hapo-G	Aury et al.[93]	https://www.genoscope.cns.fr/hapog/
Metabat 2	Kang et al.[94]	RRID:SCR_019134 https://bitbucket.org/berkeleylab/metabat/src/master/
SortMeRNA	Kopylova et al.[95]	RRID:SCR_014402 http://bioinfo.lifl.fr/RNA/sortmerna/
Velvet version 1.2.07	Zerbino and Birney[96]	RRID:SCR_010755 http://www.molecularevolution.org/software/genomics/velvet
Oases version 0.2.08	Schulz et al.[97]	RRID:SCR_011896 http://www.ebi.ac.uk/~zerbino/oases/
TransDecoder	Haas, B.J.	RRID:SCR_017647 https://github.com/TransDecoder/TransDecoder
CDDsearch	Marchler-Bauer et al.[98]	N/A
Trimmomatic version 0.38 and version 0.39	Bolger et al.[99]	RRID:SCR_011848 http://www.usadellab.org/cms/index.php?page=trimmomatic
Trinity version version 2.6.5	Grabherr et al.[100]	RRID:SCR_013048 http://trinityrnaseq.sourceforge.net/
rnaSPAdes version version 3.13.1	Bushmanova et al.[101]	RRID:SCR_016992 http://cab.spbu.ru/software/rnaspades/
RepeatMasker version 4.1.0	Smit et al.[102]	RRID:SCR_012954 http://repeatmasker.org/
Tandem repeats finder	Benson et al.[103]	RRID:SCR_022193 https://github.com/Benson-Genomics-Lab/TRF
REPET	Flutre et al.[104]	N/A
BLAT	Kent[105]	RRID:SCR_011919 http://genome.ucsc.edu/cgi-bin/hgBlat?command=start
Genewise	Birney et al.[106]	RRID:SCR_015054 http://www.ebi.ac.uk/Tools/psa/genewise/
DIAMOND version 0.9.30	Buchfink et al.[107]	RRID:SCR_009457 http://www.nitrc.org/projects/diamond/
Est2Genome	Mott[108]	https://galaxy-iuc.github.io/emboss-5.0-docs/est2genome.html
Gmove	Dubarry et al.[109]	RRID:SCR_019132 http://www.genoscope.cns.fr/gmove
votingLNC	Debit, A.	https://gitlab.com/a.debit/votinglnc
AliView version 1.26	Larsson[110]	RRID:SCR_002780 https://github.com/AliView
RAxML version 8.2.	Stamatakis[111]	RRID:SCR_006086 https://github.com/stamatak/standard-RAxML
Tracer version 1.7.2	Rambaut et al.[112]	RRID:SCR_019121 https://bioweb.pasteur.fr/packages/pack@Tracer@v1.6
OrthoFinder version 2.5.2	Emms and Kelly[113]	RRID:SCR_017118 https://github.com/davidemms/OrthoFinder
Count version 9.1106	Csűös[114]	https://www.iro.umontreal.ca/~csuros/gene_content/count.html
MUSCLE version 3.8.1551	Edgar[115]	RRID:SCR_011812 http://www.ebi.ac.uk/Tools/msa/muscle/

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
OD-Seq version 1.0	Jehl et al.[116]	https://bioconductor.org/packages/release/bioc/manuals/odseq/man/odseq.pdf
HMMER3 package versions 3.1b1 and 3.3.2	Mistry et al.[117]	RRID:SCR_005305 http://hmmer.janelia.org/
GenEra	Barrera-Redondo et al.[118]	N/A
MCL	Enright et al.[119]	RRID:SCR_024109 https://micans.org/mcl/
Foldseek	Kempen et al.[120]	https://search.foldseek.com/search
CleanBlastp	Pathmanathan et al.[121]	N/A
SEED	Overbeek et al.[122]	RRID:SCR_002129 http://www.theseed.org/wiki/Home_of_the_SEED
IPR2GO	Paysan-Lafosse et al.[123]	http://www.ebi.ac.uk/interpro/search/sequence-search
eggNOG	Huerta-Cepas et al.[124]	RRID:SCR_002456 http://eggnog.embl.de
eggNOG-mapper	Cantalapiedra et al.[125]	RRID:SCR_021165 http://eggnog-mapper.embl.de
Spearman's rank correlation analysis tool version 1.1.23-r7	P. Wessa, Free Statistics Software, Office for Research Development and Education	https://www.wessa.net/
Prodigal version 2.6.3	Hyatt et al.[126]	RRID:SCR_011936 https://github.com/hyattpd/Prodigal
ViralRecall version 2.0	Aylward et al.[127]	https://github.com/faylward/viralrecall
esl-translate version 0.48	Rivas, E.	https://github.com/EddyRivasLab/easel/blob/master/miniapps/esl-translate.man.in
bedtools version 2.29.2	Quinlan and Hall[128]	RRID:SCR_006646 https://github.com/arq5x/bedtools2
MMseqs cluster version 13.45111	Hauser et al.[129]	RRID:SCR_008184 https://github.com/eturro/mmseq#mmseq-transcript-and-gene-level-expression-analysis-using-multi-mapping-rna-seq-reads
MAFFT v7	Katoh and Standley[130]	RRID:SCR_011811 http://mafft.cbrc.jp/alignment/server/
MEGA	Tamura et al.[131]	RRID:SCR_023017 https://www.megasoftware.net/
NGphylogeny platform	Lemoine et al.[132]	https://ngphylogeny.fr/.
TrimAl	Capella-Gutiérrez et al.[133]	RRID:SCR_017334 http://trimal.cgenomics.org/
TAPscan version 4	Petroll et al.[134,135]	https://plantcode.cup.uni-freiburg.de/tapscan/
Expasy web translator	Duvaud et al.[136]	RRID:SCR_024703 https://web.expasy.org/translate/
Geneious versions 11.0.5 and 11.1.5	Geneious	RRID:SCR_010519 http://www.geneious.com/
Interproscan 94.0	Jones et al.[137]	RRID:SCR_005829 http://www.ebi.ac.uk/Tools/pfa/iprscan/
Clustal 2.1	Thompson et al.[138]	RRID:SCR_001591 http://www.ebi.ac.uk/Tools/msa/clustalo/
Gblocks	Castresana[139]	RRID:SCR_015945 http://molevol.cmima.csic.es/castresana/Gblocks_server.html
Kallisto version 0.44.0.	Bray et al.[140]	RRID:SCR_016582 https://pachterlab.github.io/kallisto/about
Deseq2	Love et al.[141]	RRID:SCR_015687 https://bioconductor.org/packages/release/bioc/html/DESeq2.html

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
FastQC	Andrews[142]	RRID:SCR_014583 http://www.bioinformatics.babraham.ac.uk/projects/fastqc/
Trim Galore version 0.6.5	Krueger et al.[143]	RRID:SCR_011847 http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/
HISAT2 version 2.1.0	Kim et al.[144]	RRID:SCR_015530 http://ccb.jhu.edu/software/hisat2/index.shtml
featureCounts	Liao et al.[145]	RRID:SCR_012919 http://bioinf.wehi.edu.au/featureCounts/
PAML version 4.9i (including MCMCTree)	Yang[146]	RRID:SCR_014932 http://abacus.gene.ucl.ac.uk/software/paml.html
phytools R package	Revell[147]	RRID:SCR_015502 https://cran.r-project.org/web/packages/phytools/index.html
VHICA package	Wallau et al.[148]	https://github.com/cran/vhica
NOVOPlasty version 3.7	Dierckxsens et al.[149]	RRID:SCR_017335 https://github.com/ndierckx/NOVOPlasty
SAMtools version 1.5	Li et al.[150]	RRID:SCR_002105 http://htslib.org/
GeSeq version 2.03	Tillich et al.[151]	RRID:SCR_017336 https://chlorobox.mpimp-golm.mpg.de/geseq.html
ARAGORN version 1.2.38	Laslett and Canback[152]	RRID:SCR_015974 http://mbio-serv2.mbioekol.lu.se/ARAGORN/
ModelFinder	Kalyaanamoorthy et al.[153]	http://www.iqtree.org/ModelFinder/
UFBoot2	Hoang et al.[154]	N/A
SynMap	Haug-Baltzell et al.[155]	https://genomevolution.org/SynMap.pl
DAGChainer	Haas et al.[156]	https://dagchainer.sourceforge.net/
CodeML	Yang et al.[146]	N/A
nwalign	Pedersen, B	https://pypi.org/project/nwalign/
BEAST version 2.7	Bouckaert et al.[157]	RRID:SCR_010228 http://beast.bio.ed.ac.uk/
StarBEAST3 version 1.1.7	Douglas et al.[158]	https://github.com/rbouckaert/starbeast3
bModelTest	Bouckaert et al.[159]	N/A
LogCombiner version 2.4.7	Bouckaert et al.[157]	N/A
TreeAnnotator version 2.4.7	Bouckaert et al.[157]	N/A
SplitsTree 4 version 4.14.6	Kloepper and Huson[160]	RRID:SCR_014734 http://www.splitstree.org/
Hectar	Gschloessl et al.[161]	https://webtools.sb-roscoff.fr/root?tool_id=abims_hectar
RShiny	R Core Team[162]	https://github.com/rstudio/shiny
IQ-TREE 2	Minh et al.[163]	https://github.com/iqtree/iqtree2
Computational analysis of gene family evolution 5 (CAFE5)	Mendes et al.[164]	https://github.com/hahnlab/CAFE5
clusterProfiler	Yu et al.[165]	RRID:SCR_016884 http://bioconductor.org/packages/release/bioc/html/clusterProfiler.html
ggplot2	Wickham et al.[166]	RRID:SCR_014601 https://cran.r-project.org/web/packages/ggplot2/index.html
tidyverse	Wickham et al.[167]	RRID:SCR_019186 https://CRAN.R-project.org/package=tidyverse
Other		
Benchmarking universal single-copy orthologue (BUSCO) analysis version 5, eukaryota_odb10	Manni et al.[168]	RRID:SCR_015008 http://busco.ezlab.org/

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
UniRef90	Suzek et al.[169]	RRID:SCR_010646 http://www.uniprot.org/help/uniref
AlphaFold protein structure database	Varadi et al.[170]	RRID:SCR_023662 https://alphafold.ebi.ac.uk/
NCVOG database	Yutin et al.[171]	N/A
VOGDB database	Trgovec-Greif et al.[172]	https://vogdb.org/
SulfAtlas database	Barbeyron et al.[172]; Stam et al.[173]	https://sulfatlas.sb-roscoff.fr/
Pfam	Mistry et al.[174]	RRID:SCR_004726 http://pfam.xfam.org/
Panther 17.0	Thomas et al.[175]	RRID:SCR_004869 http://www.pantherdb.org/
Simple Modular Architecture Research Tool (SMART)	Letunic et al.[176]	RRID:SCR_005026 http://smart.embl.de/

EXPERIMENTAL MODEL DETAILS

Ascophyllum nodosum

Species: *Ascophyllum nodosum*
Strain: field collected sperm cells
Genotype: diploid
Sex: male
Maintenance: N/A

Chordaria linearis strain ClinC8C

Species: *Chordaria linearis*
Strain: ClinC8C
Genotype: haploid
Sex: monoicous
Maintenance: Maintained in culture

Choristocarpus tenellus strain KU-1152

Species: *Choristocarpus tenellus*
Strain: KU-1152
Genotype: unknown
Sex: unknown
Maintenance: Maintained in culture

Chrysoparadoxa australica strain CS-1217

Species: *Chrysoparadoxa australica*
Strain: CS-1217
Genotype: unknown
Sex: unknown
Maintenance: Maintained in culture

Cladosiphon okamuranus strain S-strain

Species: *Cladosiphon okamuranus*
Strain: S-strain
Genotype: diploid
Sex: n/a
Maintenance: N/A

Desmarestia dudresnayi strain DdudBR16

Species: *Desmarestia dudresnayi*
Strain: DdudBR16
Genotype: haploid

Sex: monoicous
Maintenance: Maintained in culture

***Desmarestia herbacea* strain DmunF**
Species: *Desmarestia herbacea*
Strain: DmunF
Genotype: haploid
Sex: female
Maintenance: Maintained in culture

***Desmarestia herbacea* strain DmunM**
Species: *Desmarestia herbacea*
Strain: DmunM
Genotype: haploid
Sex: male
Maintenance: Maintained in culture

***Dictyota dichotoma* strain KB07f IV**
Species: *Dictyota dichotoma*
Strain: KB07f IV
Genotype: haploid
Sex: female
Maintenance: Maintained in culture

***Dictyota dichotoma* strain ODC1387m**
Species: *Dictyota dichotoma*
Strain: ODC1387m
Genotype: haploid
Sex: male
Maintenance: Maintained in culture

***Dictyota dichotoma* strain KB07m IV**
Species: *Dictyota dichotoma*
Strain: KB07m IV
Genotype: haploid
Sex: male
Maintenance: Maintained in culture

***Dictyota dichotoma* strain KB07sp VI**
Species: *Dictyota dichotoma*
Strain: KB07sp VI
Genotype: diploid
Sex: n/a
Maintenance: Maintained in culture

***Discosporangium mesarthrocarpum* strain MT17-79**
Species: *Discosporangium mesarthrocarpum*
Strain: MT17-79
Genotype: unknown
Sex: unknown
Maintenance: Maintained in culture

***Ectocarpus crouaniorum* strain Ec861**
Species: *Ectocarpus crouaniorum*
Strain: Ec861
Genotype: haploid
Sex: female
Maintenance: Maintained in culture

***Ectocarpus crouaniorum* strain Ec862**
Species: *Ectocarpus crouaniorum*
Strain: Ec862
Genotype: haploid
Sex: male
Maintenance: Maintained in culture

***Ectocarpus fasciculatus* strain Ec846**
Species: *Ectocarpus fasciculatus*
Strain: Ec846
Genotype: haploid
Sex: female
Maintenance: Maintained in culture

***Ectocarpus fasciculatus* strain Ec847**
Species: *Ectocarpus fasciculatus*
Strain: Ec847
Genotype: haploid
Sex: male
Maintenance: Maintained in culture

***Ectocarpus fasciculatus* strain EfasUO1**
Species: *Ectocarpus fasciculatus*
Strain: EfasUO1
Genotype: diploid
Sex: n/a
Maintenance: Maintained in culture

***Ectocarpus fasciculatus* strain EfasUO2**
Species: *Ectocarpus fasciculatus*
Strain: EfasUO2
Genotype: diploid
Sex: n/a
Maintenance: Maintained in culture

***Ectocarpus siliculosus* strain Ec863**
Species: *Ectocarpus siliculosus*
Strain: Ec863
Genotype: haploid
Sex: female
Maintenance: Maintained in culture

***Ectocarpus siliculosus* strain Ec864**
Species: *Ectocarpus siliculosus*
Strain: Ec864
Genotype: haploid
Sex: male
Maintenance: Maintained in culture

***Ectocarpus* species 1 strain Ec sil Puy CHCH Z9 G5f**
Species: *Ectocarpus species 1*
Strain: Ec sil Puy CHCH Z9 G5f
Genotype: haploid
Sex: female
Maintenance: Maintained in culture

***Ectocarpus* species 1 strain Ec sil Puy CHCH Z9 G3m**
Species: *Ectocarpus* species 1
Strain: Ec sil Puy CHCH Z9 G3m
Genotype: haploid
Sex: male
Maintenance: Maintained in culture

***Ectocarpus* species 1 strain Ec03**
Species: *Ectocarpus* species 1
Strain: Ec03
Genotype: haploid
Sex: male
Maintenance: Maintained in culture

***Ectocarpus* species 12 strain Ec fas CH92 Nie 2f**
Species: *Ectocarpus* species 12
Strain: Ec fas CH92 Nie 2f
Genotype: diploid
Sex: female
Maintenance: Maintained in culture

***Ectocarpus* species 12 strain Ec fas CH92 Nie 3m**
Species: *Ectocarpus* species 12
Strain: Ec fas CH92 Nie 3m
Genotype: diploid
Sex: n/a
Maintenance: Maintained in culture

***Ectocarpus* species 13 strain EcNAP12-S#4-19m**
Species: *Ectocarpus* species 13
Strain: EcNAP12-S#4-19m
Genotype: diploid
Sex: n/a
Maintenance: Maintained in culture

***Ectocarpus* species 2 strain Ec06**
Species: *Ectocarpus* species 2
Strain: Ec06
Genotype: haploid
Sex: male
Maintenance: Maintained in culture

***Ectocarpus* species 3 strain Ec10**
Species: *Ectocarpus* species 3
Strain: Ec10
Genotype: haploid
Sex: female
Maintenance: Maintained in culture

***Ectocarpus* species 3 strain Ec11**
Species: *Ectocarpus* species 3
Strain: Ec11
Genotype: haploid
Sex: male
Maintenance: Maintained in culture

***Ectocarpus* species 5 strain Ec13**
Species: *Ectocarpus* species 5
Strain: Ec13
Genotype: haploid
Sex: female
Maintenance: Maintained in culture

***Ectocarpus* species 5 strain Ec12**
Species: *Ectocarpus* species 5
Strain: Ec12
Genotype: haploid
Sex: male
Maintenance: Maintained in culture

***Ectocarpus* species 6 strain EcLAC-371f**
Species: *Ectocarpus* species 6
Strain: EcLAC-371f
Genotype: diploid
Sex: n/a
Maintenance: Maintained in culture

***Ectocarpus* species 7 strain Ec32**
Species: *Ectocarpus* species 7
Strain: Ec32
Genotype: haploid
Sex: male
Maintenance: N/A

***Ectocarpus* species 8 strain EcLAC-412m**
Species: *Ectocarpus* species 8
Strain: EcLAC-412m
Genotype: diploid
Sex: n/a
Maintenance: Maintained in culture

***Ectocarpus* species 9 strain EcSCA-722f**
Species: *Ectocarpus* species 9
Strain: EcSCA-722f
Genotype: haploid
Sex: female
Maintenance: Maintained in culture

***Ectocarpus subulatus* strain Bft15b**
Species: *Ectocarpus subulatus*
Strain: Bft15b
Genotype: haploid
Sex: male
Maintenance: N/A

***Feldmannia mitchelliae* strain KU-2106 Giff mitch BNC GA**
Species: *Feldmannia mitchelliae*
Strain: KU-2106 Giff mitch BNC GA
Genotype: haploid
Sex: monoicous
Maintenance: Maintained in culture

Fucus distichus
Species: *Fucus distichus*
Strain: field collected meristem
Genotype: diploid
Sex: n/a
Maintenance: N/A

Fucus serratus
Species: *Fucus serratus*
Strain: field collected ovule cells
Genotype: diploid
Sex: female
Maintenance: N/A

Fucus serratus
Species: *Fucus serratus*
Strain: field collected sperm cells
Genotype: diploid
Sex: male
Maintenance: N/A

***Halopteris paniculata* strain Hal grac a UBK**
Species: *Halopteris paniculata*
Strain: Hal grac a UBK
Genotype: haploid
Sex: monoicous
Maintenance: Maintained in culture

***Hapterophycus canaliculatus* strain Oshoro5f**
Species: *Hapterophycus canaliculatus*
Strain: Oshoro5f
Genotype: haploid
Sex: female
Maintenance: Maintained in culture

***Hapterophycus canaliculatus* strain Oshoro7m**
Species: *Hapterophycus canaliculatus*
Strain: Oshoro7m
Genotype: haploid
Sex: male
Maintenance: Maintained in culture

***Hapterophycus canaliculatus* strain Oshoro 3F x 9M**
Species: *Hapterophycus canaliculatus*
Strain: Oshoro 3F x 9M
Genotype: diploid
Sex: n/a
Maintenance: Maintained in culture

***Hapterophycus canaliculatus* strain Oshoro 4F x 9M**
Species: *Hapterophycus canaliculatus*
Strain: Oshoro 4F x 9M
Genotype: diploid
Sex: n/a
Maintenance: Maintained in culture

***Hapterophycus canaliculatus* strain Oshoro 6F x 6M**
Species: *Hapterophycus canaliculatus*
Strain: Oshoro 6F x 6M
Genotype: diploid
Sex: n/a
Maintenance: Maintained in culture

***Heribaudiella fluviatilis* strain SAG. 13.90**
Species: *Heribaudiella fluviatilis*
Strain: SAG. 13.90
Genotype: unknown
Sex: unknown
Maintenance: Maintained in culture

***Heterosigma akashiwo* strain CCMP452**
Species: *Heterosigma akashiwo*
Strain: CCMP452
Genotype: unknown
Sex: unknown
Maintenance: Maintained in culture

Himanthalia elongata
Species: *Himanthalia elongata*
Strain: field meristem
Genotype: diploid
Sex: n/a
Maintenance: N/A

***Laminaria digitata* strain LdigPH10-18mv**
Species: *Laminaria digitata*
Strain: LdigPH10-18mv
Genotype: haploid
Sex: male
Maintenance: Maintained in culture

***Laminarionema elsbetiae* strain ELsaHSoW15**
Species: *Laminarionema elsbetiae*
Strain: ELsaHSoW15
Genotype: unknown
Sex: unknown
Maintenance: Maintained in culture

***Macrocystis pyrifera* strain P11A1**
Species: *Macrocystis pyrifera*
Strain: P11A1
Genotype: haploid
Sex: female
Maintenance: Maintained in culture

***Macrocystis pyrifera* strain P11B4**
Species: *Macrocystis pyrifera*
Strain: P11B4
Genotype: haploid
Sex: male
Maintenance: Maintained in culture

***Myriotrichia clavaeformis* strain Myr cla04**
Species: *Myriotrichia clavaeformis*
Strain: Myr cla04
Genotype: haploid
Sex: female
Maintenance: Maintained in culture

***Myriotrichia clavaeformis* strain Myr cla05**
Species: *Myriotrichia clavaeformis*
Strain: Myr cla05
Genotype: haploid
Sex: male
Maintenance: Maintained in culture

***Myriotrichia clavaeformis* strain Myr cla12**
Species: *Myriotrichia clavaeformis*
Strain: Myr cla12
Genotype: diploid
Sex: n/a
Maintenance: Maintained in culture

Pelvetia canaliculata
Species: *Pelvetia canaliculata*
Strain: field collected meristem
Genotype: diploid
Sex: n/a
Maintenance: N/A

***Phaeothamnion wetherbeei* strain SAG 119.79**
Species: *Phaeothamnion wetherbeei*
Strain: SAG 119.79
Genotype: unknown
Sex: unknown
Maintenance: Maintained in culture

***Pleurocardia lacustris* strain SAG 25.93**
Species: *Pleurocardia lacustris*
Strain: SAG 25.93
Genotype: unknown
Sex: unknown
Maintenance: Maintained in culture

***Porterinema fluviatile* strain SAG 2381**
Species: *Porterinema fluviatile*
Strain: SAG 2381
Genotype: unknown
Sex: unknown
Maintenance: Maintained in culture

***Pylaiella littoralis* strain U1.48**
Species: *Pylaiella littoralis*
Strain: U1.48
Genotype: haploid
Sex: unknown
Maintenance: Maintained in culture

***Pylaiella littoralis* strain F24**
Species: *Pylaiella littoralis*
Strain: F24
Genotype: diploid
Sex: n/a
Maintenance: Maintained in culture

***Saccharina japonica* strain Ja**
Species: *Saccharina japonica*
Strain: Ja
Genotype: haploid
Sex: male
Maintenance: N/A

***Saccharina latissima* strain SLPER63f7**
Species: *Saccharina latissima*
Strain: SLPER63f7
Genotype: haploid
Sex: female
Maintenance: Maintained in culture

***Saccorhiza dermatodea* strain SderLü1190fm**
Species: *Saccorhiza dermatodea*
Strain: SderLü1190fm
Genotype: haploid
Sex: monoicous
Maintenance: Maintained in culture

***Saccorhiza polyschides* strain SpolBR94f**
Species: *Saccorhiza polyschides*
Strain: SpolBR94f
Genotype: haploid
Sex: female
Maintenance: Maintained in culture

***Saccorhiza polyschides* strain SpolBR94m**
Species: *Saccorhiza polyschides*
Strain: SpolBR94m
Genotype: haploid
Sex: male
Maintenance: Maintained in culture

Saccorhiza polyschides
Species: *Saccorhiza polyschides*
Strain: field collected sample (young sporophytes ~2-10cm)
Genotype: diploid
Sex: n/a
Maintenance: N/A

Sargassum fusiforme
Species: *Sargassum fusiforme*
Strain: unknown
Genotype: diploid
Sex: n/a
Maintenance: N/A

***Schizocladia ischiensis* strain KU-0333**
Species: *Schizocladia ischiensis*
Strain: KU-0333
Genotype: unknown
Sex: unknown
Maintenance: Maintained in culture

***Scytosiphon promiscuus* strain 000310-Muroran-5-female**
Species: *Scytosiphon promiscuus*
Strain: 000310-Muroran-5-female
Genotype: haploid
Sex: female
Maintenance: Maintained in culture

***Scytosiphon promiscuus* strain Ot110409-Otamoi-16-male**
Species: *Scytosiphon promiscuus*
Strain: Ot110409-Otamoi-16-male
Genotype: haploid
Sex: male
Maintenance: Maintained in culture

***Scytosiphon promiscuus* strain SXS107**
Species: *Scytosiphon promiscuus*
Strain: SXS107
Genotype: diploid
Sex: n/a
Maintenance: Maintained in culture

***Sphacelaria rigidula* strain Sph rig Cal Mo 4-1-68b**
Species: *Sphacelaria rigidula*
Strain: Sph rig Cal Mo 4-1-68b
Genotype: haploid
Sex: female
Maintenance: Maintained in culture

***Sphacelaria rigidula* strain Sph rig Cal Mo 4-1-G3b**
Species: *Sphacelaria rigidula*
Strain: Sph rig Cal Mo 4-1-G3b
Genotype: haploid
Sex: male
Maintenance: Maintained in culture

***Sphacelaria rigidula* strain Sph rig Cal Mo SP**
Species: *Sphacelaria rigidula*
Strain: Sph rig Cal Mo SP
Genotype: diploid
Sex: n/a
Maintenance: Maintained in culture

***Sphaerotrichia firma* strain ET2f**
Species: *Sphaerotrichia firma*
Strain: ET2f
Genotype: haploid
Sex: female
Maintenance: Maintained in culture

Sphaerotrichia firma strain Sfir13m
Species: *Sphaerotrichia firma*
Strain: Sfir13m
Genotype: haploid
Sex: male
Maintenance: Maintained in culture

Tribonema minus strain UTEX B 3156
Species: *Tribonema minus*
Strain: UTEX B 3156
Genotype: unknown
Sex: unknown
Maintenance: N/A

Undaria pinnatifida strain Kr2015
Species: *Undaria pinnatifida*
Strain: Kr2015
Genotype: diploid
Sex: n/a
Maintenance: N/A

METHOD DETAILS

Biological material
Sequencing brown algal genomes has been hampered by the significant challenges involved, including inherent problems with growing brown algae, the presence of molecules that interfere with sequencing reactions and complex associations with microbial symbionts. To address these problems, cultured, unialgal filamentous gametophyte material was used whenever possible (i.e. for species with haploid-diploid life cycles) and the extraction methodology was adapted for each species.

The algal strains analysed in this study are listed in Table S1A, which provides information about the sampling site for each strain. The sampling sites are shown on a world map in Figure S1D.

All strains except those belonging to the Fucales were grown under laboratory conditions. The latter cannot be maintained long-term in the laboratory so field material was harvested for extractions. The haploid gametophyte generation was grown in culture for species with characterised haploid-diploid life cycles, with the exception of *Ectocarpus* strains, for which haploid partheno-sporophytes or diploid sporophytes were cultivated. All cultures were grown either in 140 mm diameter Petri dishes or in 2–10 L bottles, the latter aerated by bubbling with sterile air. Most cultures were grown in Provasoli-enriched[104] natural seawater (PES medium) under fluorescent white light (10–30 μM photons/m2⋅s) at 13°C (or at 10°C for *Hapterophycus canaliculatus* and *Chordaria linearis* or 20°C for *Sphacelaria rigidula*, *Dictyota dichotoma*, *Schizocladia ischiensis* and *Chrysoparadoxa Australica*). Exceptions included the freshwater species *Pleurocladia lacustris*, *Porterinema fluviatile* and *Heribaudiella fluviatilis*, which were grown in natural seawater that had been diluted to 5% with distilled water (i.e., 95% distilled water / 5% seawater) before addition of ES medium (http://sagdb.uni-goettingen.de/culture_medi a/01%20Basal%20Medium.pdf) micronutrients (at 20°C for *P. lacustris*) and *Phaeothamnion wetherbeei*, which was grown in MIEB12 (medium 7 in Letunic et al.[177]). Whole thallus was extracted for all species except the Fucales, where either dissected meristematic regions or released male gametes were extracted. Tissue samples were frozen in liquid nitrogen and stored at -80°C before extraction.

DNA extraction
DNA was extracted using either the OmniPrep Genomic DNA Purification Kit (G Biosciences, St. Louis, MO, USA) or the Nucleospin Plant II midi DNA Extraction Kit (Macherey-Nagel, Düren, Germany). DNA quality was assessed using a Qubit fluorometer (Themo Fisher Scientific, Waltham, MA, USA), and fragment length was assessed by migration on a 1% agarose gel for some of the samples.

Illumina library preparation and sequencing
Libraries were prepared using the NEBNext DNA Modules Products (New England Biolabs, Ipswich, MA, USA) with an 'on bead' protocol developed by Genoscope, starting with 100 ng of genomic DNA. DNA was sonicated to a 100–800 bp size range using a Covaris E220 sonicator (Covaris, Woburn, MA, USA), end-repaired and 3′-adenylated. Illumina adapters (Bioo Scientific, Austin, TX, USA) were then added using the NEBNext Sample Reagent Set (New England Biolabs, Ipswich, MA, USA) and the DNA purified using Ampure XP (Beckmann Coulter Genomics, Danvers, MA, USA). Adapted fragments were amplified with 12 cycles of PCR using the Kapa Hifi Hotstart NGS library Amplification kit (Roche, Basel, Switzerland), followed by 0.8x AMPure XP (Beckman Coulter Genomics, Danvers, MA, USA) purification. Libraries were sequenced with Illumina MiSeq, HiSeq 4000 or NovaSeq 6000 instruments (Illumina, San Diego, CA, USA) in paired-end mode, 150 base read-length.

Oxford Nanopore library preparation and sequencing

Some samples were first purified using the Short Read Eliminator Kit (Pacific Biosciences, Menlo Park, CA, USA). All libraries were prepared using the protocol "1D Genomic DNA by Ligation" provided by Oxford Nanopore Technologies (Oxford Nanopore Technologies Ltd, Oxford, UK). Most of the libraries were prepared with the SQK-LSK109 kit (Oxford Nanopore Technologies), a few with the SQK-LSK108 or SQK-LSK110 kits (Oxford Nanopore Technologies). Three flow cells were loaded with barcoded samples. The samples were mainly sequenced on R9.4.1 MinION or PromethION flow cells.

RNA extraction, Illumina RNA-seq library preparation and sequencing

RNA was extracted using either the Qiagen RNeasy kit or the Macherey Nagel RNAplus kit (Macherey-Nagel, Düren, Germany). RNA-seq libraries were prepared using the TruSeq Stranded mRNA Sample Prep (Illumina) according to the manufacturer's protocol, starting with 500 ng to 1 μg of total RNA, or using the NEBNext Ultra II Directional RNA Library Prep for Illumina (New England BioLabs) according to the manufacturer's protocol, starting with 100 ng of total RNA. The libraries were sequenced with Illumina HiSeq 2500, HiSeq 4000 or NovaSeq 6000 instruments (Illumina, San Diego, CA, USA), in paired-end mode, 150 base read-length.

Assembly strategies

Two assembly strategies were employed: one was designed for genomes exclusively sequenced using short reads with Illumina technology, while the other was designed for genomes that underwent sequencing using a combination of long and short reads, using respectively the Nanopore and Illumina technologies.

Short-read-based genome assembly

When sequencing was performed exclusively using short reads, reads corresponding to bacterial contaminants were filtered out early in the assembly process because, typically, the initial datasets were too large to run assemblers like SPAdes. To remove bacterial contaminants, an assembly based on the initial illumina dataset was first generated for each strain using a fast and non-greedy algorithm, MEGAHIT[82] version 1.1.1 with the parameters –k-min 101 –k-max 131 –k-step 10. Assigning taxonomy is easier when working with contigs than with reads. Contigs exceeding 500 bp in each preliminary assembly underwent taxonomic classification based on gene models predicted using the *ab initio* software MetaGene[83] version 2008.8.19 with default parameters and then aligning proteins against UniprotKB using BLASTp (e-value $<10e^{-4}$). A superkingdom (Eukaryota, Archaea or Bacteria) was assigned to each gene based on the best alignment (selected using the BLASTp score). Contigs that contained more than 50% of their genes assigned to Bacteria and with at least one gene every 10 kbp were classified as bacterial sequences. For each strain, the initial Illumina sequencing reads were aligned against the corresponding bacterial sequences using latest version of the Burrows-Wheeler Aligner[85] (BWA) with default parameters and mapped short-reads were labelled as contaminants, and assembled for the purpose of obtaining more contiguous contigs. These bacterial contigs were then used to build a contaminant sequence database. Finally, the clean subset of reads was obtained by aligning the whole Illumina dataset against this strain-specific bacterial contig database, using Bowtie2[86] version 2.2.9 with default parameters. A final assembly was then generated for each strain using the contaminant-free read datasets and the SPAdes[87] assembler version 3.8.1 with the parameters -k 21,57,71,99,127 -m 2000 –only-assembler –careful. Genome assemblies based only on short-reads were more fragmented (N50 ranged from 3 kbp to 31 kbp) than assemblies that used long reads but the sizes of the former were consistent with expectations.

Long-read-based genome assemblies

A subset of the strains produced DNA of both adequate quality and quantity, enabling successful long-read sequencing. In these cases, long reads were assembled directly and the detection of possible bacterial contigs was carried out after the assembly step. To produce long-read-based genome assemblies we generated three samples of reads i) all reads, ii) 30X coverage of the longest reads and iii) 30X coverage of the filtlong (https://github.com/rrwick/Filtlong) highest-score reads. The three samples were used as input data for four different assemblers, Smartdenovo,[88] Redbean,[89] Flye[90] and Necat.[91] Based on the cumulative size and contiguity, we selected the best assembly for each strain. This assembly was then polished three times using Racon[92] with nanopore reads, and twice with Hapo-G[93] and Illumina PCR-free reads.

Assembly decontamination

Contigs from the short- and long-read genome assemblies were inspected for potential bacterial sequences. This process was carried out using a combination of several analysis and tools: GC composition, read coverage, Metabat 2 (for tetramer composition and clustering)[94] and Metagene (for gene prediction and taxonomic identification, as described previously). Contigs were manually removed based on their characteristics.

Transcriptome assembly

Ribosomal-RNA-like reads were detected using SortMeRNA[95] and filtered out. The Illumina RNA-seq short reads from each strain were assembled using Velvet[96] version 1.2.07 and Oases[97] version 0.2.08 with kmer sizes of 61, 63 and 65 bp. BUSCO[168] analysis (v5, eukaryota_odb10) was then performed on the three resulting assemblies for each strain in order to select the best assembly, i.e. the most complete at the gene level. Reads were mapped back to the contigs with BWA-mem, and only consistent paired-end reads were retained. Uncovered regions were detected and used to identify chimeric contigs. In addition, open reading frames (ORF) and domains were identified using TransDecoder (Haas, B.J., https://github.com/TransDecoder/TransDecoder) and CDDsearch,[98]

respectively. Contigs were broken into uncovered regions outside ORFs and domains. In addition, read strand information was used to correctly orient RNA-seq contigs.

De novo transcriptomes

The RNA-seq data was also used to generate *de novo* transcriptomes. For each strain, all the RNA-seq data available was cleaned to remove poor quality sequence and adapter sequences using Trimmomatic[99] v0.39 prior to being assembled using either Trinity[100] version v2.6.5 or rnaSPAdes[101] version v3.13.1. The strandness and Kmer-length parameters of the assemblers were adjusted to take into account RNA-seq read characteristics. The *de novo* transcriptomes represented an alternative source to identify and characterise genes if they were not detected in the genome assemblies. The *de novo* transcriptomes are available from the CNRS Research Data dataset (https://doi.org/10.57745/9U1J85) and from the Phaeoexplorer website (https://phaeoexplorer.sb-roscoff.fr/).

Detection and masking of repeated sequences and transposons

Prior to gene annotation, each genome assembly was masked based on the repeat library from *Ectocarpus* species 7 (formerly *Ectocarpus siliculosus*)[11] and using RepBase with RepeatMasker[102] version v.4.1.0, default parameters. Tandem repeats finder (TRF)[103] was also used to mask tandem repeat duplications. In addition, transposons were annotated in ten species using REPET[104] and the transposons detected were used as a reference to mask all genomes with RepeatMasker[102] version v4.1.0, default parameters.

Gene prediction

For each strain, gene prediction was performed using both homologous proteins and RNA-seq data. Proteins from *Ectocarpus* species 7 (https://bioinformatics.psb.ugent.be/orcae/overview/EctsiV2)[178] and UniRef90 (https://www.uniprot.org/uniref/) were aligned against each genome assembly. First, BLAT[105] with default parameters was used to quickly localise putative genes corresponding to the *Ectocarpus* species 7 proteins. The best match and matches with a score $\geq$ 90% of the best match score were retained. Second, the alignments were refined using Genewise[106] with default parameters, which is more precise for intron/exon boundary detection. Alignments were retained if more than 80% of the length of the protein was aligned to the genome. To detect conserved proteins and allow detection of horizontal gene transfer, UniRef90 proteins (without *E. siliculosus* sequences) were aligned with DIAMOND[107] (v0.9.30 with parameters –evalue 0.001 –more-sensitive) to genomic regions lacking alignments with an *Ectocarpus* species 7 protein. Only the five best matches per locus were retained, based on their bitscore. Selected proteins from UniRef90 were aligned to the whole genome using Genewise as described previously, and alignments with at least 50% of the aligned protein length were retained. The assembled transcriptome for each strain was aligned to the strain's genome assembly using BLAT[105] with default parameters. For each transcript, the best match was selected based on the alignment score, with an identity greater or equal to 90%. Selected alignments were refined using Est2Genome[108] in order to precisely detect intron boundaries. Alignments were retained if more than 80% of the length of the transcript was aligned to the genome with a minimal identity of 95%. Finally, the protein homologies and transcript mapping were integrated using a combiner called Gmove.[109] This tool can find coding sequences (CDSs) based on genome-located evidence without any calibration step. Briefly, putative exons and introns, extracted from the alignments, were used to build a simplified graph by removing redundancies. Then, Gmove extracted all paths from the graph and searched for open reading frames (ORFs) consistent with the protein evidence. Translated proteins of predicted genes were then aligned against NR prot (release 19/02/2019) and the *Ectocarpus* species 7 version v2 proteome[178] (https://bioinformatics.psb.ugent.be/orcae/overview/EctsiV2) using DIAMOND BLASTp with parameters –evalue 10-5 –more-sensitive –unal 0. All predicted genes with significant matches (the smallest protein had to be aligned for at least 50% of its length) were retained. In addition to these genes, we also retained genes with CDS size greater than 300 bp and with a coding ratio (CDS size / mRNA size) greater or equal to 0.5.

Annotation decontamination

After predicting the genes, an additional analysis was carried out to detect bacterial sequences. If a contig did not contain any genes, it was analysed with MetaGene and the predicted proteins added to the gene catalogue for the purpose of detecting bacterial sequences. Proteins generated from predicted genes (Gmove plus MetaGene) were then aligned against UniprotKB using BLASTp (e-value < $10e^{-4}$) and superkingdom (Eukaryota, Archaea or Bacteria) was assigned to each gene based on the best alignment (selected using the BLASTp score). Contigs that contained more than 80% of their genes assigned to bacteria, Archaea or viruses were classified as bacterial sequences and removed from the final assembly file. Genes belonging to these contigs were also removed from the final gene catalogue. Finally, completeness of each predicted gene catalogue was assessed using BUSCO[168] (v5.0.0; eukaryota_odb10).

In addition, the quality of the annotations was assessed by comparing the length of coding regions in pairs of orthologous proteins (best reciprocal hits) between each genome and *Ectocarpus* species 7, which was used as a reference because its high-quality annotation has been extensively curated.[178] The correlation between orthologous CDS lengths was higher for genomes sequenced with long reads than for genomes only sequenced with short reads (Figure S1B). This difference was probably principally due to a higher proportion of underestimated protein lengths in the latter (Table S1B) which likely corresponded to fragmented genes. The qualities of Ectocarpales genome annotations were very high (BUSCO and length of predicted CDS) even when the genomes were sequenced

using only short reads, probably because their phylogenetic proximity to *Ectocarpus* species 7 facilitated the building of good quality gene models.

Analyses aimed at deducing functional characteristics of predicted proteins

Several different analyses of the predicted proteomes of each species were carried out to provide information about the cellular functions of the encoded proteins. These included eggNOG-mapper[125] analyses (v2.1.8 or v2.0.1, with emapperDB v5.0.2 or v4.5.1) to provide multiple functional annotations (Gene Ontology, Kyoto Encyclopedia of genes and genomes, Clusters of Orthologous Genes, Pfam), Interproscan[137] analyses (versions v5.55-88.0, v5.51-85.0 or v5.36-75.0) to detect functional domains, Hectar[161] (v1.3) predictions of protein subcellular localisation and various DIAMOND[107] (v2.0.15 vs UniRef90 2022_03, with parameter "evalue" set to $10e^{-5}$) sequence similarity searches aimed at identifying homologous proteins with functional annotations.

Detection of tandemly duplicated genes

Starting with the protein alignments that had been constructed to build the orthogroups, matches between proteins within the same genome with an e-value of $\leq 10-20$ and which covered at least 80% of the smallest protein were extracted. Two genes were considered to be tandemly duplicated if they were localised on the same genomic contig separated by five or less intervening genes, regardless of their orientation. The tandemly-duplicated genes were clustered using a single linkage clustering approach. A contingency test was applied to compare the proportion of tandemly-duplicated genes in each orthogroup with the global proportion of tandemly-duplicated genes (p=0.0532792). The p-values are shown in Table S1.

Relative orientation of adjacent genes and lengths of intergenic regions

For each species, the proportion of pairs of adjacent genes localized on opposite strands was compared to the expected proportion of 0.5 using a binomial test (with p=0.5). The p-values are shown in Table S1B (p-values of <0.05 correspond to cases where the proportion is significantly higher than 0.5).

The lengths of intergenic regions between pairs of adjacent genes located on opposite strands (i.e. divergently or convergently transcribed) were compared with the lengths of intergenic regions between genes located on the same strand (i.e. transcribed in the same direction). Contingency tables were constructed for each species using a threshold of 1000 bp for the intergenic length and the number of intergenic regions in each of four categories were counted: 1) same strand genes, intergenic <1000 bp, 2) opposite strand genes, intergenic <1000 bp, 3) same strand genes, intergenic $\geq$1000 bp, 4) opposite strand genes, intergenic $\geq$1000 bp. Fisher exact tests were applied to the contingency tables (alternative hypothesis: true odds ratio is greater than 1). The p-values are shown in Table S1. When p-values are <0.05, short intergenic lengths are significantly associated with pairs of genes on opposite strands. All calculations were performed with R[162] (version 4.3.0).

Detection of long non-coding RNAs

Transcriptome data for 11 species (Table S1F), including nine brown algal strains and two outgroup taxa, was analysed to identify lncRNAs. Any transcripts with invalid nucleotide DNA symbols were discarded and sequences shorter than 200 nucleotides were removed to avoid the detection of small RNA transcripts. The transcriptome sequences in Fasta format were analysed with votingLNC (https://gitlab.com/a.debit/votinglnc) to detect lncRNA transcripts and assign a confidence level for each transcript. A similar approach was used to detect lncRNAs in the lncPlankton database.[179] VotingLNC is a meta-classifier combining the predictions of the ten most commonly used coding potential tools. Based on a majority voting ensemble procedure, the meta-tool assigns the final coding potential class to a transcript as the class label predicted most frequently by the ten classification models included in the ensemble. Alongside the majority voting class, a reliability score was calculated for each transcript. A cut-off non-coding reliability score of $p > 0.5$ was chosen to treat a transcript as lncRNA and to decrease false-positive identification. The set of transcripts predicted as lncRNA by the majority-voting procedure and having an ORF(s) encoding peptide(s) with length $\geq$ 100aa were discarded. lncRNA transcripts that had significant matches in either the Pfam[174] (hmmscan e-value < 0.001) or SwissProt (BLASTp e-value < $1e^{-5}$ and similarity $\geq$ 90%) databases were removed from the dataset. Transcript length, GC content, and the length of the longest ORF were compared between lncRNAs and protein-coding RNAs. The comparison was carried out using a Wilcoxon test. R version V.4.1.2 was used for all the analyses and ggplot2[166] (version 3.4.0) for plotting.

Intron conservation

Intron positions were compared in a set of single copy genes that are conserved across all the Phaeophyceae and the outgroup species. The analysis focused on the 21 reference genomes (Table S1F) and on orthogroups that occurred exactly once in at least 20 of the 21 genomes, allowing the gene to be absent from only one of the 21 genomes. In addition, orthogroups were discarded if more than three copies had been annotated in the other Phaeophyceae genomes. These filters produced a set of 235 conserved (ancestral) orthogroups. Multiple alignments were carried out for each orthogroup using MUSCLE[115] version 3.8.1551 with default parameters and conserved blocks were identified with Gblocks[139] version 0.91b with the parameters -p=t -s=n -b5=a -b2=[nsp] -b1=[nsp] -b3=6, where "nsp" is equal to 90% of the number of proteins aligned. A shell script was then used to compare intron positions in the alignments. For each intron in the multiple sequence alignment, we obtained a corresponding conservation profile listing which species contains an intron at that position. The profiles obtained for the 949 introns that are in conserved blocks of the multiple alignments are

shown in Figure S4B. Both phase and length of ancestral introns (e.g. that were conserved in most Phaeophyceae and at least two sister clades) were compared to the phase and length of *Ectocarpus* species 7 introns as a reference. The same approach was used to compare intron positions across 11 *Ectocarpus* species, with *Scytosiphon promiscuus* as an outgroup, by selecting 831 conserved monocopy orthogroups. The number of introns per gene in brown algae and in closely-related outgroup species were compared using a contingency test (Table S1C).

Phylogenomic tree of the Phaeophyceae

To provide a phylogenetic framework for the analyses of the Phaeoexplorer genome dataset, the 41-species phylogenomic tree reported by Akita et al.[180] was updated by adding 15 additional species using the same methodology. Briefly, for the additional species, amino acid sequences were recovered for the 32 single-copy orthologous genes used to construct the published tree and these were aligned manually with the existing sequences using the alignment software AliView[110] v.1.26. The aligned sequences of the final 56 species were concatenated and maximum likelihood analysis was carried out with 10,000 rapid bootstraps using RAxML[111] v.8.2.9 and the gamma model. The best-fit evolutionary model for each gene was determined using AIC.

Bayesian divergence time estimation for the brown algae

An estimation of brown algal divergence time was carried out using the 32 orthologous nuclear genes (see above and) for 51 brown algae and five non-brown species (16,185 amino acids, 56 spp.) and MCMCTree (PAML package v4.9j) with the approximate likelihood method. The WAG protein model was selected based on the AIC and BIC criteria of ModelFinder.[153] The independent clock model was selected based on previous work on the brown algal timeline by Choi et al.[1] One hundred million years was set to correspond to 1 in the MCMCTree calculation. A secondary calibration for the root was based on Choi et al.[1] using a gamma distribution of 70.2 alpha and 10.22 beta. A kelp holdfast fossil[55] was used to date the crown node of kelps with a minimum bound of 0.31, and a *Julescraneia* fossil[181] for the *Macrocystis*/*Saccharina* clade with a minimum bound of 0.13 (Figure S2A). MCMC chains were run 1.5 million generations, with the first 200,000 MCMC chains being discarded as burn-in, and the convergence of MCMC chains was checked with Tracer v1.7.2.[112] This analysis estimated that Schizocladiophyceae and brown algae diverged 457.88 Mya (95% HPD: 321.29-592.66 Ma), similar to (about 8 Mya older than) the previous estimate using plastid genes[1] and that diversification of the major brown algal linages began about 220 million years later, after the origin of DFI clade (235.97 Mya, 95% HPD: 158.88-312.48 Mya), about 12 Ma earlier than the previous estimate.[1] The fossil-calibrated phylogenetic tree for 11 *Ectocarpus* species (Figure S2C) was extracted from the brown algal tree (Figure S2A).

Detection of orthologous groups

Predicted proteins from the 60 strains sequenced in Phaeoexplorer complemented with 16 public proteomes covering the Ochrophytina subphylum and the terrestrial oomycetes were clustered using OrthoFinder[113] v2.5.2 with default parameters. This generated 56,340 orthogroups that contained 90.1% of the proteins (1,415,341 of the 1,571,648). Seventy-one of the 76 strains had more than 75% of their proteins in an orthogroup shared with at least one other strain. The orthogroups contain between 2 and 6,220 proteins with a mean of 25.1 proteins and a median of three.

Dollo analysis of orthogroup gain and loss

An analysis of evolutionary events of gene family gain and loss was carried out on a selection of strains covering the brown algal phylogeny and sister groups as distant as the Raphidophyceae under the Dollo parsimony law using orthogroups as proxies for gene families. To limit possible problems due to the fragmentation of predicted proteins in some assemblies, we selected 24,410 orthogroups present in at least one of 17 strains that had both good quality genome assembly and good quality gene predictions. Dollo parsimony analysis was then run using Count[114] version v9.1106 based on a cladogram of a subset of 24 species representative of the Phaeoexplorer project and excluding all public outgroups more distant than *Heterosigma akashiwo*. The cladogram was based on the topology of the brown algae phylogenetic tree published by Akita et al.[182]

Phylostratigraphy analysis

GenEra[118] was used to estimate gene family founder events for each genome assembly by running DIAMOND[107] in ultra-sensitive mode against the Phaeoexplorer protein dataset and the NCBI non-redundant database. All sequence matches with e-values < 10^{-5} were treated as being homologous with the query genes in the target genomes. The NCBI taxonomy was used as an initial template to infer the evolutionary relationships of each query gene with their matches in the sequence database but taxonomic assignments within the PX clade and Phaeophyceae were then modified to reflect the evolutionary relationships that were inferred in the maximum likelihood tree. Gene families were predicted based on a clustering analysis of the query proteins against themselves using an e-value cutoff of 10^{-5} in DIAMOND and an inflation parameter of 1.5 with MCL.[119] Estimated evolutionary distances were extracted for each pair of species from the maximum likelihood species tree (substitutions/site) to calculate homology detection failure probabilities.[183] Taxonomic sampling of the species tree enabled homology detection failure tests to be carried out within the PX clade. Gene families whose ages could not be explained by homology detection failure were analysed by inspecting the functional and domain annotations for *Ectocarpus* species.[7,179] Structural alignments were performed using Foldseek[120] against the AlphaFold protein structure database.[170]

Detection of gene family amplifications

A binomial test with a parameter of 17/21 was carried out to detect gene families (OGs) that had significantly expanded in 17 Phaeophyceae reference genomes compared with four closely-related outgroup species (*Schizocladia ischiensis*, *Tribonema minus*, *Chrysoparadoxa australica* and *Heterosigma akashiwo*; Table S1F). Expanded gene families deviated significantly from the expected proportion (17/21 under the null hypothesis where there are equal gene numbers in all species). Benjamini–Hochberg FDR correction for multiple testing was then applied and 233 candidate OGs with corrected *p*-values of < 0.001 were retained. All calculations were performed with R (version 4.1.0).

The set of 233 candidate OGs was then filtered to limit counting errors due to annotation artefacts (e.g. genes missed or fragmented) using the following procedure:

1) A protein consensus was first deduced for each orthogroup. Protein sequences representative of all lineages were extracted and aligned using MUSCLE[115] version 3.8.1551 with default parameters and the multiple alignments were filtered using OD-Seq[116] version 1.0 to remove outlier sequences, with parameter –score set to 1.5. The consensus sequences were then extracted from the multiple alignments of non-outlier sequences using hmmemit in the HMMER3[117] package version 3.1b1 with default parameters.
2) In order to estimate gene family copy number independently of the assembly and annotation processes, short read sequences for each genome were mapped onto the orthogroup consensus sequences using DIAMOND.[107] Unique matches were retained for each read and depth of coverage was calculated for each consensus orthogroup. The depth obtained for each orthogroup was normalised for each species by dividing by the depth obtained on a set of conserved single-copy genes, so that the final value obtained was representative of the gene copy number. Then, for each candidate amplified orthogroup, the average depth for the 17 Phaeophyceae species and the average depth for the four outgroup species was calculated and OGs where the depth for outgroups was more than half the depth for the Phaeophyceae were discarded. We retained 227 out of 233 orthogroups after this step.
3) Finally, functional annotations were used to remove orthogroups that were likely to correspond to transposable elements. A final list of 180 OGs was retained (Table S3).

The amplified gene families were manually categorised into functional classes based on the output of automatic functional annotation programs (InterProScan,[137] EggNOG,[124] nr BLASTp) and an amplification profile was assigned to each orthogroup by identifying the taxonomic group where the amplification of the family was most marked (Table S3).

In addition to the binomial tests, we also ran CAFE5[164] to reconstruct the history of gene family amplifications. Such reconstructions rely on a species tree and require that all gene families are present at the root of the tree. However, of the 180 amplified OGs that were strongly amplified in Phaeophyceae (see above and listed in Table S3) only 19 were present at the ancestral node. The majority (161) of the 180 families were gained during the early evolution of the lineage, most (105) at the origin of the PX clade (i.e. a collapsed node corresponding to nodes n1 and n2 in Figure S2B) or of the Phaeophyceae/FDI clades (i.e. a collapsed node corresponding to nodes n5 and n6; Figure S2B). To determine whether the 180 amplified OGs were significantly enriched in genes that were gained early during Phaeophyceae evolution (i.e. at nodes n1/n2, n4, n5/n6 in Figure S2B), a Chi-squared test was carried out using the R *chisq.test* function on a contingency table containing the proportions of OGs gained at various periods during brown algal evolution for both the amplified OGs and for the entire set of OGs as a reference dataset (Table S1C). Twelve independent CAFE5 reconstructions were carried out on the OG subsets gained at 12 different nodes (n0, n1/2, n4, n5/n6, n8, n9, n10/n11, n13, n15, n18, n19, n20), using the subtrees rooted at these nodes so that the sets of OGs gained at each node would be placed at the root of the tree for one of the 12 analyses (Figure S3E). The analysis focused on the 19 highest quality genomes (Table S1F), which is why some pairs of nodes were collapsed (e.g. nodes n1 and n2 to give n1/n2). Several parameters were tested for CAFE5: the –p option (Poisson distribution) resulted in better likelihood scores than default, but we observed a weak effect when increasing the value of lambda (–k). Consequently, all reconstructions were performed with –p (and no k, i.e. k=1) for efficiency purposes. As recommended by Mendes et al.,[164] very large gene families were discarded as these can cause the program to fail to initialize the parameters. The twelve reconstructions were then aggregated and the proportions of amplified and reduced gene families were calculated for each node (Table S3). Only results on internal nodes were considered, since leaves are more subject to artefactual amplifications/reductions due to genes being missed, fused or split in the annotations.

Composite genes

The amino-acid sequences of all 530,598 genes present in the selected genomes were compared in an all-against-all pairwise alignment using DIAMOND BLASTp[107] version 2.0.11; "very-sensitive" mode; e-value threshold $1e^{-5}$. This raw alignment was then filtered using CleanBlastp, from the CompositeSearch suite,[121] to remove sequence alignments with under 30% residue identity and produce the final sequence similarity network. CompositeSearch was then used on this network to identify putative composite gene families among the orthologous groups (OGs) previously computed by OrthoFinder.[113] Composite OGs containing two or more genes and having non-overlapping regions aligned to their component OGs were retained for further analysis, while singleton composite OGs and composites with overlapping component regions were discarded. A phylogeny-based approach,[184] which uses information from extant genomes to apply a Dollo parsimony model in Count,[114] was used to reconstruct the evolutionary events

(domain fusions and fissions) that led to structural rearrangements of composite genes, allowing them to be labelled as fusion or fission events (or as complex events when sequentiality could not be clearly deduced).

Horizontal gene transfer (HGT)

Dataset and experimental approach

Uneven data collection across taxa can impact HGT identification. The phylogeny-based HGT screening approach used here requires the establishment of a comprehensive and taxonomically diverse reference dataset. The analysis focused on the Phaeoexplorer genomes using a background database called REFAL and an automated bioinformatics tool called RoutineTree, which screens for HGTs using phylogenetics. The background database was built using a starting database, GNM1157, which includes a diverse set of 17,250,679 protein sequences from 1157 genomes spanning various prokaryotic and eukaryotic lineages (540 bacteria, 45 archaea, 431 Opisthokonta, 15 Rhodophyta, 83 Viridiplantae, and 43 genomes from CRASH lineages). Data from NCBI RefSeq (updated as of May 2020) and MMETSP were integrated into GNM1157 to form the background database REFAL. To enhance data quality and reduce redundancy, CD-HIT version 4.5.4 was used to remove highly similar sequences (with sequence identity $\geq$90%) within each taxonomic order. This curation process resulted in a protein database consisting of 39.9 million sequences, representing over 7,786 taxa and providing comprehensive coverage across the diverse branches of the tree of life. To obtain the best assembled genome within a genus, the latest version was selected if multiple versions were available. In addition, the dataset was expanded by searching for genomes in other repositories such as the Joint Genome Institute. Special attention was paid to achieving balanced representation of the Rhodophyta and Viridiplantae, which are particularly crucial for HGT analysis within the Chromalveolate group. To accomplish this, protein data from six red algal transcriptomes sourced from MMETSP was added. The HGT search was applied to 72 Stramenopile genomes, including 45 newly sequenced and 27 public genomes.

Phylogenetic Tree Reconstruction

The pipeline for constructing phylogenetic trees splits fasta files into individual sequence files and then carries out a search for homologous sequences, followed by multiple sequence alignment and tree-building. Nested positions within the trees were identified using artificial intelligence and hU and hBL methods were used for HGT verification. Instead of using all available sequences, sequences with the best BLAST hit scores from each kingdom, phylum, and class were used for tree construction to expedite tree-building and enhance clarity. Each gene, regardless of whether it was a copy or not, was used as a query for tree construction. To improve precision, four different methods were used for tree building: neighbour-joining, maximum parsimony, maximum likelihood and Bayesian. As a result, each node within a tree was associated with four support values. To create single-gene phylogenetic trees, a BLASTp[84] search was carried out against the background database, employing an e-value cutoff of $1e^{-05}$. For each query, the top 1,000 significant matches were sorted by bit-score in descending order as the default criterion. Matching sequences were then retrieved from the database, with a constraint of no more than three sequences per genus and no more than 12 sequences per phylum. To further refine the selection, significant matches with a query-subject alignment length of at least 120 amino acids were re-sorted based on query-subject identity in descending order. A second set of homologous sequences was then retrieved from the database following the same procedure. These two sets of homologous sequences, along with the query, were merged and aligned using MUSCLE[115] version 3.8.31 with default settings. The resulting alignments, trimmed to a minimum length of 50 amino acids using TrimAl[133] version 1.2 in automated mode (-automated1), were used to construct phylogenetic trees with FastTree version 2.1.7, with the 'WAG + CAT' model and four rounds of minimum-evolution SPR moves (-spr 4) along with exhaustive ML nearest-neighbour interchanges (-mlacc 2 -slownni). Branch supports were estimated using the Shimodaira-Hasegawa (SH)-test.

Inferring HGT based on tree topology

Phylogenetic trees were examined to identify specific topologies where Phaeoexplorer query sequences were nested among other sequences, defined as a situation where two or more monophyletic clades consist of both queries and prokaryotic sequences, supported by distinct nodes within the tree. These monophyletic clades are considered to group together if they share the same set of prokaryotic sequences but differ in sequences from optional taxa. Singletons for both the donor and receptor genes were excluded to minimise contamination and recent HGT interference. To retain only robustly supported nested positions, positions were required to be multiply supported, with a minimum of $\geq$0.70 for the SH-test and aByes-test support from at least two Phaeoexplorer receptor nodes and three donor supporting nodes. Furthermore, queries that displayed significantly different amino acid compositions ($P < 0.05$) compared to the remaining sequences in the alignment were discarded. Queries from the CRASH category that nested among sequences from other kingdoms (supported by >70% UFBoot at one or more supporting nodes) were retained.

Enhancing accuracy and establishing the timing of HGTs

To enhance accuracy, a minimum requirement was imposed for all supporting nodes and for strongly supported nodes that indicate query-donor monophyly. To determine the timing of HGT events, temporal information, primarily derived from the timetree database, was incorporated into each node. We assigned the "smallest boundary" role to pinpoint the most recent common ancestor at the time of the HGT event. Essentially, if all descendants of a given query protein sequence can be traced back to the initial HGT event, a common ancestral node can be identified whose occurrence time can be inferred using a molecular clock approach based on archaeological and fossil evidence. The taxonomy boundaries of HGT descendants were determined by identifying the smallest ancestor shared by both the donor and receptor taxa from the monophyletic clades within the tree. By considering the emergence

times of both taxa, the timing of the transfer of genes from earlier taxa to later taxa can be determined, as the reverse scenario is not considered plausible.

Verification of HGTs

Verification of HGT used the following contamination assessment criteria: i) HGT candidates were excluded if they were located in a contig where 50% of the genes had better matches with other kingdoms, ii) HGT candidates were excluded if they were located in a contig where 50% of the genes were primarily identified as HGT genes, iii) HGT candidates were excluded if one of their five closest flanking genes, both upstream and downstream, had a better match with other kingdoms. AI, hU and the hBL value were used to further validate HGT events. This process was supplemented with annotation and functional predictions for the identified HGTs.

Further validation was based on the following concepts:

OUTGROUP. This comprises all biological donors present in a tree, excluding the query species if it belongs to biological donors.

SKIP. This includes all biological receptors (species belonging to optional taxa) in a tree, again excluding the query species if it belongs to biological receptors.

INGROUP. This encompasses species from SKIP's upper level, excluding SKIP itself and the query species (if it belongs to biological receptors).

AI (Alien Index). computed for each query gene using e-values from BLAST hits:

$$AI = (E - value\ of\ best\ BLAST\ hit\ in\ the\ INGROUP\ lineage) / (E - value\ of\ best\ BLAST\ hit\ in\ the\ OUTGROUP\ lineage)$$

The AI score quantifies how similar queries are to their homologs in the OUTGROUP compared to homologs in the INGROUP. We apply a relatively lenient cut-off (AI > 0) for initial screening, which can be adjusted in the second screening as needed.

hU (HGT Score Support Index). calculated for each query gene based on the best bit scores of INGROUP vs. OUTGROUP:

$$hU = (Best - hit\ bitscore\ of\ OUTGROUP) - (Best - hit\ bitscore\ of\ INGROUP)$$

A lenient cut-off (hU > 0) is used for initial screening, with flexibility for adjustment in the second screening.

hBL (HGT Branch Length Support Index). calculated based on the minimum branch length to the query within INGROUP vs. OUTGROUP:

$$hBL = (Minimum\ branch\ length\ to\ the\ query\ within\ INGROUP) - (Minimum\ branch\ length\ to\ the\ query\ within\ OUTGROUP)$$

A lenient cut-off (hBL > 0) is applied initially, with the option for modification in the second screening.

CHE, CHS, CHBL (Consensus Hit Support). To mitigate the possibility that the best bit score for either INGROUP or OUTGROUP is influenced by contamination, we consider alternative matches. We introduce consensus hit support (CHE, CHS, and CHBL) to assess the reliability of AI, hU, and hBL, respectively.

For example, if AI > 0, CHE evaluates the likelihood that "AI remains greater than 0" when using the e-value of each sequence in OUTGROUP instead of the e-value of the best BLAST hit in the OUTGROUP lineage (bbhO). A similar approach applies to CHS for hU and CHBL for hBL. This additional layer of evaluation helps ensure the robustness of the HGT verification process.

Gene codon usage, functional annotation and expression

Indices of codon usage and GC content were calculated using Codonw 1.4.4 (http://codonw.sourceforge.net). Gene functions were assigned by searching against the Gene Ontology (GO) database using blast2GO (ref blast2GO 08) and the KEGG database using blastKOALA (http://www.kegg.jp/blastkoala/) with default parameters. The full gene sets of each species were set as the background for KEGG and GO enrichment analyses by applying Student's t-test (p-value cutoff = 0.01). HGTs were also analysed with SEED (http://www.theseed.org/wiki/Home_of_the_SEED), IPR2GO (http://www.ebi.ac.uk/interpro/search/sequence-search), eggNOG[124] (http://eggnogdb.embl.de/#/app/home) and Pfam.[175] For each species, the differences between mean gene expression levels for HGTs and non-HGT genes with common GO terms were accessed using Student's t-test. Go terms with less than five genes in either gene category were ignored. The differences in expression dispersal (coefficient of variation: standard deviation across genes or samples / mean value) and expression specificity (frequencies of a gene to be detected as unexpressed, defined as transcripts per kilobase million (TPM) = 2, in any condition) were accessed in a similar manner. Given the variable experimental conditions associated with different transcriptome data for each species, gene expression values for a gene were used indiscriminately regardless of the conditions. Correlation tests between the codon adaptation index (CAI) and gene expression were carried out using the Spearman's rank correlation analysis tool (P. Wessa, Free Statistics Software, Office for Research Development and Education, version 1.1.23-r7, https://www.wessa.net/).

Comparative analysis of gene sets identified by genome-wide analyses of evolutionary history

Genes identified as belonging to orthogroups that were predicted to be gained at specific nodes of the phylogenetic tree based on the Dollo parsimony analysis, to belong to either significantly amplified gene families (binomial analysis) or to belong to gene families that have significantly changed in size over evolutionary time (CAFE5 analysis), to correspond to founder events (Phylostratigraphy analysis), to have been remodelled (composite gene analysis) or to have been derived from an HGT (HGT analysis) were extracted from the output of each of these analysis and aggregated in a single datatable. Correspondences were established manually between phylogenetic tree nodes and phylostrata and this information was integrated into the datatable. Counting and calculations of the

frequency of events at specific time points were carried out using *ad hoc* R scripts (R version 4.4.1) and the tidyverse[167] package (version 2.0.0). Graphs were generated using the ggplot2[166] package (version 3.5.1). For each gene, a COG functional category was retrieved from the eggNOG mapper output and the COG enrichment analysis was carried out in R using the clusterProfiler[165] package (version 4.6.2) by comparing each set of gene families with the full set of gene families.

Detection of viral genome insertions and viral regions in algal genomes

To reduce the dataset size for analysis, 64 Phaeoexplorer and eight public genomes were initially filtered to retain only contigs that were more than 10 kbp in length. Gene prediction was then carried out on all contigs using Prodigal[126] (V2.6.3, settings: default, meta) and the resulting proteins were used as queries against the NCVOG[171] and VOGDB[185] databases using hmmscan (HMMER 3.3.2 with default settings). The contigs detected by hmmscan were then filtered to retain only sequences with at least one match to either viral database at a defined e-value cutoff ($1e^{-20}$ for NCVOG, and $1e^{-80}$ for VOGDB). The resulting positive 4,951 contigs were then analysed using ViralRecall[127] version 2.0 with settings -w 50 -g 1 -b -f -m 2 using the built-in Nucleocytoviricota (NCV) database GVOG and a window size of 50 kbp. To ensure that viral genes were not missed because they had not been annotated by Prodigal, six-frame translations of the contigs were generated using esl-translate (version 0.48 with default settings), and the resulting proteins queried against the same databases used by ViralRecall using hmmsearch (HMMER 3.3.2, settings: -E 1e-10). The ViralRecall results were then parsed using an in-house workflow. Six-frame translations were removed from the results if they overlapped (even partially) with any Prodigal gene prediction, as identified using bedtools[128] (v2.29.2; intersect). Likewise, overlapping six-frame translations and gene predictions with the same NCVOG match were removed to reduce redundancy. Based on the distance between query sequences with the same GVOG hit, queries were flagged as frame-shifted (less than 100 bp gap), intron-containing (100-5,000 bp gap) or mono-exonic (greater than 5,000 bp gap). All queries were also checked for overlaps with multi-exonic genes that had been annotated by the Phaeoexplorer gene prediction procedure (using Gmove[109]), and flagged if they did. All queries were then filtered to retain only those that matched a set of key NCV marker genes, identified by NCVOG code (A32, D5 helicase, D5 DNA primase, MCP, DNA polymerase B, SFII and VLTF3) or some Phaeovirus integrase genes (integrase recombinase, integrase resolvase and RNR). The marker gene proteins were clustered with the protein sequences of NCVOGs using MMseqs cluster[129] (version 13.45111 with settings –min-seq-id 0.3 -c 0.8). Finally, the parsed results of the NCV marker gene set identified by the ViralRecall screen were manually curated, retaining only those queries with varying combinations of the following properties: placement within a viral region as identified by ViralRecall, similar hmmsearch results (score and e-value) and gene length to that of known NCV genes, not part of a multi-exonic gene, lack of Pfam HMM matches to cellular domains sharing homology to the marker gene (specific to certain marker genes), and clustered with an NCVOG in the MMseqs analysis. We noted that the median number of viral regions found in genomes assembled with long reads was very similar to that for genomes assembled with short reads (9 and 10, respectively). The marker gene content of the viral regions was manually assessed to estimate the number of complete or partial inserted viruses in each genome. VRs were considered to be complete proviruses if they contained all seven of the key NCV marker genes listed above. VRs were classed as partial proviruses if they only contained a subset of the seven key NCV marker genes, the presence of the MCP and DNA polymerase B genes being particularly strong indicators of a partial provirus.

To classify genes in VRs (Figure 6B), viral sequences were removed for the NBCI RefSeq non-redundant protein database (NR) by removing proteins assigned to the "Viruses" category and by comparing the database with RVDB using BLASTp and removing any proteins that matched with an e-value cut-off of < 1-e40 to create a "virus-free NR" database. Deduced proteins were then compared with the RVDB and the virus-free NR databases using BLASTp and relative bitscores (rbitscores) were calculated by dividing the BLASTp bitscore for the best match in each database by the query protein's self-hit bitscore.[186] Self-hit scores were acquired by comparing the complete deduced proteomes with themselves using BLASTp. Proteins with a RVDB rbitscore at least 20% greater than its virus-free NR rbitscore were designated as "viral". Proteins with a virus-free NR rbitscore at least 20% greater than its RVDB rbitscore were designated as "cellular" (i.e. corresponding to a gene from a cellular organism). Ambiguous cases without a 20% differential were designated as "viral or cellular" and proteins with no significant matches were designated as ORFans (i.e. unknown proteins).

The presence of host regions flanking the viral regions was evaluated based on the ViralRecall output (Table S5C). The percentages of viral regions with two, one or zero flanking regions (longer than 2 kbp) were 25.8%, 15.0% and 59.2%, respectively (i.e. 40.8% of viral regions had at least one flanking region). Of the viral regions that had two flanking regions, 89.5%, 7.0% and 3.5% had flanking regions with a total length of >200 kbp, between 20 and 200 kbp or between 2 and 20 kbp, respectively. For the viral regions that had one flanking region, the corresponding percentages were 25.3%, 36.7% and 38.0%.

Phylogenetic analysis of viral genes

Amino acid sequences of manually-curated collections of major capsid protein (MCP) and DNA polymerase B proteins were aligned using MAFFT (v7.520, settings: –adjustdirectionaccurately –auto –maxiterate 1000) and phylogenetic trees were generated using IQ-TREE (v 2.2.2.3, settings: -m MFP -B 1000).

Metabolic networks

Genome-scale metabolic networks were reconstructed using AuCoMe[46] version 0.5.1 using the MetaCyc[187] version 26 database. A first dataset, consisting of the 60 species listed in Table S1F (column "Metabolic networks") plus two public diatom genomes

already used in the initial AuCoMe study (*Fragilariopsis cylindrus* and *Fistulifera solaris*) was processed to build the largest possible database (phaeogem) for exploratory comparisons (https://gem-aureme.genouest.org/phaeogem/). Then, a second comparison was performed on all long-read species plus outgroups. Based on Multidimensional-scaling (MDS) analyses, the most divergent long-read species (*Choristocarpus tenellus*, *Laminaria digitata*, *Phaeothamnion wetherbeei* and the public genome of *Sargassum fusiforme*) were excluded to construct a 16 species dataset, balancing assembly quality and phylogenetic coverage (https://gem-aureme.genouest.org/16bestgem/). MDS plots were build using the vegan package, version 2.6-4 (https://github.com/vegandevs/vegan) with R 4.1.2,[162] using Jaccard distances. A third stricter dataset (fwgem), enriched in high-quality long-read Ectocarpales, was built to address questions related to freshwater adaptation (https://gem-aureme.genouest.org/fwgem/). A set of reactions that were overrepresented in brown algae compared to the outgroup was created by taking reactions present in 100% of brown algae and less than 70% of outgroups. Reactions corresponding to genes lost in freshwater species were also extracted. These reaction sets were extracted from all the networks using the Aucomana library (https://github.com/PaulineGHG/aucomana). Online wikis (phaeogem, 16bestgem and fwgem) were generated using AuReMe.[188]

CAZymes

CAZyme genes were identified based on shared homology with biochemically characterised proteins, either individually or as hidden Markov model (HMM) profiles. For phylogenetic analyses, proteins were aligned using MAFFT[130] with the iterative refinement method and the scoring matrix Blosum62. The alignments were manually refined and trees were constructed using the maximum likelihood approach. Alignment reliability was tested by a bootstrap analysis using 100 resamplings of the dataset. Only bootstrap values above 60% are shown. The phylogenetic trees were displayed with MEGA.[131] The annotated genes are listed in Table S4B with accession numbers.

Sulfatases

The sulfatases encoded by each brown algal genome were identified and assigned to their respective family and subfamily using the SulfAtlas database[173,174] (https://sulfatlas.sb-roscoff.fr/). Each predicted proteome was first submitted to the SulfAtlas HHM server (https://sulfatlas.sb-roscoff.fr/sulfatlashmm/), which allows rapid identification of sulfatase candidates and (sub)family assignment using hidden Markov model profiles for each SulfAtlas (sub)family. Each sulfatase candidate sequence was then used as a query in a BLASTp[84] search against the SulfAtlas database (https://blast.sb-roscoff.fr/sulfatlas/). Sequences with at least 50% identity with sulfatases from marine bacteria or other marine microorganisms were considered to be contaminants. Below this threshold, additional examination of the predicted gene structure and genomic context of the candidate sequence was undertaken to identify possible horizontal gene transfers.

Haloperoxidases

vHPO genes were identified based on sequence homology and active site conservation. Maximum likelihood phylogenetic analyses were carried out using the NGphylogeny platform at https://ngphylogeny.fr/. MAFFT was used to align vHPO sequences and alignments were automatically curated with TrimAl,[133] leading to the selection of 444 informative positions from the initial 1450 positions for the algal-type vHPOs and 402 informative positions from the initial 1078 positions for the bacterial-type vHPOs. Maximum likelihood trees were constructed using FastTree with the WAG+G gene model and 1000 bootstrap replicates. Maximum likelihood Newick files were formatted as circular representations using iTOL. Only bootstrap values between 0.7 and 1 were conserved. The lists of annotated vHPO genes are in Tables S4C and S4D.

Ion channels

A search was carried out for 12 classes of ion channel in the predicted proteomes of the 21 Phaeoexplorer reference genomes plus those of two diatoms, *Phaeodactylum tricornutum* and *Thalassiosira pseudonana*. Predicted proteomes were screened using BLASTp[84] and query sequences from *Ectocarpus* species 7 and seven other species from diverse eukaryotic taxa.

Membrane-localised proteins

Membrane protein family genes were identified either by carrying out BLASTp[84] searches of the predicted Phaeoexplorer proteomes using *Ectocarpus* species 7 sequences as queries or by recovering orthogroups containing the relevant *Ectocarpus* species 7 sequences as members. The BLASTp approach was used for DEK1-like calpains, fasciclins, tetraspanins, CHASE, ethylene-binding-domain-like and MASE1 domain histidine kinases whereas the orthogroup approach was used to recover other members of the histidine kinase family. Both approaches were used to search for integrins and transmembrane receptor kinases. For integrins the two methods detected exactly the same set of proteins. For receptor kinases the BLASTp and orthogroup analyses detected 99.3% and 98.3% of the 269 genes, respectively. For these analyses, either the whole genome dataset was analysed or only the set of 21 reference genomes (Table S1F), depending on the size of the gene family.

Manually-curated histidine kinase protein families were aligned with Muscle[115] before phylogenetic tree construction using IQ-TREE 2[163] (version 2.3.4) with automatic model selection and 1000 bootstraps.

Transcription-associated proteins

TAPscan v4[135] was used to analyse the transcription-associated protein (TAP) complements of 21 species. TAPscan[134] is a comprehensive tool for annotating TAPs based on the detection of highly conserved protein domains using HMM profiles with specific thresholds and coverage cut-offs. Following detection, specialised rules are applied to assign protein sequences to TAP families based on the detected domains. TAPscan v4 can assign proteins to 138 different TAP (sub)families with high accuracy.

EsV-1-7 domain proteins

EsV-1-7 domain proteins were identified in the 31 brown algal and sister taxa genomes (Table S1F) by recovering the members of all orthogroups (with the exception of OG0000001, which is a very large OG that consisting principally of transposon sequences) that either contained one or more of a curated set of 101 EsV-1-7 domain proteins[62] for *Ectocarpus* species 7 or contained an EsV-1-7 domain protein based on a match to the Pfam EsV-1-7 motif PF19114. The recovered proteins were screened manually for the presence of at least one EsV-1-7 domain and a total of 2018 were finally identified as members of the EsV-1-7 family.

To identify orthologues of the EsV-1-7 protein IMMEDIATE UPRIGHT[62] (IMM), BLASTp searches of 25 brown algal and four sister taxa predicted proteomes were carried out with the amino-terminal domain of the IMM protein minus the five EsV-1-7 repeats as this domain is unique to IMM. Proteins were retained as IMM orthologues if they were more similar to IMM than to the most closely-related protein in *Ectocarpus* species 7, Ec-17_002150.

Histones

Histone protein sequences were analysed in *Ascophyllum nodosum*, *Chordaria linearis*, *Chrysoparadoxa australica*, *Desmarestia herbacea*, *Dictyota dichotoma*, *Discosporangium mesarthrocarpum*, *Ectocarpus crouaniorum*, *Ectocarpus fasciculatus*, *Ectocarpus siliculosus*, *Fucus serratus*, *Heterosigma akashiwo*, *Pleurocladia lacustris*, *Porterinema fluviatile*, *Pylaiella littoralis*, *Saccharina latissima*, *Sargassum fusiform*, *Schizocladia ischiensis*, *Scytosiphon promiscuus*, *Sphacelaria rigidula*, *Tribonema minus* and *Undaria pinnatifida* using BLASTp against the complete predicted proteomes (https://blast.sb-roscoff.fr/phaeoexplorer/) with the histone protein sequences from the diatom *Phaeodactylum tricornutum* as queries. The genes and transcripts coding for the identified histones were then retrieved from the genomes and predicted transcripts using BLAST (https://blast.sb-roscoff.fr/phaeoexplorer/). The proteins encoded by the identified genes and transcripts were predicted with the Expasy web translator (https://web.expasy.org/translate/). In order to identify truncated proteins or incorrect start codons, the following constraints were applied: H2A proteins must start with the SGKGKGGR sequence, H2B with AKTP, canonical H3.1 and variants H3.3 with ARTKQT and H4 with SGRGKGGKGLGKGG. For the linker histone H1, protein sequences had to be lysine-rich and sequences with incorrect start codons were determined by alignments of all identified H1 proteins. For proteins with incorrect start codons, the region upstream of the correct start codon was removed. For truncated proteins, *i.e.* proteins whose transcripts lacked either the start (no methionine) or stop codons, the protein sequence was completed based on alignment with the corresponding genomic region using the Geneious 11.0.5 software. When the sequence could not be completed, a BLAST was performed against the Phaeoexplorer *de novo* transcriptomes (https://blast.sb-roscoff.fr/phaeoexplorer_denovo/) when this data was available (this was not possible for the public genomes *T. minus*, *U. pinnatifida* and *S. fusiforme*). Based on the nomenclature established by,[189] H3 histones were classified as follows: canonical H3.1 proteins harbour AT residues at positions 31-32 while histone variants H3.3 harbour TA residues, H3 proteins with other residues at positions 31-32 were named H3.4 and so on. CenH3 variants of H3 were identified by analysis with Panther 17.0 (www.pantherdb.org/tools/sequenceSearchForm.jsp?) and/or Interproscan[137] 94.0 (www.ebi.ac.uk/interpro/search/sequence/).

Species abbreviations used in histone phylogenetic trees are: Atr, *Amborella trichopoda*; At, *Arabidopsis thaliana*; Ce, *Caenorhabditis elegans*; Di, *Dictyostellium discoideum*; Dr, *Danio rerio*; Dm, *Drosophila melanogaster*; Hs, *Homo sapiens*; Mm, *Mus musculus*; Pp, *Physarum polycephalum*; Ppa *Physcomitrium patens*; Sc, *Saccharomyces cerevisiae*; Tm, *Tetrahymena thermophila*; Zm, *Zea mays*; Mp, *Marchantia polymorpha* subsp. *Ruderalis*; Bd, *Brachypodium distachyon*; Ccr, *Chondrus crispus*; Gs, *Galdieria sulphuraria*; Cm, *Cyanidioschyzon merolae*; Cr, *Chlamydomonas reinhardtii*; Ol, *Ostreococcus luciminarinus*; Ot, *Ostreococcus tauri*; To, Thalassiosira oceanica; Pt, *Phaeodactylum tricornutum*; An, *Ascophyllum nodosum*; Cl, *Chordaria linearis*; Ca, *Chrysoparadoxa australica*; Dh, *Desmarestia herbacea*; Ddi, *Dictyota dichotoma*; Dme, *Discosporangium mesarthrocarpum*; Ec, *Ectocarpus crouaniorum*; Ef, *Ectocarpus fasciculatus*; Es, *Ectocarpus siliculosus*; Fse, *Fucus serratus*; Ha, *Heterosigma akashiwo*; Pla, *Pleurocladia lacustris*; Pf, *Porterinema fluviatile*; Pli, *Pylaiella littoralis*; Sl, *Saccharina latissima*; Sf, *Sargassum fusiform*; Si, *Schizocladia ischiensis*; Sp, *Scytosiphon promiscuus*; Sri, *Sphacelaria rigidula*; Tm, *Tribonema minus*; Up, *Undaria pinnatifida*.

DNA methyltransferases

Searches were carried out for methyltransferases and demethylases in the predicted proteomes of 20 of the high quality brown algal reference genome assemblies (based on Nanopore long-read sequence) plus the sister taxa *Chrysoparadoxa australica* and *Schizocladia ischiensis* using BLASTp (Table S1F). A methyltransferase reference database was constructed by recovering sequences from NCBI, ENSEMBL and UniProtKB. Methyltransferase sequences were recovered for stramenopiles such as *Nannochloropsis gaditana*, the diatom *Phaeodactylum tricornutum*, the oomycete *Phytophthora infestans* and for species from more distant lineages including *Arabidopsis thaliana*, *Homo sapiens* and the fungus *Neurospora crassa*. The proteomes of the selected brown algal strains were then queried against this database using BLASTp and matches with an e-value of < 0.001, a bit score > 70, a maximum gap of 5 and percentage identity of >30% were retained. The retained matches were screened against the NCBI, UniProt and

SwissProt databases to identify and remove contaminating bacterial or viral proteins. Methyltransferase domains were detected in the retained matches using the Simple Modular Architecture Research Tool (SMART)[177] domain architecture analysis and InterPro searches (https://www.ebi.ac.uk/interpro/). Sequences with methyltransferase domains were retained for further analysis. Validated brown algal methyltransferases were aligned with reference methyltransferases using Clustal[138] 2.1.

Spliceosome

Components of the Major Spliceosome were identified using a reference set of 147 human components (https://www.genenames.org/data/genegroup/#!/group/1518), excluding the five small nuclear RNAs (snRNAs). Including isoforms, this query set consisted of 626 proteins. These proteins were used to screen the predicted proteomes of 54 genomes (Table S1F) using BLASTp and matches were retained if they had an e-value of at most $1e^{-40}$ and coverage >30%. Searches were also carried out for components of LSm and Sm complexes which have roles as scaffolds in the formation of ribonucleoprotein particles (RNPs), in the maturation of mRNAs (including splicing, such as the cytoplasmic complex LSm1-7, LSm2-8 which is part of the core U6 snRNP and other complexes important for the formation of the 3' ends of histone transcripts), in the assembly of P-Bodies and in the maintenance of telomeres.

Flagella proteins

A previous proteomic analysis of anterior and posterior flagella of the brown alga *Colpomenia bullosa* identified a total of 592 proteins across the two proteomes.[41] Here the *Ectocarpus* species 7 orthologues of 70 of these proteins that had been detected with a very high level of confidence were used to identify the corresponding orthogroups and the presence or absence of these orthogroups was scored for seven representative species (Table S1F).

Detection of *Porterinema fluviatile* genes differentially expressed in freshwater and seawater

Six independent cultures of *Porterinema fluviatile* were cultivated for four weeks in 140 mm Petri dishes with Provasoli-enriched culture medium,[190] which was renewed every two weeks. For three Petri dishes, the culture medium was based on autoclaved natural seawater (high salinity treatment), for the other three Petri Dishes natural seawater was diluted 1:19 vol/vol with distilled water (low salinity treatment). Cultures were harvested with 40 μm nylon sieves, dried with a paper towel, and immediately frozen in liquid nitrogen. RNA extraction library construction and sequencing were carried out as described in section "RNA extraction, Illumina RNA-seq library preparation and sequencing". RNA-seq reads were cleaned with Trimmomatic[99] V0.38 and then mapped to the *P. fluviatile* genome using Kallisto[140] version 0.44.0. Differentially expressed genes were identified using the DESeq2 package[141] included in Bioconductor version 3.11, considering genes with an adjusted $p < 0.05$ and a $\log_2$ fold-change > 1 as differentially expressed. To compare the differentially expressed genes in *P. fluviatile* with an equivalent set previously identified for *Ectocarpus subulatus* in a microarray experiment using nearly identical growth conditions,[191] orthologues in the two species were detected using Orthofinder version 2.3.3. Of the 10,066 shared orthogroups, 6,606 had microarray expression data for *E. subulatus*. This information was used to classify differentially expressed genes for the two species as either shared orthologues or as lineage-specific.

Identification of genes with generation-biased expression patterns

RNA-seq data (two to five replicates per condition) was recovered for gametophyte and sporophyte generations of ten species (Table S1F). Data quality was assessed with FastQC[142] version 0.11.9 and sequences were then trimmed with Trim Galore version 0.6.5 with the parameters –length 50, - quality 24, –stringency 6, –max_n 3. The cleaned reads were mapped onto the corresponding genome for each species using HISAT2 version 2.1.0 with default options. Counting was carried out with featureCounts[145] from the subread package (version 2.0.1) on CDS features grouped by Parent. Transcript Per Kilobase Million (TPM) tables were generated for all conditions and differentially expressed genes were detected using DESeq2[141] version 1.30.1. Genes were classified into six categories based on the differential expression analysis and the TPM values: gametophyte-biased, mean TPM ≥1 in gametophyte and sporophyte, $\log_2$(fold change) ≥1, adjusted *p*-value <0.05; sporophyte-biased: mean TPM ≥1 in gametophyte and sporophyte, $\log_2$(fold change) ≤-1, adjusted *p*-value <0.05; gametophyte-specific, mean TPM <1 in sporophyte and ≥1 in gametophyte, $\log_2$ (fold change) ≥1, adjusted *p*-value <0.05; sporophyte-specific, mean TPM <1 in sporophyte and ≥1 in gametophyte, $\log_2$(-fold change) ≤-1, adjusted p-value <0.05; unbiased genes: mean gametophyte and sporophyte TPMs ≥1, $\log_2$(fold change) <1 or >-1 and/or adjusted p-value ≥0.05; unexpressed genes, mean gametophyte and sporophyte TPM <1.

Life cycle and thallus architecture

Genome dataset and traits

To study the impact of body architecture, the brown algae were divided into three categories: 22 filamentous species, eight simple parenchymatous species and 13 species with elaborate thalli (Table S1F). For the life-cycle-based assessment, the groups were: 30 haploid-diploid species and six diploid species (Table S1F). Body architecture information was available for 43 species, and life cycle information was available for 36 species; species without body plan or life cycle information were not used in subsequent analyses. Two approaches were used to estimate selection intensity across the phylogeny, (i) a model-based method, and (ii) by evaluating codon usage bias and nucleotide composition. Two evolutionary models were used, one based on architecture and the other based on life cycle. For model-based methods the phylogeny was categorised based on the above traits, and selection intensity parameters were estimated using PAML[146] version 4.9i. Rate estimates were obtained for non-synonymous substitutions (dN), synonymous

substitutions (dS) and omega (dN/dS) for the multiple sequence alignments of all genes within each orthogroup using the variable-ratio model of CODEML from PAML, which allows different omegas for different branch categories. The traits were assigned to the branches of the phylogeny using ancestral state estimation by stochastic mapping with the phytools R package.[147,162]

Evolutionary models to study impacts of body architecture

To study variation in selection intensity as a function of body architecture, we devised a model with the following trait categories: filamentous/pseudoparenchymatous (simple cell division and organisation on a single plane), parenchymatous (cell division and organisation on multiple planes) and elaborate thallus (tissue differentiation). To ensure that at least 50% of the species in each category were used in the analysis, we selected orthogroups (OGs) that contained at least 11 members for filamentous, at least four members for parenchymatous and at least six members for elaborate thallus algae. Using this filter, 1068 OGs were obtained, on which the model based on body architecture was fitted. Selection intensity parameters [rate of non-synonymous substitution (dN), rate of synonymous substitution (dS) and omega (dN/dS)] were estimated for the three trait categories for each gene alignment. We used the Wilcoxon signed-rank test to evaluate the statistical significance of differences between the selection intensity parameters (dN, dS and dN/dS) for each category.

Evolutionary models to study the impacts of life cycle

The impact of life cycle on molecular evolution was assessed using a model with two categories consisting of diplontic and haplodiplontic species. For this model we used 1,058 OGs that contained at least three members for diploid species and at least 15 members for haploid-diploid species. Using alignments of the gene within the OGs, we estimated the selection intensity parameters for the different categories and applied the Wilcoxon signed-rank test to assess the statistical significance of differences in selection intensity between the diploid and haploid-diploid life cycles.

Selection of intensity parameters

Omega (dN/dS) provides an estimate of the ratio of substitutions at sites under selection compared to neutral sites, and is generally used to infer the strength of purifying selection. Omega needs to be interpreted with caution because not all synonymous sites are neutral[192] and also synonymous substitutions are often underestimated due to saturation of synonymous sites, which might in turn impact the omega ratios.[193] Omega values lower than one indicate substitutions are less frequent at sites under selection compared to neutral sites and are characteristic of highly conserved genes or genes evolving under strong purifying selection. As we used primarily low copy number genes in this study, the analysed genes were expected to evolve under strong purifying selection, with omega values much lower than one. Using omega for near neutral studies is challenging because near neutral sites are determined by effective population size, that is to say, sites under mild selection constraint in larger populations can behave as neutral sites in smaller populations. It is therefore difficult to infer the amount of mutation from relative values of omega. In order to obtain better insight into selection intensity, mutation accumulation was not only investigated using rates of synonymous (dS) and non-synonymous (dN) substitutions but also by estimating codon bias and nucleotide composition. Codon usage bias was used, in addition to omega, to infer selection intensity across species as the former reflects selection efficacy at synonymous sites.[194–196] We inferred codon usage bias by estimating the effective number of codons (ENC) for each species using the enc method from the VHICA package.[148,182] The effective number of codons (ENC) quantifies the extent of deviation of codon usage of a gene from equal usage of synonymous codons. For the standard genetic code, ENC values range from 20 (where a single codon is used per amino acid implying strong codon usage bias) to 61 (implies that all synonymous codons are equally used for each amino acid[197]). Low ENC indicates constrained use of codons, which potentially highlights stronger codon bias due to stronger selection at synonymous sites. As nucleotide composition can also influence codon bias, we calculated the overall GC composition, GC at the third codon position (GC3) and the theoretical expected ENC (EENC) based on GC3 using local R scripts. The lower the observed ENC (OENC, estimated from the gene sequence) relative to EENC, the stronger the influence of selection due to translation on codon usage. This was studied by estimating the difference (DENC = EENC - OENC) between the expected ENC and the observed ENC.[198] Positive DENC indicates a role for selection constraints on codon usage in addition to the influence of nucleotide composition. DENC values of zero or less indicate that codon bias is entirely driven by nucleotide composition. DENC values were used to study the influence of translation selection and nucleotide composition on codon usage bias.

Assembly and analysis of organellar genomes

Plastid and mitochondrial genomes were assembled *de novo* using NOVOPlasty[149] v3.7 and *rbcL* and *cox1* nucleotide sequences as seeds. Assembled genomes were checked by aligning reads using Bowtie2[86] v2.3.5.1 and processed with SAMtools[150] v1.5. Annotation of protein-coding genes was performed with GeSeq[151] v2.03. Annotation of tRNAs, tmRNAs and rRNAs was performed with ARAGORN[152] v1.2.38.

Maximum-likelihood (ML) phylogenetic trees were constructed using 92 plastid genomes (11 non-brown outgroup sequences) and 89 mitochondrial genomes (seven non-brown outgroup sequences). The conserved coding-region amino acid sequences of 139 plastid genes (31,159 amino acids) and 35 mitochondrial genes (7,461 amino acids) were used to construct these phylogenetic trees. The sequence for each gene was aligned individually using MAFFT[130] v7 (–maxiterate 1000) and then concatenated. Alignment partitions were assigned based on genes. Each of the aligned gene sequences was trimmed with trimAl[133] v1.2 (-automated1). ML phylogenetic trees were constructed with IQ-TREE 2.[163] The protein substitution models in each gene partition were selected using ModelFinder.[153] Statistical support for tree branches was assessed with 1,000 replicates of ultrafast bootstrap (UFBoot2).[154]

Analysis of *Ectocarpus* genome synteny

Global genome synteny analysis was performed using SynMap[155] on the CoGe platform (https://genomevolution.org/coge/) with the following genomes: *Ectocarpus crouaniorum* male, *Ectocarpus fasciculatus* male, *Ectocarpus siliculosus* male, *Ectocarpus* species 7 male and *Ectocarpus subulatus*. SynMap identifies syntenic regions between two or more genomes using a combination of sequence similarity and collinearity algorithms. Last[199] was used as the BLAST algorithm and syntenic gene pairs were identified using DAGChainer[156] with settings "Relative Gene Order", -D = 20, -A =5. Neighbouring syntenic blocks were merged into larger blocks. Substitution rates between the synthetic CDS pairs were calculated using CodeML,[146] which was also implemented in SynMap, CoGe. In detail, protein sequences were aligned using the Needleman-Wunsch algorithm implemented in nwalign (https://pypi.org/project/nwalign/) and then translated back to aligned codons. CodeML was run five times for each alignment using the default parameters and the lowest dS was retained, with the upper cutoff for dS values set at 2. *Ectocarpus* genes were grouped according to their age based on the phylostratigraphic analysis and by chromosomal location based on their chromosome position in *Ectocarpus* species 7. All plots and statistical analysis were carried out in R version v.4.3.1. Local synteny analysis was based on orthologous genes as identified by Orthofinder.

Analysis of *Ectocarpus* gene evolution

Protein sequence alignments were used to remove gaps with trimAl[133] and then translated back to DNA with backtranseq.[200] Only DNA fasta files with a minimum of 70 bp were retained (831 single-copy orthologs). PhyML trees were built with Geneious v11.1.5 (https://www.geneious.com). Maximum likelihood analysis was carried out to detect site specific, branch-site specific and branch specific positive selection as well as sites under negative selection, using PAML.[201]

Phylogenetic analysis of *Ectocarpus* species

Phylogenetic analysis was carried out for 11 *Ectocarpus* species plus *Scytosiphon promiscuus* as an outgroup (Table S1F). Of the 933 single-copy orthogroups identified for these 12 species, 261 high-confidence alignments were retained for gene tree and species tree inferences following the removal of low-quality alignments using BMGE.[202] Bayesian inference of the phylogeny of the *Ectocarpus* species complex was performed using BEAST[157] v2.7. The analysis was conducted under the multi-species coalescent (MSC) model, implemented in StarBEAST3[158] v1.1.7. The MSC model coestimates gene trees and the species tree within a multispecies coalescent framework, enabling the assessment of incongruences among genes with respect to the species tree. To account for substitution model uncertainty, bModelTest[159] was employed to average over a set of substitution models for each alignment. StarBEAST3 was run under both the Yule model and the strict clock model. A total of 300,000,000 Markov Chain Monte Carlo (MCMC) generations were conducted, with tree states stored every 50,000 iterations. Posterior tree samples were combined, discarding the initial 10% burn-in, using LogCombiner v2.4.7. A maximum clade credibility tree was generated using TreeAnnotator[157] v2.4.7.

Ectocarpus introgression analysis

To distinguish introgression from shared ancestry, D estimates (i.e. ABBA-BABA tests) were generated from 36 four-taxon combinations[203]: four to test the level of introgression within clade 1 (i.e. *E. subulatus*, *E. crouaniorum*, *Ectocarpus* species 1, *Ectocarpus* species 2), 20 to test the level of introgression within clade 2 (i.e. *Ectocarpus* species 6, *Ectocarpus* species 7, *Ectocarpus* species 5, *Ectocarpus* species 9, *E. siliculosus*, *Ectocarpus* species 3) and 12 to test the level of introgression between these two clades. Tests were designed using a four-taxon fixed phylogeny (((P1,P2)P3)O), where P1 and P2 are closely related species from the same clade, P3 is a more divergent species that may have experienced admixture with one or both of the (P1,P2) taxa, and an out-group (O). *E. fasciculatus* was used as the out-group taxon for all ABBA-BABA tests. Details about how P1, P2 and P3 taxa were selected for each test are given in Table S6. Previous results of species tree inference were used to inform subsequent ABBA–BABA tests and to define the (((P1,P2)P3)O) phylogenies. ABBAs are sites at which the derived allele (called B) is shared between the taxa P2 and P3, whereas P1 carries the ancestral allele (called A), as defined by the outgroup while BABAs are sites at which the derived allele is shared between P1 and P3, whereas P2 carries the ancestral allele. Under incomplete lineage sorting, conflicting ABBA and BABA patterns should occur in equal frequencies, resulting in a D statistic equal to zero. Historical gene flow between P2 and P3 causes an excess of ABBA, generating positive values of D. Historical gene flow between P1 and P3 causes an excess of BABA, generating negative values of D. Patterson's D-statistic was calculated for the concatenated alignments of the 261 orthologroups. Significance was detected using a block-jackknifing approach,[203–205] with a block size of 5 kbp. For the jackknife procedure, one block of adjacent sites was removed n times. A Z-score was finally obtained by dividing the value of the D statistic by the standard error over n sequences of 5 kbp. The ParimonySplits network was reconstructed for the genus *Ectocarpus* using SplitsTree 4[160] (version 4.14.6) with 1000 bootstrap replicates.

QUANTIFICATION AND STATISTICAL ANALYSIS

All statistical analyses are described in detail in the relevant sections of the "method details" section and the results of statistical tests are shown in the tables and figures.

ADDITIONAL RESOURCES

The Phaeoexplorer website (https://phaeoexplorer.sb-roscoff.fr) provides access to all the annotated genome assemblies described in this study as downloadable files. The output files from the Orthofinder,[113] Interproscan,[137] Hectar[161] and eggNOG-mapper[125] analyses, together with the results of the various DIAMOND[107] sequence similarity analyses (see section "Analyses aimed at deducing functional characteristics of predicted proteins"), can also be downloaded. In addition, the site provides genome browser interfaces for the genomes and multiple additional tools and resources including BLAST interfaces for genomes, proteomes and *de novo* transcriptomes, various experimental protocols, an AskOmics genomic data query interface (PhaeoAskOmics), an RShiny-based transcriptomic aggregator for the model brown alga *Ectocarpus* species 7 strain Ec32, a link to genome-wide metabolic networks for the Phaeoexplorer species and a list of project-related publications.

Additional data and results have been deposited in the CNRS Research Data depository under the title "Data for Phaeoexplorer publication: Evolutionary genomics of the emergence of brown algae as key components of coastal ecosystems" (DOI: https://doi.org/10.57745/9U1J85). Dataset description: "The Phaeoexplorer project sequenced 60 genomes corresponding to 44 brown algal and sister species. This dataset corresponds to supplementary information relating to the initial annotation of the Phaeoexplorer genomes and multiple analyses of the genome data. The dataset includes additional results of the project, together with accompanying additional figures and tables, (Additional_results.tar.gz), presubmission (v0) versions of the Phaeoexplorer genome annotation (GFF) files (GFF_v0.tar.gz) and genome-wide predicted proteomes as fasta files (Proteomes_v0.tar.gz), de novo transcriptome assemblies for the Phaeoexplorer species (RNA-seq data assembled with Trinity or rnaSPAdes; de-novo-transcriptomes.tar.gz), RepeatMasker analyses of repeat sequences (RepeatMasker.tar.gz), alignment files used to generate a phylogenetic tree for the Phaeoexplorer species (PhylogeneticTree.tar.gz), alignments used to build a densitree specifically for *Ectocarpus* species (Microevolution_Ectocarpus.tar.gz), an Orthofinder-based analysis of shared orthologues (Orthogroups.tar.gz) together with a Dollo-logic-based analysis of orthogroup gain and loss during evolution (Dollo_analysis.tar.gz), a Phylostratigraphy analysis of brown algal genes (Phylostratigraphy.tar.gz), an analysis of protein functional domain fissions and fusions (CompositeGenes.tar.gz), Interproscan analyses of protein domains (InterProScan.tar.gz), Hectar predictions of protein subcellular localisations (Hectar.tar.gz), eggNOG output providing information about predicted protein functions (eggNOG.tar.gz), RNA-seq-based data on gene expression levels (mRNAexpression.tar.gz), results of a search for genes acquired via horizontal gene transfer (HGT.tar.gz), analyses of intron conservation across genomes (Introns_conservation.tar.gz), an analysis of tandem gene duplications (Tandemely_duplicated_genes.tar.gz), CAFE5 reconstruction of gene family amplifications (CAFE5.tar.gz), comparisons of CDS size with the *Ectocarpus* reference genome that were used to evaluate gene model completeness (CDS_size.tar.gz), a DESeq2 analysis of differential gene expression between the sporophyte and gametophyte generations of several brown algal species (DEG_LifeCycle.tar.gz), information about orthogroups selected to analyse the effects of morphological complexity and life cycle structure on gene evolution (Genes_selection.tar.gz). Each individual dataset contains a README file explaining its content. Detailed information about the methodology used for each analysis can be found in the STAR Methods section of the manuscript preprint (https://doi.org/10.1101/2024.02.19.579948). The majority of these analyses and datasets can also be accessed via the Phaeoexplorer website (https://phaeoexplorer.sb-roscoff.fr/)."

4. Punctuated, repeated evolution of remodelled genes in the animal kingdom

We also took part in a research project that studied the impact of gene remodelling on the evolution of animals. This study was led by Mary O'Connell and James McInerney at the University of Nottingham, and a draft article has been written to present the results, which is currently in the process of submission. Our main contribution to this work consisted in applying the polarisation method that we developed to composite families that had already been identified. This allowed for a finer understanding and interpretation of the results, as composite genes would otherwise be ambiguously attributed to gene fusions or gene fission events.

This study was based on a dataset of ~1.2 million protein-coding genes, from a set of 63 species covering all major clades of metazoans. We found that composite genes represented around 5% of all genes in animal genomes, with gene fusions responsible for the emergence of 73.3% of all composite genes, and 25.4% of composites corresponding to fission events. Only a fifth (21%) of fusion composites were compatible with the scenario of a single fusion event with no ulterior remodelling, whereas the other 79% of gene fusions showed signs of having undergone later events of fission or other remodelling. This suggests that gene remodelling in animals is a particularly dynamic process that frequently revisits gene families that had already been involved in previous remodelling events. The SSN of animal gene families is also highly modular, which corroborates this observation: 87% of composite families are also components for another composite, and even contribute, on average, to more remodelling events than non-composite families that are also components. Remodelling events thus appear to involve a specific pool of families in animal genomes that are regularly reused towards new genetic rearrangements. The high levels of intermingling and recombination within this specific gene subset may also be responsible for another trend of gene remodelling in animals, which is its remarkable repeatability. Indeed, of all composite families in the dataset, 41% had a polyphyletic distribution in the species tree, suggesting that these composites may have evolved convergently in distinct lineages, which may reflect an adaptive advantage granted by these composites.

Rather than being evenly spread throughout the animal tree of life, gene remodelling events (or at least retained composite genes) seem to occur in punctuated bursts at specific nodes of the phylogeny (Figure 29). Deuterostomia (including chordates, hemichordates and echinoderms), in particular, display higher amounts of remodelled genes than Protostomia (including arthropods,

molluscs, and most worm-like animals) or non-bilaterians[16]. The maximal amount of remodelled gene gain at a single position in the animal phylogeny occurs on the branch leading to the ancestor of Euteleostomi (around 450 million years ago), coinciding with a number of important phenotypic changes such as the transition from a cartilaginous skeleton to one of mineralised bone, as well as an overall increase in genome complexity [Sacerdot et al. 2018, Simakov et al. 2020]. Relative rates of composite gene formation per time unit highlight some more recent lineages that have experienced particularly high rates of gene remodelling, especially in Hominoidea and Caenorhabditis. This punctuated gain pattern at specific points of the animal phylogeny contrasts with overall trends of gene family acquisition in animals, which predominantly occurred in branches leading to the common metazoan ancestor and soon after. This substantial wave of gene origination early in the evolution of animals could thus have created a repertoire of genetic "building blocks" that gene remodelling would have later exploited for further genetic innovation, especially at certain key points of animal evolution.

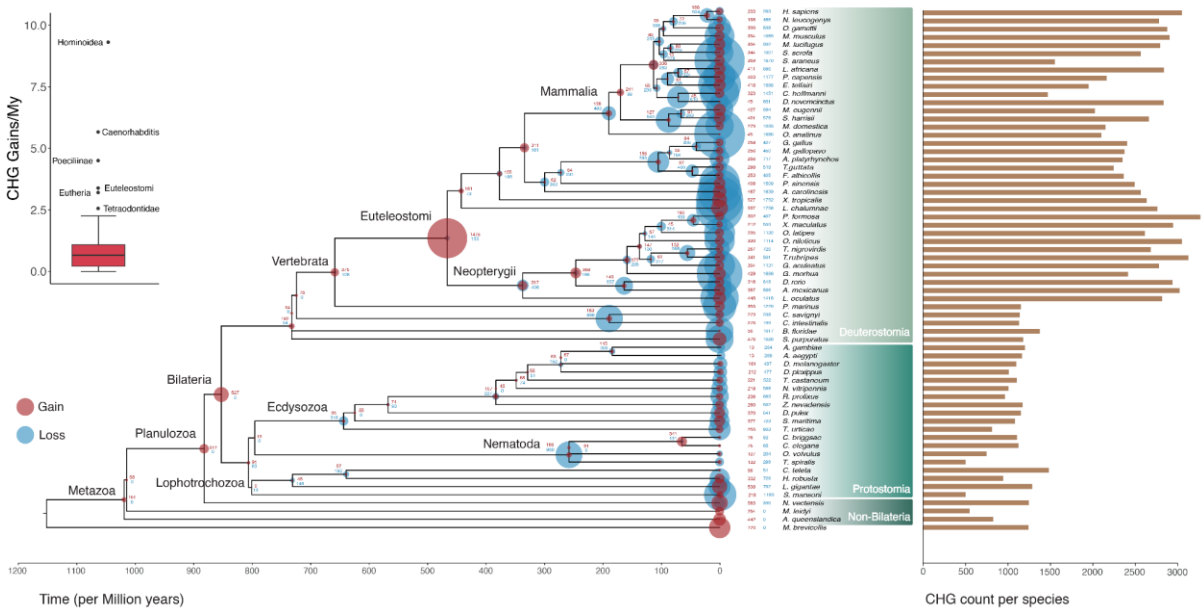


Figure 29: Emergence and loss of remodelled genes in the evolution of animals.
Phylogenetic distribution of the gains and losses of fused and split genes across the animal tree of life. The area of each circle is proportional to the amount of gain/loss at the corresponding node. The bar plot on the right shows the number of composite gene families in each extant genome. The inset box plot shows the distribution of the rate of composite gene gain per time unit.

[16] Deuterostomia and Protostomia are the two main clades of bilaterian animals that have a distinct bilateral symmetry at the embryonic stage. Most Bilateria maintain this symmetry as adults, with the exception of echinoderms, which become pentamerous as adults (e.g. starfish). Animals that are not bilateral include Porifera (sea sponges), Ctenophora (comb jellies), Placozoa, and Cnidaria (jellyfish, corals, anemones, etc.).

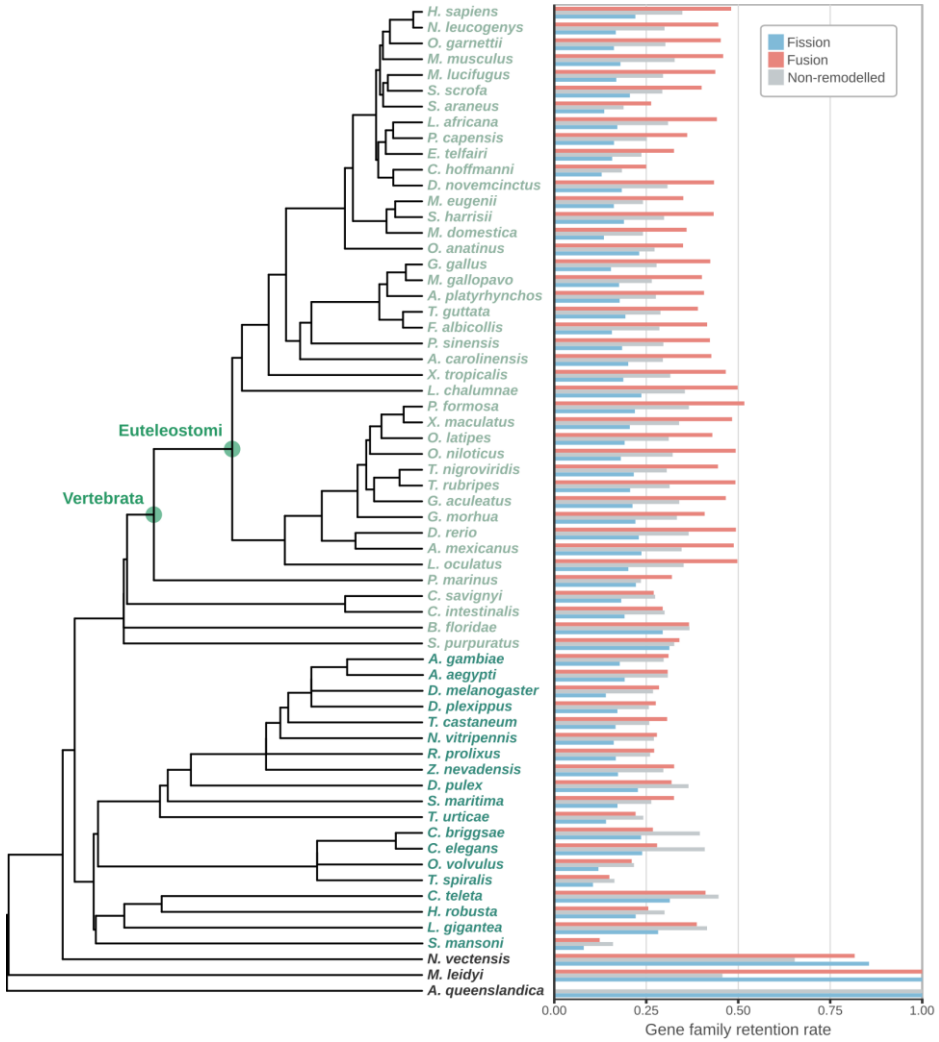


Figure 30: Retention of remodelled genes in extant animal genomes.
This bar plot indicates the rates of retention for fused, split and non-remodelled gene families in extant animal genomes. Species names are coloured according to their inclusion in Deuterostomia (light green), Protostomia (dark green) or non-bilaterian clades (black). The internal nodes corresponding to the emergence of Vertebrata and Euteleostomi are highlighted as they represent a shift in the pattern of retention, compared to non-vertebrate groups.

We also found that fused genes were more likely to be retained in present-day genomes of vertebrate species than non-remodelled genes, with Euteleostomi in particular showing markedly high rates of fusion gene retention (Figure 30). In non-vertebrate lineages, on the other hand, fused genes were lost at rates comparable (or sometimes higher, e.g. in Caenorhabditis) to non-remodelled genes. Conversely, split genes that originated from fission events were substantially less stable than fused and non-remodelled genes in most animal lineages. These results suggest that gene fissions and fusions have played distinct roles in the evolution of animals, with fissions creating gene products that were largely volatile and rarely selected for, whereas gene fusions may have resulted in more neutral or even beneficial (especially in vertebrates) genetic innovations. Lastly, from a functional standpoint,

fusion genes were found to be predominantly involved in transcription (COG category K), post-translational modification, protein turnover and chaperone functions (O), signal transduction (T), extracellular structures (W) and inorganic ion transport and metabolism (P). Functional innovations in these categories (especially K, O, T) could have contributed to a complexification of gene regulation pathways, which in animals are particularly associated with morphological development and the evolution of body plans [Davidson and Erwin 2006]. Likewise, the emergence of new functions associated with extracellular structures (W) may have played a role in the extant diversity of biological tissues and organs in animals. Although gene remodelling may not have been central to the initial emergence of animals [Ocaña-Pallarès et al. 2022], later remodelling events might therefore have contributed to the diversification of animal lineages. In particular, some major physiological and morphological changes in animal evolution, such as bilateral body plans and the axial endoskeleton of vertebrates, are coincident with bursts of gene gains from fusion events at specific nodes in the animal tree of life.

Bursts of novel composite gene families at major nodes in animal evolution

Peter O. Mulhair[1,2,a], Raymond J. Moran[1,3,a], Jananan S. Pathmanathan[4,5,a], Duncan Sussfeld[4,6], Christopher J. Creevey[7], Karen Siu-Ting[7], Fiona J. Whelan[8], Davide Pisani[9], Bede Constantinides[1,10], Eric Pelletier[6,11], Philippe Lopez[4], Eric Bapteste[4], James O. McInerney[8*], Mary J. O'Connell[1,2*]

[1]*Computational and Molecular Evolutionary Biology Group, School of Biology, Faculty of Biological Sciences, University of Leeds, Leeds LS2 9JT, United Kingdom.*

[2]*Computational and Molecular Evolutionary Biology Group, School of Life Sciences, Faculty of Medicine and Health Sciences, University of Nottingham, Nottingham, NG7 2RD, United Kingdom.*

[3]*Bioinformatics and Molecular Evolution Group, School of Biotechnology, Dublin City University, Glasnevin, Dublin 9, Ireland.*

[4]*Institut de Systématique, Evolution, Biodiversité (ISYEB), Sorbonne Université, CNRS, Museum National d'Histoire Naturelle, EPHE, Université des Antilles, 75005 Paris, France.*

[5]*Departments of Biochemistry & Molecular Biophysics and Biological Sciences, Columbia University, New York, NY, 10032, USA.*

[6]*Génomique Métabolique, Genoscope, Institut François-Jacob, CEA, CNRS, Univ Evry, Université Paris-Saclay, 91000 Evry, France.*

[7]*School of Biological Sciences/Institute for Global Food Security, Queen's University Belfast, Belfast, BT9 5DL, United Kingdom.*

[8]*School of Life Sciences, Faculty of Medicine and Health Sciences, The University of Nottingham, Nottingham, NG7 2RD, United Kingdom.*

[9]*Schools of Earth and Biological Sciences, University of Bristol, Life Sciences Building, Tyndall Avenue, Bristol, BS8 1TQ, United Kingdom.*

[10]*Nuffield Department of Clinical Medicine, John Radcliffe Hospital, University of Oxford, Oxford, OX3 9DU, United Kingdom.*

[11] *Research Federation for the Study of Global Ocean Systems Ecology and Evolution, FR2022/Tara Oceans GOSEE, 75016 Paris, France.*

[a] These authors contributed equally to this work

* Corresponding Authors: mary.o'connell@nottingham.ac.uk

and james.mcinerney@nottingham.ac.uk

Author Contributions: MJO'C and JMcI conceived of the study. MJO'C, JMcI, EB, PL, CC, RM, JP, and PM designed the specific experiments. PM, RM, and JP performed most experiments. RM, JP, PM, MJO'C, EB, JMcI, and PL designed, performed, and interpreted the homology and composite searching analyses. DS, EP and PM performed functional enrichment analyses, and DS, EB and PL performed the fission vs fusion analyses. PM, RM, JP, CC, KST, BC, PL and FJW contributed code for data analyses and data visualisation. MJO'C, PM, RM, JP, JMcI, DP, KST, DP and CC contributed to the interpretation of the results. PM, RM, JMcI and MJO'C took the lead in writing the manuscript. All authors provided critical feedback and helped shape the overall manuscript.

Competing interest statement: The authors declare no competing interests.

Classification: Biological Sciences, Evolution.

Keywords: Gene Fusion, Major animal groups, Convergent Evolution, Animal Evolution.

This file includes: Main text and Figures 1- 4, Supplementary Figures S1-6 and Supplementary Tables S1-2.

Abstract

A molecular level perspective on how novel phenotypes evolve is contingent on our understanding of how genomes evolve through time, and of particular interest is how novel elements emerge or are lost. Mechanisms of protein evolution such as gene duplication have been well established. Studies of gene fusion events show they often generate novel functions and adaptive benefits. Identifying gene fusion and fission events on a genome scale allows us to establish the mode and tempo of emergence of composite genes across the animal tree of life, and allows us to test the repeatability of evolution in terms of determining how often composite genes can arise independently. Here we show that ~5% of all animal gene families are composite, and their phylogenetic distribution suggests an abrupt, rather than gradual, emergence during animal evolution. We find that gene fusion occurs at a higher rate than fission (73.3% vs 25.4%) in animal composite genes, but many gene fusions (79% of the 73.3%) have more complex patterns including subsequent fission or loss. We demonstrate that nodes such as Bilateria, Euteleostomi, and Eutheria, have significantly higher rates of accumulation of composite genes. We observe that in general deuterostomes have a greater amount of composite genes as compared to protostomes. Intriguingly, up to 41% of composite gene families have evolved independently in different clades showing that the same solutions to protein innovation have evolved time and again in animals.

Significance statement

New genes emerge and are lost from genomes over time. Mechanisms that can produce new genes include, but are not limited to, gene duplication, retrotransposition, *de novo* gene genesis, and gene fusion/fission. In this work, we show that new genes formed by fusing distinct homologous gene families together comprise a significant portion of the animal proteome. Their pattern of emergence through time is not gradual throughout the animal phylogeny - it is intensified on nodes of major transition in animal phylogeny. Interestingly, we see that evolution replays the tape frequently in these genes with 41% of gene fusion/fission events occurring independently throughout animal evolution.

Introduction

Composite genes emerge by fusion of distinct protein coding sequences ("components"), or by the fission of protein coding sequences into components. Often composite genes establish novel domain architectures, expression profiles, and functions (1–7). For example, the fusion gene *Jingwei* is remodelled from *yellow emperor* and *alcohol dehydrogenase* genes combining activity on both long chain alcohols and diols, including growth hormones and pheromones, and establishing a novel developmental function in Drosophila (8). In addition, the *kua-UEV* fusion gene in human has facilitated cytoplasmic localization of an otherwise solely nuclear polyubiquitination co-effector (9). Whilst there is mounting evidence for the role of gene fusion in driving adaptive evolution (10–15), there are a number of outstanding questions about the evolution of composite genes in animals. Specifically, how prevalent composite gene formation has been, whether the emergence of composite genes occurs in bursts rather than gradually, whether the pattern of emergence of composite genes correlates with the origin of major animal groups, and at what rate fusions or fissions occur. In addition, whilst the convergent evolution of phenotypes is well established in animals, the extent to which composite genes can arise independently in different lineages is largely unknown. Given the divergence in morphologies, niches, lifestyles, and indeed genomes, it is not clear whether repeated evolution of the same molecular components would be precluded, or whether the deterministic effects, *i.e.* the benefits of particular kinds of composites, would overcome any contingent effects of prior genome evolution.

Animal genome evolution has been shaped by regulatory innovations (16), by *de novo* gene genesis (17–21), and by gene duplication and gene loss (20, 22). Composite genes have been particularly challenging to study as they simultaneously reside in more than one homologous gene cluster which complicates gene family assignment and phylogenetic analyses. Studies of protein coding aspects of animal evolution have necessarily relied upon strict definitions of gene families that limit our view to purely furcating processes (23, 24). An alternative view is provided by retaining the connections between composite genes and their components in sequence similarity networks (SSNs), permitting genes to be members of more than one family simultaneously. Taking advantage of the unique network motif typical of composite genes, *i.e.* they form "non-transitive triplets" thereby connecting gene families that are otherwise unconnected through partial sequence homology, we can identify candidate composite genes from genome scale data (25).

The protein domain space that comprises all animal proteins (components and composites alike) is limited, and it is conceivable that the same composite gene forming event could occur multiple times independently (26). Indeed, comparative empirical studies reveal surprisingly repeatable evolutionary fates in closely related lineages, a trend considered to reduce with increasing distance (27, 28). To date, the rate of independent evolution observed within multi-domain proteins has varied dramatically from very low, *i.e.* 0.4% (29), to much higher, 5-25% (13, 27). The lower range of estimates of independent evolution is thought to be caused by limited taxon sampling (13, 27). Using a large representative dataset, we provide a statistically sound framework to elucidate the rate of independent evolution of composite genes in animals.

Materials and methods

Dataset assembly

A dataset of 1,217,174 protein coding genes from a sample of 63 animal species representing all major clades within the animal tree was obtained from the OMA orthology database (70). Taxa were sampled to capture the known periods of major transition within animal evolution, and species representing all major nodes in the animal tree were included. The quality of data used was of particular importance in this study (given the potential for misidentification of composite genes) therefore taxon sampling was guided by the quality of gene annotation of the available species genomes using two filtering steps of genomes. First, we searched for protein coding genes known to be present across all of Metazoa (412 genes in total) (71), ranking genomes as high quality if they possessed >70% of the conserved set, while low quality genomes had less. Next, a smaller set of 40 protein coding genes that are annotated as being present across all of life (72) were used as queries to search for their presence in the set of animal genomes. As this set of protein coding genes is more conserved, this allowed for stricter filtering for quality of the genomes. All homology searching for the “core set of metazoans” and “all of life” protein coding genes was carried out using a reciprocal BLASTp approach (73). Searching for the set of conserved genes within sampled genomes in Metazoa and all of life, ensured that genomes of high quality (deemed by the presence of these sets of conserved genes) were used in our analysis.

To construct a time-calibrated species tree, node dates and topology were obtained from TimeTree (53), and contentious groupings (such as the branching order at the root of the animal lineage) were resolved based on current literature on the animal phylogeny (50–52, 74). We also included an alternative topology for the root position on the animal tree to test if our results are robust to the position of the deepest divergences. Twelve of the species in our dataset were missing from the TimeTree database, and so to place their position and time of divergence, closely related species to these lineages were used as replacements. In most cases, sister species from the same genus were present, and a list of the closely related species used to replace them can be found in (**Supplementary Table S2**). With other species, such as the case of *Ciona savignyi,* which was not present in TimeTree, the divergence time between it and its sister lineage *Ciona intestinalis* was taken as 176 MYA from the literature (75).

Generation and filtering of the sequence similarity network

An all versus all BLASTp (Altschul et al. 1990) was carried out (E-value <= 1e-5, percent identity >= 30%). The statements of homology output from BLASTp were used to generate a Sequence Similarity Network (SSN), using the cleanBlastp step in CompositeSearch (25). CompositeSearch applies a modified Depth First Search (DFS) algorithm to annotate gene families followed by subsequent network searching to define composite genes and gene families, then takes this SSN as input and identifies composite gene clusters which are denoted by non-transitive triplet patterns in the SSN. We used an E-value cutoff of 1e-5, percent identity cutoff of 30%, and coverage threshold of 80%. This provides output files on all HGs - both the gene families detected, and the gene families annotated as composite (CHGs). The composite gene family’s annotation file also provides information on the size of the composite gene families, the number of component families associated with the

composite family, the size of the component gene families and the connectivity of the subgraph of composite genes within a family. Information such as the number of composite genes within a family, and the amount of overlap of the homologous regions between the distinct component genes and the composite gene is all made readily available. As discussed previously the detection of composite genes may be prone to misidentification and false positives. Therefore, as an initial filtering step, a series of quality filters were applied to the putative composite gene families identified in the CompositeSearch analysis: firstly, singleton CHGs, *i.e.* those with only a single member in the CHG, were removed. In total this filtering step removed 48,640 CHGs out of the total of 77,085 putative CHGs. In addition, genes where the mapping of components to the composite was ambiguous or where the mapping of the components overlapped, were also removed (this removed a further 14,813), leaving a total of 13,632 remaining putative CHGs.

Finding evidence for expression of unique joining-points of composite genes using publicly available transcriptome data

We collated a dataset of all available transcriptomes from RNA sequencing studies for each taxon (52 out of 63 taxa had RNAseq data available). This allowed us to assess the putative composite genes at two levels: (i) their validity - making sure we do not report putative composite genes that are misassembly or misannotation artefacts, and (ii) whether they have evidence of expression. RNAseq reads were mapped to the unique joining-point region of the composite genes (*i.e.* the junction of component genes) using bowtie2 (76). RNAseq datasets were selected based on their robustness (as measured by the number of time points and tissues sampled) and the phylogenetic distribution of composites. For example, for widely distributed composite families, representative taxa from across the lineages containing the composite family were chosen based on the robustness of their available RNAseq datasets. The representative taxa for each of the Bilaterian clades included humans (Deuterostomia) and fruit-fly (Protostomia). Coverage across the composite joining-point was assessed using BEDTools (77). Evidence for transcription of the composite gene was determined by the coverage of at least one read across the joining-point.

Domain architecture analysis

For all of our HG datasets (composite HGs, component HGs, and non-composite associated HGs) we first annotated protein domains from the Pfam database using domain-specific hidden markov models (31), using pfam_scan.pl and parsing using PfamScanner with an e-value threshold of 1e-3. We calculated the proportion of retention and loss during gene fusion for each domain by dividing every time it is present or lost in a composite gene by the number of times it is seen in all component genes (this analysis was carried out on just the domain type architecture of the proteins rather than the full protein architecture which may include repeat domains). No statistically significant correlation is observed between the number of times a domain type is seen in a CHG vs the proportion of time it is present or lost (**Supplementary Figure S1**). Similarly, when we annotate the domain by function or size, there does not seem to be a correlation between the presence of a domain and these traits.

To assess whether there were any domains enriched in the set of composite genes which were annotated as emerging independently multiple times, we compared

domain sets between single event CHGs and convergently formed CHGs. We obtained domain lists and correlating functions for both sets of CHGs and applied the find_enrichment.py script from Goatools (78) setting the full set of domains from all CHGs as the background.

Assessment of rate of convergent evolution of CHGs

The pipeline for determining which CHGs emerged in a single event or multiple events is highlighted in **Figure 3**. In summary, we first retained all CHGs that had a simple 2-to-1 relationship between composite and component genes. Next, we extracted the homologous regions between both parts of the composite gene and their respective component sequences, using information from the tabular all-v-all BLAST output. The homologous regions between composite and components were aligned using MAFFT (79) and trimmed using trimAl, using the -gappyout parameter (80). Corresponding gene trees for both parts of the composite gene and respective component genes were constructed using IQTree (42, 43), applying ModelFinder (81) to find the model of best fit and carrying out 1000 ultrafast bootstrap replicates. We used clan-check to classify composite genes according to whether they appeared at face-value to have a single or multiple origins (82). Next, we constructed constrained gene trees in IQTree by forcing the composite sequences to be monophyletic and applying the model of best fit as inferred from the previous gene tree construction step. This ensures that gene tree construction is consistent between the two approaches, the only exception being that the composite genes are forced to be monophyletic. Finally, an AU test was carried out using IQTree, applying the -au parameter to compare support levels for the inferred gene tree with the constrained gene tree.

We also assessed the conservation of the joining-points between composite genes in each CHG tested. This involved determining the location of the joining-point for each composite gene by annotating where the sequence homology of the component genes mapped to the composite gene. Then, each composite gene in each CHG was split into four non overlapping but equal length regions (proportional to overall length of the composites) and we assessed whether the joining-points for all composite genes in a CHG fell within the same region. The assumption being that, while there may be some variation in the exact location of the joining-point, those in the same region of the composite gene provide more support for a single origin of a given CHG. This test was carried out on all CHGs. Leading on from this we could then address the question of whether we observe different joining-points in CHGs of multiple origin.

Mapping composite gene gain and loss onto species tree

For the taxa in our dataset, most of the branching patterns are well resolved allowing the analysis of the rate of emergence of CHGs on each branch across the tree to determine the patterns of gain and loss. We reconstructed the gain/loss history of the CHGs and used a constrained timetree (53) to determine their rates of gain and loss. The pattern of gains and losses of CHGs across the tree was assessed using one of two models; if CHGs were determined to have been formed in a single event, we used Claddis (46) an R package which operates in a maximum likelihood framework to describe characteristics of binary data. Specifically we used the map_dollo_changes.R function, which was developed to generate a stochastic character map for Dollo characters, allowing for a single gain event followed by any

number of losses (47). Alternatively, if CHGs were annotated as evolving convergently, we implemented the Mk model using a stochastic character mapping approach, this time implemented in RevBayes (49). Setting the root to zero and using a Mk model with unequal transition rates, allowing a character to be lost and gained a number of times at different rates across the tree, we measured the rate of gain and loss of each CHG individually. For each CHG we ran two mcmc chains for 5,000 generations each, allowing us to measure the precise timing of gain along each branch stochastically. Visualisation of the numbers of gains and losses on the species tree was carried out using an edited version of the R package, RevGadgets (83). To determine the rate of gain and loss of the CHGs mapped to the species phylogeny, we divided the number of CHGs gained or lost at a given node by the age of that node. The rates of gain and loss were added to each branch in the tree using ETE3 (84) and the tree was plotted using ggtree (85).

Characterising composite formation events

Composite genes may have originated from either fusion or fission. Fusion events merge two or more pre-existing components (*e.g.* gene families without direct connections in sequence similarity networks). Fission results in the subsequent appearance of split forms (*i.e.* components) of the gene. CompositeSearch (25) does not natively provide a classification of detected composites as corresponding to a fusion or fission event, yet this step is pivotal for a deeper biological interpretation of the computational outcome. We therefore applied a phylogeny-based method to infer the relationship of evolutionary precedence between composites and their respective components, and deduce the type of gene remodelling that was detected in the network (*i.e.* the evolution of composite genes by fusion vs the evolution of component genes by fission).

The last common ancestor of each CHG and the last common ancestor of each of its components were mapped onto the reference species tree of our sample set. These last common ancestors represent the putative points of appearance (assuming a unique origin) of each composite family and their associated component families. A simple heuristic was then applied to label CHGs as fused (originated from a gene fusion event) or split (underwent a gene fission event). CHGs for which components existed prior to the composite origin were considered as fused, as the inference of a component evolutionarily older than the composite indicates that at least one “building block” of the composite was present in its ancestral lineage before its appearance, and thus was unlikely the result of the composite fission. Conversely, CHGs for which components appeared only below the composite origin were marked as having undergone gene fission, as split forms of the composite persist in extant lineages ancestrally carrying the non-split gene. Many CHGs exhibited a particular pattern with both a component existing prior to the composite form and another component evolving only after the composite origin. Such cases are difficult to ascribe to a single fusion or fission event and may be the result of a more complex evolutionary path: the existence of a component predating the CHG indicates that it likely originated from a fusion event, possibly followed by a subsequent gene fission or loss that would have given rise to the later-evolving component.

Functional annotation of composite genes

All proteins used in our starting dataset were functionally annotated using eggNOG-mapper (86) v2.1.6, employing DIAMOND v2.0.11 to align sequences to the

eggNOG database v5.0.2. Cluster of Orthologous Groups (COG) functional categories were extracted for each sequence, and where required these were used to annotate the representative set of functions per CHG. To first test whether there were differences between the functional groups represented in fusion genes versus genes never associated with CHGs, we compared each COG category for each gene annotated as fusion or annotated as neither fusion or fission, and plotted the relative proportions (*i.e.* the number of COG categories divided by the number of genes) for each category. Next, to compare proportion of functional categories which emerged on each node of the tree, including major animal nodes, we took the proportions of categories from each CHG, and inferred the overall contribution at each node by dividing all the categories by the number of CHGs gained on the given node. These proportions for each node were plotted individually for the major nodes and combined to compare against all other internal nodes in the tree.

Results

Sequence similarity networks uncover a large number of composite genes

From a set of 1.2 million protein coding sequences across 63 animal genomes we identified 297,806 homologous groups (HGs) of which 77,085 contained putative composite homologous groups (CHGs). We removed (i) all singleton CHGs (of which there were 48,640), and (ii) all putative CHGs where the contributing component sequences did not map to specific and non-overlapping regions of the composite gene (14,813 CHGs in total). Under these strict criteria we identified 13,632 CHGs, or ~5% of all the gene families in animals, and these groups included 157,206 individual composite genes. To further mitigate against annotation and assembly artefacts, we assessed whether putative composites have associated evidence of gene expression. We mapped the unique “joining-point” in each composite gene to available transcriptome datasets (the “joining-point” is the location within the composite sequence where the contributing component sequences meet). Transcriptome data was available for 52 of the 63 species and 12,048 of the 13,632 CHGs (see Materials and Methods). A total of 7,774 CHGs (65%) had evidence of expression for at least one composite gene member of a given CHG family. The proportion with evidence of expression (*i.e.* 65%) is what we might expect from large scale RNAseq studies on temporal and spatial variation in expression in animal protein coding genes (30).

The 13,632 CHGs were related to a total of 40,217 component HGs, with the majority of CHGs (*i.e.* 10,855 (80%)) having just 2 component genes (or parts thereof). We identified a nested characteristic of composite formation in that CHGs once formed tend to contribute to further composite formation. In total, 11,805 out of 13,632 (87%) CHGs are also components (**Figure 1 a&b**). Indeed, when compared to genes that did not arise by composite formation, genes that arose by composite formation are more likely to subsequently contribute to other novel composite events (on average 10 vs 17 subsequent events respectively), suggesting there is a pool of genes prone to remodelling in animal genomes. On assessing protein length and domain content across the ~1.2 million protein coding genes in the network we observe that composite genes display a wider range of domain combinations (as classified by Pfam (31)) than either (i) component genes ($p < 2.2e-16$, Wilcoxon

signed-rank test), or (ii) non-composite associated genes (*i.e.* those genes that are not composite and not component) ($p < 2.2e-16$). Additionally, composite genes tend to have longer protein coding sequences than either component genes ($p < 2.2e-16$) or non-composite associated genes ($p < 2.2e-16$). On average component genes contribute 100 amino acids during composite formation - corresponding closely to the average length of a domain in all component genes in our dataset (*i.e.* 118 amino acids). The most common proportion of component gene sequence to be present following a fusion event is 20% or 100%, suggesting that the domain unit places the strongest constraint on the size and architecture of composite genes (**Figure 1c**). Comparing domain types between component and composite genes, we find that no domain is significantly over-represented across all CHGs and thus present at a higher rate than others during composite formation (27). However, whilst not statistically significant, domains WD40 and zf-C2H2 are present at a higher rate in CHGs than any other domain, *i.e.* WD40 is present in 391 components and is present in the resulting CHGs 58% of the time, and zf-C2H2 is present in 339 components and present in CHGs 58% of the time (**Supplementary Figure S1**). This trend possibly reflects the abundance and promiscuity of these two domains: the WD40 domain is one of the most abundant and also amongst the top interacting domains in eukaryotic genome (32), whilst the zf-C2H2 domain is amongst the most numerous of domains in metazoa (33).

For most CHGs (82% or 11,211 of 13,632 CHGs), the contributing elements (*i.e.* the parents of the gene fusion or the gene parent for the fission), are lost from the host genome. There are 2,421/13,632 CHGs where the composite and at least one component reside in the same genome simultaneously. For example, whilst previous studies of insulin-like growth factor-binding protein gene family (IGFBP) have characterised the functional domains, our analyses identify that the formation of this gene was via gene fusion on the stem chordate lineage (**Figure 2a**). We also show that following formation of this gene fusion its component genes were not lost from the genome. The process of gene fusion involved the C terminal IGFBP domain which functions to regulate IGF, and the N terminal Thyroglobulin-1 domain which contains nuclear localization sequences (**Figure 2b**). The IGFBP fusion and subsequent duplication resulted in many novel IGF-independent actions in a new cellular functional landscape (34–38). This example provides new insights into the process of gene fusion which involves retention of both component and composite genes in most chordates sampled (**Figure 2**). Conversely, the Nitrilase and fragile histidine triad fusion protein (NitFhit) demonstrates the loss of ancestral components following gene fusion. Given that the separate components have been found to be expressed and localised at similar time points and are also involved in similar interaction networks and functions, this fusion represents a coordination of biochemical pathways (39, 40). The NitFhit fusion was initially proposed to have originated by gene fusion in *C. elegans* and *D. melanogaster* (39), and we identify it in both Ecdysozoa and Lophotrochoza, placing the origin of the NitFhit fusion at the base of the Protostomia.

Larger number of composite genes are formed by gene fusion

As composite gene losses may be conflated with lineage-specific gene fission following a fusion event, we further categorised the mode of origin of all composite

genes in each CHG as having emerged by gene fusion or fission (see Materials and methods, **Supplementary Figure S2**). Briefly, if the last common ancestor of the component genes was an older node than that of the composite gene, the composite was categorised as formed by gene fusion; and the converse for fissions. Out of the 13,632 CHGs in this analysis, 9,994 CHGs (73.3%) were found to have originated from a fusion event. Of these, 2,096 (21%) fit the scenario of a single fusion event with no subsequent fission, while 7,898 (79%) were inferred to have undergone a more complex pattern of independent fusion and/or subsequent fission events (**Supplementary Figure S2B**). Interestingly, there were 3,460 CHGs (25.4%) that underwent a single, unique fission event (**Supplementary Figure S2C)**. Finally, 178 CHGs (1.3%) could not be assigned to any of the above categories, suggesting a more complex evolutionary history. Overall, the relative rates of fusion and fission observed mirror recent findings which suggest that gene fusion played a greater role in metazoan gene family evolution as compared to other eukaryotes, *i.e.* Fungi (21, 41) (**Supplementary Figure S2)**. These findings also suggest that gene fission occurs at a previously underestimated rate in animal genomes, particularly following gene fusion events.

Composite genes evolve multiple times independently

We quantified the rates at which evolution converged on the same composite gene using a phylogenetic approach, thus allowing us to overcome the bias related to gene loss. Using the phylogenetic signal within the homologous regions of the component and composite gene alignments, we determined the rate of independent origins of composite genes (**Figure 3**). We analysed component-composite gene trees to distinguish composite sequences that form monophyletic groups (which were most likely formed by a single event), from polyphyletic composite sequences (which represent possible multiple independent origins of that composite family, *e.g.*, multiple independent fusions/fissions). From among the 13,632 CHGs we selected families that met two criteria: whether they involved only two component families, and whether they contained more than three species in the alignment. In total, 10,829 of 13,632 CHGs satisfied these criteria.

Briefly, for each CHG we first aligned and trimmed the sequences and then built gene trees using IQTree (42, 43) (v2.03; using automatic model selection and carrying out 1000 ultrafast bootstrap replicates) for all homologous blocks of sequences of component and composite genes (**Figure 3a-b**). We then selected those maximum likelihood trees where the composite genes do not appear as a monophyletic group (9,124/10,829 or 84%). We re-ran the analysis using the same phylogenetic models as before, but this time we imposed topological constraints on the search for optimal trees, where we forced the composite sequences to form a monophyletic group (**Figure 3c**). The Approximately Unbiased (AU) test (44) was used to measure the significance in the difference in support for the unconstrained (polyphyletic) versus the constraint (monophyletic) tree (**Figure 3d**). The null hypothesis is that there is no significant difference in likelihood score for the constraint tree and the unconstrained tree. This approach provides a robust statistical framework to infer the rate of independent evolution of composite genes. Out of the 9,124 CHGs of putative independent origin, 5,631 rejected the null hypothesis, *i.e.* the monophyletic constraint is significantly worse than the

polyphyletic gene tree implying independent origin of the same composite event. For the remaining 3,493 CHGs tested for single or multiple origins, there was no significant difference between the constraint and gene trees, *i.e.* a monophyletic origin could not be discounted and in these cases we parsimoniously assumed monophyletic origin. Whilst different joining-points are possible in both monophyletic and polyphyletic CHGs, there should be less constraint on identical joining-points in polyphyletic cases. To test this we annotated the joining-point in all composite genes for each CHG and asked whether this location fell within the same general region of the protein for each gene in a CHG (see Materials and methods). Indeed, an analysis of the joining-points of all CHGs shows that 70% of polyphyletic CHGs have different joining-points as compared to 44% for monophyletic CHGs. To summarise, across all 13,632 CHGs identified, 5,198 CHGs (38%) were most likely monophyletic, 5,631 CHGs (41%) emerge independently more than once across the animal phylogeny, and 2,803 CHGs (21%) could not be assessed in this way as a consequence of having more than two components and/or insufficient species in the alignment.

Next, we assessed whether there was a significant difference in the types of protein domains present in composite genes of multiple origin as compared to those of single origin. We found that domains with functions related to protein binding and binding (GO:0005515, GO:0005488) were enriched in the set of composite genes which emerged more than once ($p<0.05$, Benjamini-Hochberg correction) (**Supplementary Table S1**).

Composite gene gain and loss events are concentrated at specific nodes on the phylogeny

In order to determine the tempo of CHG gain and loss across the animal phylogeny we analysed a subset of 10,829 (from a total of 13,632) CHGs where we could categorise the CHG as single or multiple origin. For the 5,198 CHGs of single origin we analysed their evolutionary history using the irreversible Dollo model of evolution (45) implemented in the R package Claddis (46, 47). For the 5,631 CHGs of multiple origin we used the reversible Mk model (48) implemented in the RevBayes software program (49). We used the species tree shown in **Figure 4** (50–52), the most appropriate model as described above, and applied time calibrations extracted from the TimeTree database (53). Using TimeTree as a source for divergence times allowed us to use a detailed phylogeny including all taxa of interest. However, TimeTree divergence times are summary statistics from a diversity of studies some of which are by now only of historical value. In particular, for the deep part of the animal phylogeny, TimeTree estimates are most likely too old (54). Accordingly, rates of origin and loss of composite genes estimated here should be interpreted as minimal estimated values, as reducing the branch length in the timetree following more recent animal divergence time studies will cause the inference of higher rates of origin and losses. We also employed an alternative rooting for the tree placing the Ctenophore as the earliest diverging group (55–57) and found no significant change to the results presented here (**Supplementary Figure S3**).

The first major observation is that CHG gain is not evenly distributed across the tree (**Figure 4**). Within Deuterostomia for example there are 21,381 separate CHG gain events across all branches compared to 6,295 CHGs gains in all branches within

Protostomia. There are 1,322 gains in total across the five nodes preceding the divergence of Deuterostomia and Protostomia (*i.e.* Bilateria, Panulozoa, Eumetazoa, Metazoa, and Metazoa + Choanoflagellates). While there is a clear disparity between the number of composite genes between protostome and deuterostome species, this trend does not hold at the level of total gene count per species (**Supplementary Figure S4A&B**) suggesting underlying variation in the rate of CHG formation between these clades. To account for the difference in the number of species sampled between Protostomia and Deuterostomia, CHG counts for a random sample of 10 species were carried out 100 times, and we find that the number of CHGs present in the Deuterostome clade is significantly higher (Wilcoxon rank-sum test, p=0) (**Supplementary Figure S4C**). Finally, the distribution of CHGs across the tree suggests that a large proportion may be clade-specific. For example, there are 341, 338, and 1,475 CHGs unique to Caenorhabditis, Eutheria and Euteleostomi, respectively. The ancestral node with the largest number of CHGs gained was Eutelostomi with 1,475 gains, followed by Bilateria with 527 gains (**Figure 4**).

Across the tree we find that the rates of composite gene gain and loss per million years (MY) are highly variable and non-clock like (**Figure 4**). Note that the absolute values presented for rates are likely to be underestimated given the tree, but the comparisons of gains and losses remain valid. Within the internal branches of the phylogeny the average rate of CHG gain is 1.03/MY, compared to a loss rate of 2.81/MY. The branch leading to the Hominoidea clade displays the highest rate of CHG gain (9.31 CHG gains/MY). The branches leading to Caenorhabditis (5.67 gains/MY), Euteleostomi (3.39 gains/MY), Poecilinae (4.50 gains/MY), Eutheria (3.20 gains/MY), and Tetraodontidae (2.56 gains/MY) all display rates of CHG gain above the average plus the SD observed across the whole phylogeny (1.03 ± 1.51 CHG gains/MY) (**Figure 4 inset, Supplementary Figure S5A**). Contrastingly, higher rates of CHG loss tend to be found towards the tips of the tree, consistent with the observation of loss most often relating to loss of a single composite gene within a specific species rather than complete loss of the CHG (**Figure 4**). Branches with rates of loss higher than the average plus SD (2.81 ± 4.33 losses/MY) include: Hominoidea (24.96 losses/MY), Xenarthra (12.27 losses/MY), and Passeriformes (10.53 losses/MY), and the tip lineages: *Nomascus leucogenys* (23.08 losses/MY), *Choloepus hoffmanni* (21.98 losses/MY) (**Supplementary Figure S5B**). When assessing rates of gain and loss of composite genes, it must be noted that loss can relate to a complete loss of a composite gene from a species, or a reversal of the composite formation event (*e.g.* subsequent fission in a lineage following a fusion event). Gain of an individual composite gene member also occurs but at a lower rate than CHG member loss.

Finally, to assess the functional contribution of gene fusions specifically from our composite gene repertoire we compared the functional categories of fusion genes versus genes never associated with composite formation (*i.e.* neither composite nor components). We calculated the relative representation of Cluster of Orthologous Groups (COG) categories for genes of each type (*i.e.* fusion genes versus non-composite associated genes). We found that, in comparison to non-composite associated genes, fusion genes had a larger number of genes involved in transcription (K), post-translational modification, protein turnover, and chaperone functions (O), inorganic ion transport and metabolism (P), signal transduction (T), and extracellular structures (W) (**Supplementary Figure S6A**). To assess the potential functional impact of gene fusion events at major nodes in the animal tree,

we measured the relative proportions of COG categories for fusion genes gained at these particular nodes (**Supplementary Figure S6B**). The overall proportions of COG categories are similar for all nodes tested, with signal transduction (T), transcription (K), and post-translational modification, protein turnover, and chaperone functions (O) representing the highest proportion of COG categories in all nodes. These patterns overlap with the broader contribution of fusion genes to these functions relative to non-composite genes, as seen above (**Supplementary Figure S6A**). Some clade specific shifts in COG proportions were observed; with overall larger proportion of genes involved in RNA processing and modification (A) present in the nodes Deuterostomia, Tetrapoda, Mammalia, and Eutheria; a larger contribution of extracellular structures (W) category in Euteleostomi and Tetrapoda; and a higher proportion of genes involved in defence mechanisms (V) in Tetrapoda, relative to other major nodes in the tree (**Supplementary Figure S6B**). Generally, the functional impact of gene fusion seems to be specific to certain broad functional categories throughout the animal phylogeny, pointing to a specific, persistent, role of gene fusion in driving the evolution of certain functions important for animal evolution (58).

Discussion

The most recognisable part of evolutionary biology is the Tree of Life with its continually diverging branches emerging from the root. This narrative has hugely influenced how we think about evolutionary history, and it influences what we expect to see when we examine genomes. In addition to, and perhaps influenced by, the Tree of Life perspective, there is a feeling that evolution rarely, if ever, repeats itself. This last idea was most forcefully expressed by the palaeontologist Stephen Jay Gould who asked whether the tape of life was replayable (59) – a question to which Gould answered: No.

Fortunately, with the sequencing of an extensive array of genomes from many taxa across the diversity of animals we can address issues relating to the non-treelike aspects of evolution on one hand, as well as whether genome evolution is contingent on genetic background. If evolution is contingent on prior evolutionary events, then we expect that with increasingly divergent genetic backgrounds we are less likely to see repeated evolution, while on the other hand if evolution is largely deterministic, then despite differences in genetic backgrounds we expect to see the same evolutionary events occurring in different lineages. In the end, we find that the repeatability of evolution falls short of being hugely deterministic – only a small proportion of the overall animal gene repertoire shows evidence of repeated evolution, but by the same token, we see that repeatability has happened and it has happened across animal evolution many, many times.

Composite gene evolution is largely characterised by high rates of gene turnover, unequal rates of gain and loss between animal phyla, and significant bursts of composite gains intensified at particular nodes in the animal phylogeny. The rate of composite gene formation varies drastically between the major animal groups, with a greater number of composite genes found in Deuterostomia, compared to Protostomia or the non-bilaterian lineages. The single largest number of CHG gain events observed are on the branch leading to the Euteleostomi ancestor - a branch

synonymous with major phenotypic innovations such as the emergence of mineralised bone and the development of a more complex immune system. The higher rates of CHG birth on this branch may also be related to increases in genome complexity at this point in animal evolution (60, 61). Multiple whole genome duplication (WGD) events within the vertebrate clade (62), coincide with nodes containing a large number of composite genes. In terms of rate of composite gain per million years, the branch leading to Euteleostomi also shows a significantly higher rate of gain than the average across the whole tree, reinforcing the contribution of gene fusion and fission at this point in animal evolution. The branch leading to the Caenorhabditis species also shows significantly higher rates of CHG gain. While this branch may not represent a point of major phenotypic or genomic change, species of this phylum are known to have a higher rate of recombination within their genomes (63). This may provide a molecular basis for the increased rates of gene fusion and fission in this group. Compared to patterns of gain and loss of other gene types, which show significant gain in early metazoan branches and pre-metazoan branches in particular (17, 18, 20), we find that the highest rates of composite gene gain correlate with nodes that emerge subsequent to these deep nodes, suggesting that the evolution of genetic content through mechanisms other than fusion and fission may be followed by subsequent higher rates of composite gene formation.

In animals, convergent fusion events are known, for example the TRIM5-CYPA gene in New World monkeys (64) and Old World monkeys (65), the repeated fusion of β-globin genes in Laurasiatheria (66), and the recurrent fusion of transcription factors and transposons in vertebrates (67). More broadly speaking, 25% of all multi-domain proteins in eukaryotes are thought to have emerged independently, and 71% of domain combinations in the human genome have been found to be gained independently in at least one other eukaryotic genome (27). There have also been several examples of recurrent gene fusion in different eukaryote lineages (68, 69). Our estimate for the rate of convergent evolution of composite genes in the evolution of animals suggests that selection for the same combinations of gene sequences in composite genes is indeed common, with 41% (5,631 CHGs) having likely evolved independently more than once on the tree. Given our approach, using phylogenetic signal within the composite and component sequences, we could be confident that our results are not skewed by taxon sampling or data quality issues. The data presented here suggests that there are high levels of CHG formation and loss across animal evolution, that the same composites form independently across the tree, and that these CHG likely contribute substantially to the emergence of animal gene repertoires providing functional innovation, *e.g.* the IGFBP fusion protein. This work has important implications for our understanding of how protein coding genes evolve in animals, the prevalence of convergent evolution, how we construct gene families, and how we annotate function between homologous genes.

References

1. M. Long, E. Betrán, K. Thornton, W. Wang, The origin of new genes: glimpses from the young and old. *Nat. Rev. Genet.* **4**, 865–875 (2003).

2. D. Ekman, A. K. Björklund, A. Elofsson, Quantification of the elevated rate of domain rearrangements in metazoa. *J. Mol. Biol.* **372**, 1337–1348 (2007).

3. A. D. Moore, A. K. Björklund, D. Ekman, E. Bornberg-Bauer, A. Elofsson, Arrangements in the modular evolution of proteins. *Trends Biochem. Sci.* **33**, 444–451 (2008).

4. R. L. Rogers, D. L. Hartl, Chimeric genes as a source of rapid evolution in Drosophila melanogaster. *Mol. Biol. Evol.* **29**, 517–529 (2012).

5. E. Bornberg-Bauer, M. M. Albà, Dynamics and adaptive benefits of modular protein evolution. *Curr. Opin. Struct. Biol.* **23**, 459–466 (2013).

6. A. M. McCartney, *et al.*, Gene Fusions Derived by Transcriptional Readthrough are Driven by Segmental Duplication in Human. *Genome Biol. Evol.* **11**, 2678–2690 (2019).

7. N. B. Stewart, R. L. Rogers, Chromosomal rearrangements as a source of new gene formation in Drosophila yakuba. *PLoS Genet.* **15**, e1008314 (2019).

8. M. Long, C. H. Langley, Natural selection and the origin of jingwei, a chimeric processed functional gene in Drosophila. *Science* **260**, 91–95 (1993).

9. T. M. Thomson, *et al.*, Fusion of the human gene for the polyubiquitination coeffector UEV1 with Kua, a newly identified gene. *Genome Res.* **10**, 1743–1756 (2000).

10. L. Patthy, Evolution of the proteases of blood coagulation and fibrinolysis by assembly from modules. *Cell* **41**, 657–663 (1985).

11. L. Patthy, Exon shuffling and other ways of module exchange. *Matrix Biol.* **15**, 301–10; discussion 311–2 (1996).

12. C. Vogel, S. A. Teichmann, C. Chothia, The immunoglobulin superfamily in Drosophila melanogaster and Caenorhabditis elegans and the evolution of complexity. *Development* **130**, 6317–6328 (2003).

13. T. Kawashima, *et al.*, Domain shuffling and the evolution of vertebrates. *Genome Res.* **19**, 1393–1403 (2009).

14. G. W. C. Thomas, *et al.*, Gene content evolution in the arthropods. *Genome Biol.* **21**, 15 (2020).

15. E. Dohmen, S. Klasberg, E. Bornberg-Bauer, S. Perrey, C. Kemena, The modular nature of protein evolution: domain rearrangement rates across eukaryotic life. *BMC Evol. Biol.* **20**, 30 (2020).

16. A. M. Heimberg, L. F. Sempere, V. N. Moy, P. C. J. Donoghue, K. J. Peterson, MicroRNAs and the advent of vertebrate morphological complexity. *Proc. Natl. Acad. Sci. U. S. A.* **105**, 2946–2950 (2008).

17. X. Grau-Bové, *et al.*, Dynamics of genomic innovation in the unicellular ancestry of animals. *Elife* **6** (2017).

18. J. Paps, P. W. H. Holland, Reconstruction of the ancestral metazoan genome reveals

an increase in genomic novelty. *Nat. Commun.* **9**, 1730 (2018).

19. D. J. Richter, P. Fozouni, M. B. Eisen, N. King, Gene family innovation, conservation and loss on the animal stem lineage. *Elife* **7** (2018).

20. R. Fernández, T. Gabaldón, Gene gain and loss across the metazoan tree of life. *Nat Ecol Evol* (2020) https:/doi.org/10.1038/s41559-019-1069-x.

21. E. Ocaña-Pallarès, *et al.*, Divergent genomic trajectories predate the origin of animals and fungi. *Nature* **609**, 747–753 (2022).

22. C. Guijarro-Clarke, P. W. H. Holland, J. Paps, Widespread patterns of gene loss in the evolution of the animal kingdom. *Nat Ecol Evol* (2020) https:/doi.org/10.1038/s41559-020-1129-2.

23. L. S. Haggerty, *et al.*, A pluralistic account of homology: adapting the models to the data. *Mol. Biol. Evol.* **31**, 501–516 (2014).

24. T. H. Oakley, Furcation and fusion: The phylogenetics of evolutionary novelty. *Dev. Biol.* **431**, 69–76 (2017).

25. J. S. Pathmanathan, P. Lopez, F.-J. Lapointe, E. Bapteste, CompositeSearch: A Generalized Network Approach for Composite Gene Families Detection. *Mol. Biol. Evol.* **35**, 252–255 (2018).

26. E. Bolotin, D. Melamed, A. Livnat, Genes that are Used Together are More Likely to be Fused Together in Evolution by Mutational Mechanisms: A Bioinformatic Test of the Used-Fused Hypothesis. *Evol. Biol.* **50**, 30–55 (2023).

27. C. M. Zmasek, A. Godzik, This Déjà vu feeling--analysis of multidomain protein evolution in eukaryotic genomes. *PLoS Comput. Biol.* **8**, e1002701 (2012).

28. Z. D. Blount, R. E. Lenski, J. B. Losos, Contingency and determinism in evolution: Replaying life's tape. *Science* **362** (2018).

29. J. Gough, Convergent evolution of domain architectures (is rare). *Bioinformatics* **21**, 1464–1471 (2005).

30. L. Fagerberg, *et al.*, Analysis of the Human Tissue-specific Expression by Genome-wide Integration of Transcriptomics and Antibody-based Proteomics*. *Mol. Cell. Proteomics* **13**, 397–406 (2014).

31. S. El-Gebali, *et al.*, The Pfam protein families database in 2019. *Nucleic Acids Res.* **47**, D427–D432 (2019).

32. C. Xu, J. Min, Structure and function of WD40 domain proteins. *Protein Cell* **2**, 202–214 (2011).

33. H. S. Najafabadi, *et al.*, Non-base-contacting residues enable kaleidoscopic evolution of metazoan C2H2 zinc finger DNA binding. *Genome Biol.* **18**, 167 (2017).

34. L. J. Schedlich, *et al.*, Nuclear Import of Insulin-like Growth Factor-binding Protein-3 and -5 Is Mediated by the Importin β Subunit *. *J. Biol. Chem.* **275**, 23462–23470 (2000).

35. C. Iosef, T. Gkourasas, C. Y. H. Jia, S. S.-C. Li, V. K. M. Han, A functional nuclear localization signal in insulin-like growth factor binding protein-6 mediates its nuclear import. *Endocrinology* **149**, 1214–1226 (2008).

36. J. Zhou, J. Xiang, S. Zhang, C. Duan, Structural and functional analysis of the amphioxus IGFBP gene uncovers ancient origin of IGF-independent functions. *Endocrinology* **154**, 3753–3763 (2013).

37. D. O. Daza, G. Sundström, C. A. Bergqvist, C. Duan, D. Larhammar, Evolution of the insulin-like growth factor binding protein (IGFBP) family. *Endocrinology* **152**, 2278–2289 (2011).

38. D. J. Macqueen, D. Garcia de la Serrana, I. A. Johnston, Evolution of ancient functions in the vertebrate insulin-like growth factor system uncovered by study of duplicated salmonid fish genomes. *Mol. Biol. Evol.* **30**, 1060–1076 (2013).

39. Y. Pekarsky, *et al.*, Nitrilase and Fhit homologs are encoded as fusion proteins in *Drosophila melanogaster* and *Caenorhabditis elegans*. *Proc. Natl. Acad. Sci. U. S. A.* **95**, 8744–8749 (1998).

40. H. C. Pace, C. Brenner, The nitrilase superfamily: classification, structure and function. *Genome Biol.* **2**, REVIEWS0001 (2001).

41. G. Leonard, T. A. Richards, Genome-scale comparative analysis of gene fusions, gene fissions, and the fungal tree of life. *Proc. Natl. Acad. Sci. U. S. A.* **109**, 21402–21407 (2012).

42. L.-T. Nguyen, H. A. Schmidt, A. von Haeseler, B. Q. Minh, IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* **32**, 268–274 (2015).

43. B. Q. Minh, *et al.*, IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Mol. Biol. Evol.* **37**, 1530–1534 (2020).

44. H. Shimodaira, An approximately unbiased test of phylogenetic tree selection. *Syst. Biol.* **51**, 492–508 (2002).

45. A. V. Alekseyenko, C. J. Lee, M. A. Suchard, Wagner and Dollo: a stochastic duet by composing two parsimonious solos. *Syst. Biol.* **57**, 772–784 (2008).

46. G. T. Lloyd, Estimating morphological diversity and tempo with discrete character-taxon matrices: implementation, challenges, progress, and future directions. *Biol. J. Linn. Soc. Lond.* **118**, 131–151 (2016).

47. J. E. Tarver, *et al.*, Well-Annotated microRNAomes Do Not Evidence Pervasive miRNA Loss. *Genome Biol. Evol.* **10**, 1457–1470 (2018).

48. P. O. Lewis, A likelihood approach to estimating phylogeny from discrete morphological character data. *Syst. Biol.* **50**, 913–925 (2001).

49. S. Höhna, *et al.*, RevBayes: Bayesian Phylogenetic Inference Using Graphical Models and an Interactive Model-Specification Language. *Syst. Biol.* **65**, 726–736 (2016).

50. D. Pisani, *et al.*, Genomic data do not support comb jellies as the sister group to all other animals. *Proc. Natl. Acad. Sci. U. S. A.* **112**, 15402–15407 (2015).

51. P. Simion, *et al.*, A Large and Consistent Phylogenomic Dataset Supports Sponges as the Sister Group to All Other Animals. *Curr. Biol.* **27**, 958–967 (2017).

52. R. Feuda, *et al.*, Improved Modeling of Compositional Heterogeneity Supports Sponges as Sister to All Other Animals. *Curr. Biol.* **27**, 3864–3870.e4 (2017).

53. S. Kumar, G. Stecher, M. Suleski, S. B. Hedges, TimeTree: A Resource for Timelines, Timetrees, and Divergence Times. *Mol. Biol. Evol.* **34**, 1812–1819 (2017).

54. M. dos Reis, *et al.*, Uncertainty in the Timing of Origin of Animals and the Limits of Precision in Molecular Timescales. *Curr. Biol.* **25**, 2939–2950 (2015).

55. N. V. Whelan, *et al.*, Ctenophore relationships and their placement as the sister group to all other animals. *Nat Ecol Evol* **1**, 1737–1746 (2017).

56. Y. Li, X.-X. Shen, B. Evans, C. W. Dunn, A. Rokas, Rooting the Animal Tree of Life. *Mol. Biol. Evol.* **38**, 4322–4333 (2021).

57. D. T. Schultz, *et al.*, Ancient gene linkages support ctenophores as sister to other animals. *Nature* **618**, 110–117 (2023).

58. H. Suga, *et al.*, The Capsaspora genome reveals a complex unicellular prehistory of animals. *Nat. Commun.* **4**, 2325 (2013).

59. S. J. Gould, *Wonderful life: the Burgess Shale and the nature of history* (WW Norton & Company, 1989).

60. C. Sacerdot, A. Louis, C. Bon, C. Berthelot, H. Roest Crollius, Chromosome evolution at the origin of the ancestral vertebrate genome. *Genome Biol.* **19**, 166 (2018).

61. O. Simakov, *et al.*, Deeply conserved synteny resolves early events in vertebrate evolution. *Nat Ecol Evol* (2020) https:/doi.org/10.1038/s41559-020-1156-z.

62. L. Z. Holland, D. Ocampo Daza, A new look at an old question: when did the second whole genome duplication occur in vertebrate evolution? *Genome Biol.* **19**, 209 (2018).

63. J. Stapley, P. G. D. Feulner, S. E. Johnston, A. W. Santure, C. M. Smadja, Variation in recombination frequency and distribution across eukaryotes: patterns and processes. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **372** (2017).

64. D. M. Sayah, E. Sokolskaja, L. Berthoux, J. Luban, Cyclophilin A retrotransposition into TRIM5 explains owl monkey resistance to HIV-1. *Nature* **430**, 569–573 (2004).

65. G. Brennan, Y. Kozyrev, S.-L. Hu, TRIMCyp expression in Old World primates Macaca nemestrina and Macaca fascicularis. *Proc. Natl. Acad. Sci. U. S. A.* **105**, 3569–3574 (2008).

66. M. J. Gaudry, J. F. Storz, G. T. Butts, K. L. Campbell, F. G. Hoffmann, Repeated evolution of chimeric fusion genes in the β-globin gene family of laurasiatherian mammals. *Genome Biol. Evol.* **6**, 1219–1234 (2014).

67. R. L. Cosby, *et al.*, Recurrent evolution of vertebrate transcription factors by transposase capture. *Science* **371** (2021).

68. N. A. Stover, A. R. O. Cavalcanti, A. J. Li, B. C. Richardson, L. F. Landweber, Reciprocal fusions of two genes in the formaldehyde detoxification pathway in ciliates and diatoms. *Mol. Biol. Evol.* **22**, 1539–1542 (2005).

69. F. Maguire, *et al.*, Complex patterns of gene fission in the eukaryotic folate biosynthesis pathway. *Genome Biol. Evol.* **6**, 2709–2720 (2014).

70. A. M. Altenhoff, *et al.*, The OMA orthology database in 2015: function predictions, better plant support, synteny view and other improvements. *Nucleic Acids Res.* **43**, D240–9

(2015).

71. S. Powell, *et al.*, eggNOG v3.0: orthologous groups covering 1133 organisms at 41 different taxonomic ranges. *Nucleic Acids Res.* **40**, D284–9 (2012).

72. F. D. Ciccarelli, *et al.*, Toward automatic reconstruction of a highly resolved tree of life. *Science* **311**, 1283–1287 (2006).

73. S. F. Altschul, W. Gish, W. Miller, E. W. Myers, D. J. Lipman, Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).

74. A. K. Redmond, A. McLysaght, Evidence for sponges as sister to all other animals from partitioned phylogenomics with mixture models and recoding. *Nat. Commun.* **12**, 1783 (2021).

75. F. Delsuc, *et al.*, A phylogenomic framework and timescale for comparative studies of tunicates. *BMC Biol.* **16**, 39 (2018).

76. B. Langmead, S. L. Salzberg, Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).

77. A. R. Quinlan, I. M. Hall, BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).

78. D. V. Klopfenstein, *et al.*, GOATOOLS: A Python library for Gene Ontology analyses. *Sci. Rep.* **8**, 10872 (2018).

79. K. Katoh, K.-I. Kuma, H. Toh, T. Miyata, MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucleic Acids Res.* **33**, 511–518 (2005).

80. S. Capella-Gutiérrez, J. M. Silla-Martínez, T. Gabaldón, trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**, 1972–1973 (2009).

81. S. Kalyaanamoorthy, B. Q. Minh, T. K. F. Wong, A. von Haeseler, L. S. Jermiin, ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat. Methods* **14**, 587–589 (2017).

82. K. Siu-Ting, *et al.*, Inadvertent Paralog Inclusion Drives Artifactual Topologies and Timetree Estimates in Phylogenomics. *Mol. Biol. Evol.* **36**, 1344–1356 (2019).

83. C. M. Tribble, *et al.*, RevGadgets: An R package for visualizing Bayesian phylogenetic analyses from RevBayes. *Methods Ecol. Evol.* (2021) https:/doi.org/10.1111/2041-210x.13750.

84. J. Huerta-Cepas, F. Serra, P. Bork, ETE 3: Reconstruction, Analysis, and Visualization of Phylogenomic Data. *Mol. Biol. Evol.* **33**, 1635–1638 (2016).

85. G. Yu, D. K. Smith, H. Zhu, Y. Guan, T. T.-Y. Lam, ggtree: an r package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods Ecol. Evol.* **8**, 28–36 (2017).

86. J. Huerta-Cepas, *et al.*, Fast Genome-Wide Functional Annotation through Orthology Assignment by eggNOG-Mapper. *Mol. Biol. Evol.* **34**, 2115–2122 (2017).

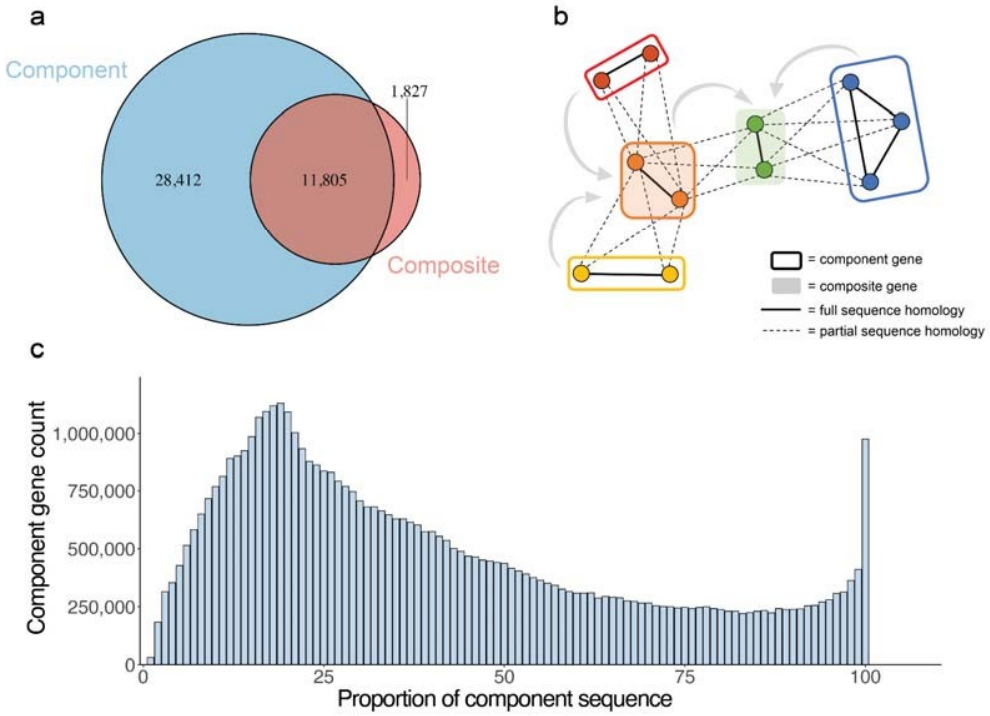


Figure 1. Characterisation and contributions of composite genes and their components. (a) The number of component CHGs (blue) and composite CHGs (red). Overlap represents component CHGs which are themselves composite. **(b)** Cartoon network demonstrating the nested nature of composite formation, whereby *e.g.* a CHG (orange) formed from distinct component CHGs in red and yellow, may itself be involved in a separate fusion with another HG component family (in blue) to form a new CHG (green). **(c)** Distribution of the component sequences showing the proportion of the component that is present in the composite.

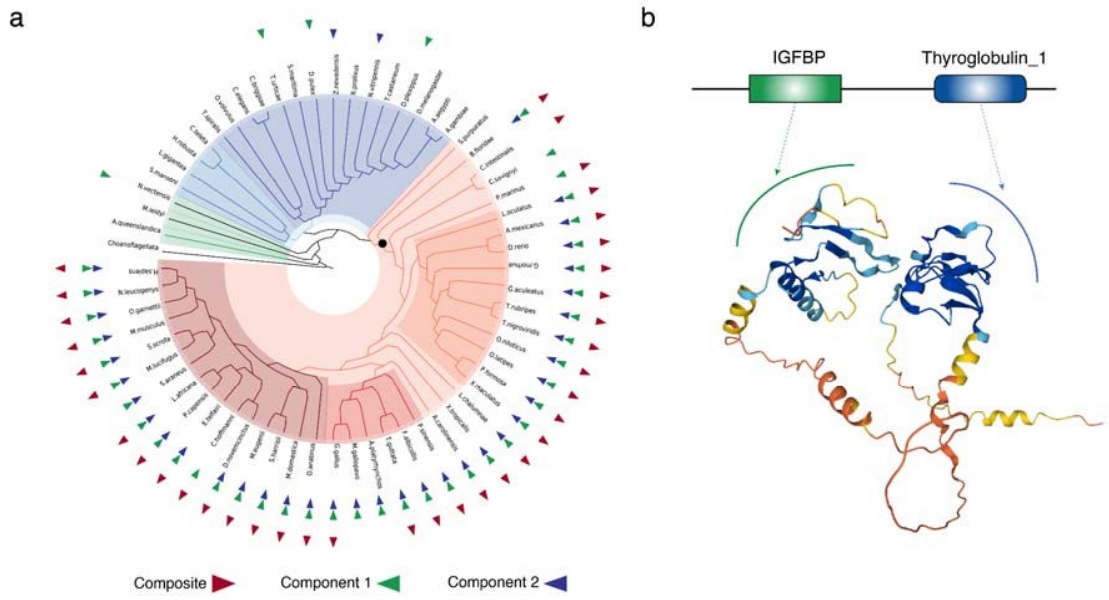


Figure 2. Evolution of the composite IGFBP gene in Chordates. (a) The chordate clade is highlighted with a red toned background on the circular tree, and the outgroups in blue and green toned backgrounds. The presence/absence of the composite and components are denoted with coloured triangles on the leaf nodes: the presence of the IGFBP composite gene is denoted as a red triangle, and the two component genes IGFBP and Thyroglobulin 1 are denoted as blue and green triangles respectively. The node of origin of the IGFBP composite genes is annotated by an in-filled circle on the species tree. **(b)** A cartoon of the constituent domains IGFBP (green) and Thyroglobin_1 (blue). Arrows point to the corresponding regions in the IGFBP composite gene protein structure. The structure colour gradient represents regions of high (blue) and low (red) confidence, note the two component protein domains are linked by a sequence of low structural confidence.

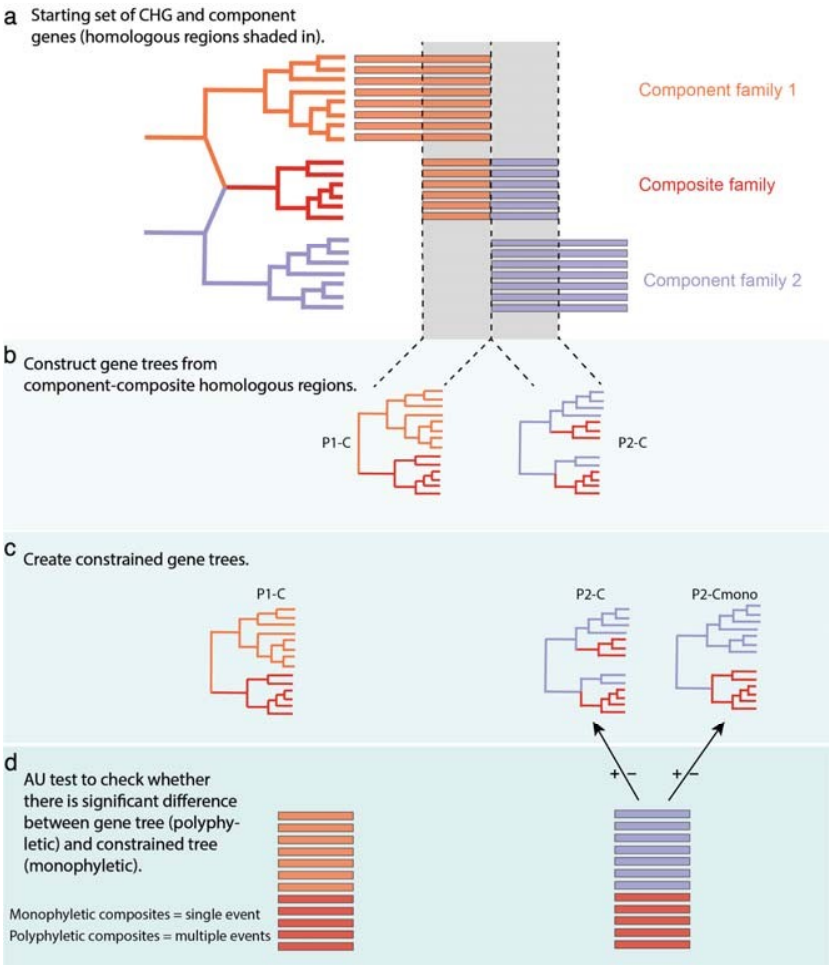


Figure 3. Pipeline used to identify composite genes that emerged in a single event or multiple independent events. (a) Summary of component and composite homologous sequences used for gene tree inference. **(b)** Gene trees inferred from component-composite sequences showing an example where the composite sequences (in red) form a monophyletic group (left) and an example where they do not (right). **(c)** Constrained trees inferred using the topology of the gene tree inferred from the previous step but forcing the composite sequences to be monophyletic. **(d)** Approximately unbiased test: measuring the significance in the difference in support for the constraint tree (monophyletic) versus the unconstrained (polyphyletic) tree. Where H_0 = no significant difference in likelihood score for the constraint and unconstrained trees.

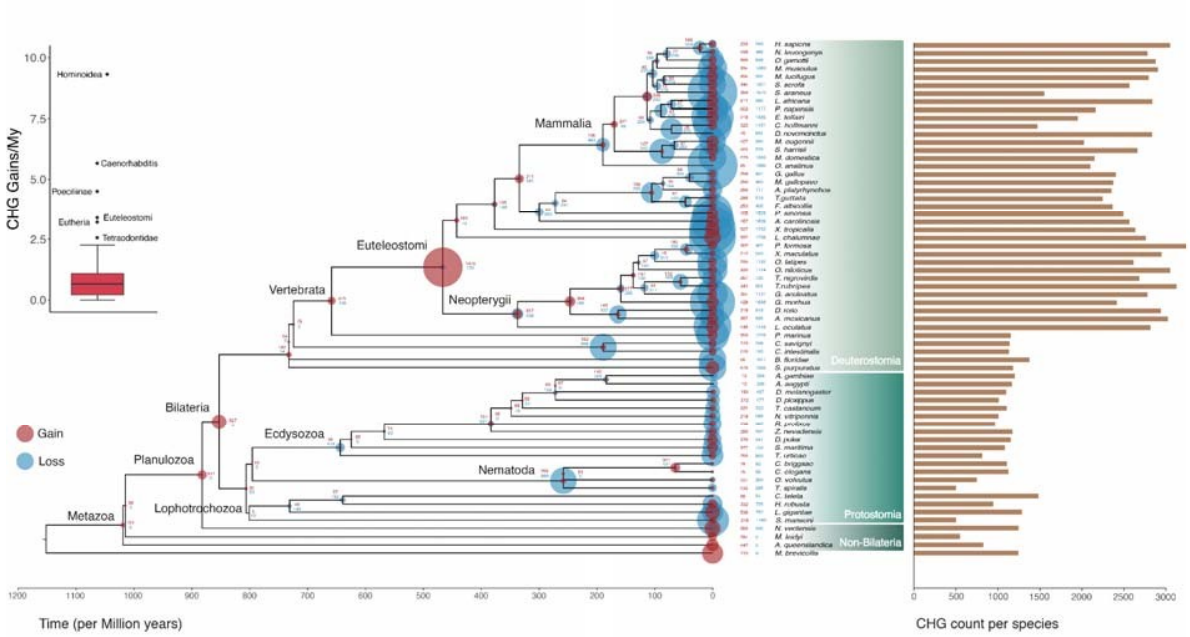


Figure 4. Distribution of the gain and loss of CHGs across the animal tree. The species phylogeny for our sample set is shown in the centre, with divergence time estimates in millions of years ago (MYA) on the x-axis (taken from TimeTree (53)). Gains of CHGs are shown as red discs and losses as blue. The size of the disc on the node is proportional to the amount of gain/loss at that node. The associated histogram on the right shows the total number of CHGs identifiable in the genomes of extant species. **Inset:** boxplot showing overall rate of gain for each node in the phylogeny, outlier nodes are named.

Competing Interests

The authors declare no conflict of interest.

Data availability

All data and code used in this study are publicly available. Find all necessary information deposited at:
https://figshare.com/projects/CompositeGenesMetazoa_Mulhair_et_al_/127943

Acknowledgements

This work was undertaken on ARC3, part of the High-Performance Computing facilities at the University of Leeds, UK. We thank Martin Callaghan and all members of the ARC team for their excellent technical support. PM was funded through a University Academic Fellowship to MOC at the University of Leeds. RM was funded by the IRCSET PhD scholarship (GOIPG/2014/306). The authors wish to acknowledge the Irish Centre for High-End computing (ICHEC) for the provision of computational facilities and support. EB and JP were supported by a FP7/2007-2013 Grant Agreement # 615274, category LS8). C.J.C. wishes to acknowledge funding from the European Commission via Horizon 2020 (818368, MASTER with K.S.T. and 101000213 HoloRuminant); FJW was supported by a Marie Skłodowska-Curie Individual Fellowship (GA no. 793818) and a University of Nottingham Anne McLaren Fellowship. DP wishes to acknowledge funding from the John Templeton Foundation (#62220 although the opinions expressed in this paper are those of the authors and not those of the John Templeton Foundation) and the Gordon and Betty Moore Foundation (GBMF9741). We would also like to thank all current and former members of the O'Connell and McInerney research groups for invaluable discussions and insights over the course of this work. We are publishing under an open access license.

Chapter IV. Conclusion and perspectives

Over the course of my doctoral studies, I have developed and applied several methods of gene family analysis, based on representations of data as sequence similarity networks, to study particular types of homology relationships. The first part of my work focused on detecting and characterising remote homologues of known gene families. To that end, I have contributed to the development of SHIFT, a new tool using iterative sequence alignments to retrieve increasingly divergent homologues of an input gene or protein family. I have then applied this protocol to explore the diversity of a large marine metagenome, targeting in particular divergent variants of highly conserved gene families. The second focus of my PhD work centred on combinatorial evolution, and specifically gene fusion and fission events. I have designed a polarisation method, complementary to composite gene detection techniques already in place in our lab, which allows the classification of composite families in fusion and fission events. These methods were then used to study the influence of gene remodelling in the evolution of two different eukaryotic lineages that share the particularity of having evolved complex multicellularity independently. In this conclusion, I summarise and discuss my contribution to both of these research themes.

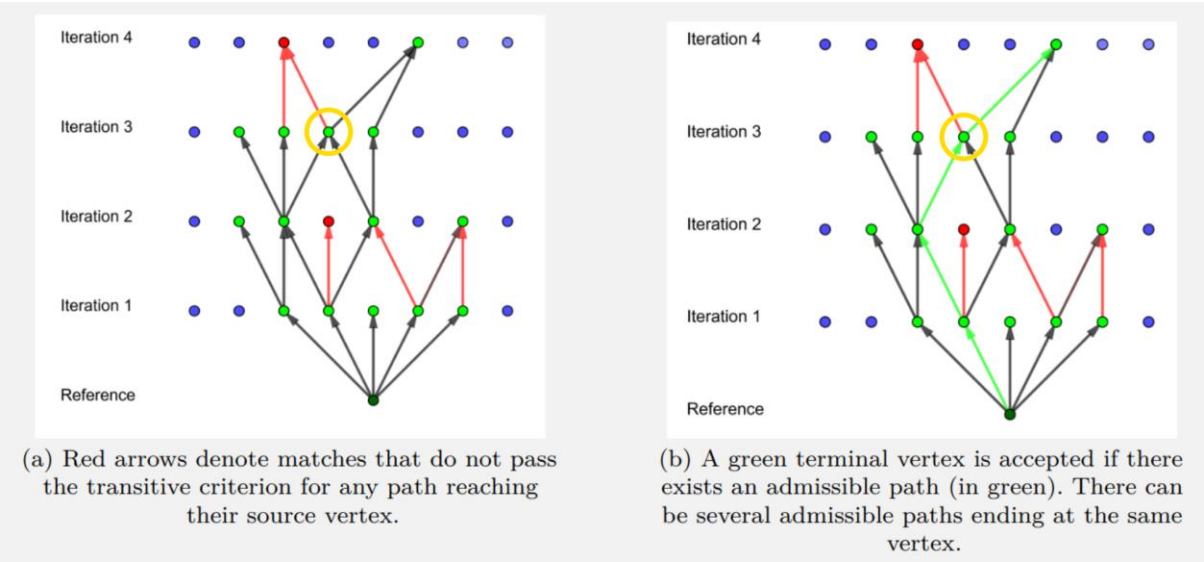
1. Exploring the oceanic microbial dark matter with remote homology searches

1.1 – Detecting distant homologues with SHIFT

In the Chapter I of this thesis, I detailed the study that we conducted on surveying the microbial dark matter inside an ocean metagenome. This analysis consisted in identifying highly divergent variants of universally conserved genes that are considered to be as old as LUCA. The methodological backbone of this study was SHIFT, a programme we developed to identify remote homologues of an input set of sequences (e.g. from the same gene family).

The foundational ideas behind SHIFT were first laid out by two of my supervisors in a 2015 article [Lopez, Halary, and Bapteste 2015], in which they conducted two rounds of BLAST searches in a collection of metagenomes: a first round gathered the direct environmental homologues of query sequences, which were then used as queries themselves for a second round of search, to retrieve their own homologues in the target dataset. The study found that the second-degree homologues of query sequences were more divergent from the known microbial diversity than first-degree homologues.

Given those results, the logical extension of this approach would be to ask: if we looked for third-degree homologues, and beyond, could we find even more divergent variants? This idea was implemented by Romain Lannes, who completed his PhD in the lab before my arrival and created a first prototype of SHIFT, which built on this principle and allowed multiple iterations of search to take place instead of just two [Lannes 2019]. The version of SHIFT that I developed during my doctoral studies (in close collaboration with Eduardo Corel) is based on the same idea of iterative searches: the homologues retrieved in the *n*-th search step are used as queries in step *n+1* to find their own homologues, such that increasingly distant homologues of the seed sequences are reached step by step[17].



(a) Red arrows denote matches that do not pass the transitive criterion for any path reaching their source vertex.

(b) A green terminal vertex is accepted if there exists an admissible path (in green). There can be several admissible paths ending at the same vertex.

Figure 31: Homology searches by SHIFT can converge to the same sequence via different paths.

The crux of SHIFT resides in the 'sanity check' step, which ensures that any retrieved homologue can be mapped back onto at least 80% of a seed sequence. This step is essential to limit the risk of overextending the search into regions of the sequence space that are not bona fide homologous to the initial sequence set. Projecting the aligned region of a newly matched sequence back onto a seed sequence, through a chain of alignments, is relatively trivial in the case where sequences are aligned without any gaps: aligned positions define 'columns' of the alignment stack, and the columns corresponding to the alignment on the final sequence can be checked to cover a sufficiently large interval of the initial sequence. However, sequence alignments are rarely ungapped in practice, and the mutual cover check becomes much less trivial in the general case (with gapped alignments). This

[17] Here I wish to voice a mild annoyance of mine around the locution "de proche en proche", which is commonplace in French and explains the intuition behind SHIFT very efficiently, but which somehow has no direct equivalent in English – I cannot count the number of different formulations I have used or considered to express this idea during the writing of this manuscript!

specific issue is the one that was targeted by most of the improvements to the prototype version of SHIFT, in order to formalise as much as possible this crucial part of the algorithm while preserving the computational efficiency of our implementation. We have identified a sound criterion to apply in order to check the mutual coverage of all sequences along an alignment stack, but some unsolved issues remain in the implementation of this criterion, which at the moment is still quite cumbersome computationally. In a nutshell, when using a retrieved environmental sequence as a query for a new round of search, some information must be stored with respect to how it was retrieved (i.e. via which alignment stack/s). Looking at Figure 31, we see that three different paths are admissible to reach the sequence circled in yellow (in addition to the path highlighted in green, there is one to its left going through a different sequence at Iteration 1, and one to its right going through a different sequence at Iteration 2). Each of these paths defines a different alignment stack, which may correspond to slightly different regions on the sequence circled. Once that sequence is used as query in Iteration 4, it may match new sequences that could be admissible for one alignment stack, but not for others. Therefore it is necessary to keep track, for each query sequence, of all the admissible paths that lead to it from a seed, and of the regions covered on intermediate sequences along each of these paths. Storing and processing this information explicitly would require a lot of time and memory space, and we have lowered this complexity by only saving a reduced representation of it, i.e. the complete paths that match positions on every sequence in the stack. Still, there is a significant computational cost when many alignment stacks are acceptable and have many complete paths. The main upcoming challenge in the development of SHIFT is now to reduce this complexity further, either by identifying an efficient heuristic that can circumvent suboptimal alignment stacks, or by using a more performant representation of the alignment data that could apply the mutual cover check to newly matched sequences with fewer computations.

The problem of remote homology is in no way new in the field of bioinformatics, and multiple algorithms have been proposed to overcome the boundary of the twilight zone of protein similarity. Different ways to represent and compare biological sequences, in particular, have been developed to improve the sensitivity of aligners: rather than using direct pairwise sequence-sequence comparisons (e.g. Smith-Waterman, BLAST), sequences can be compared to profiles (PSI-BLAST) or HMMs (HMMER) that encapsulate the information of a multiple-sequence alignment between sequences that are already considered homologous. Taking the idea one step further, target sequences can also be represented as profiles or HMMs, allowing for pairwise profile-profile (COMPASS) and HMM-HMM (HHblits) alignments. In tandem with innovations in the types of comparisons carried out, another algorithmic avenue to identify remote homologues can be found in iterative searches, alternating between phases of (i) searching for new homologues and (ii) updating the search model (e.g. the PSSM

profile or HMM) with new sequences found in the search phase. Notable examples of remote homology methods based on iterative model refinements include PSI-BLAST (profile-sequence comparisons), JackHMMER (HMM-sequence) and HHblits (HMM-HMM).

SHIFT is fundamentally similar to these methods, namely in the use of iteration to further the search for homologues from an initial gene or protein family, either until no new sequence is found or a maximum number of iterations is reached. However, SHIFT also differs from those, chiefly in that it does not use (representations of) multiple sequence alignments as queries, but rather individual sequences – albeit with implicit constraints relative to the stacking of successive alignments. This presents a certain computational advantage, because pairwise alignments are faster to construct than comparisons to sequence profiles or HMMs. The 'model actualisation' phase is also highly simplified, as sequences matched at a given search step essentially *are* the model for the next search. This also provides a level of traceability to the output of SHIFT, in the sense that we can reconstruct the path leading to the retrieval of any homologue from a seed sequence. Lastly, because SHIFT uses primary sequences as targets, it can be readily applied to mine any sequence dataset, without requiring pre-processing the target data to format it into profiles or HMMs, as is the case for HHblits for instance.

In the simulations that we conducted to test the precision and recall power of SHIFT, we found that it gathered homologues in a relatively conservative way, without ever calling a false-positive homology. However, this comes at the necessary expense of its sensitivity, meaning that SHIFT fails to gather extremely distant homologues when their divergence rate exceeds a certain point. This is preferable to the opposite situation of high sensitivity and low precision, as this would result in considering as homologues many sequences that are not, but improvements to the recall power could still be desirable, for instance by identifying better heuristics to perform the mutual coverage check between matched and seed sequences. Furthermore, the risk of homology overextension is low, but not entirely absent, and indeed in our use of SHIFT on real-world data we identified a few cases where sequences were likely retrieved beyond the boundaries of homology. In such cases of overextension, we can typically see the number of sequences retrieved in each round decrease, then significantly increase again before eventually converging to zero. This can be expected when a small number of non-homologues are erroneously retained, after which many other sequences from the same 'foreign' family are also included; generally, these sequences will also have higher similarity to their published counterparts than some bona fide homologues of seed sequences.

Arguably, this pattern can also arise from possibly desirable cases where a distant paralogue of the seed family is retrieved during the search. We observed an instance of this due to the presence of two ancient paralogues in our selection of conserved families, namely the SecD and SecF translocation

proteins [Pogliano and Beckwith 1994, Hand et al. 2006]. These proteins were retained as separate families in our initial dataset, and were thus used independently for SHIFT homology searches. However, the environmental homologues that were retrieved for these families largely overlapped, amounting to 73% of all sequences matched by SecD and/or SecF. This shows that SHIFT is able to recover ancient paralogy relationships, and further analyses of these families' SSNs enabled us to map back 80% of sequences matched by both SecD and SecF to one family or the other. However, although we were able to identify this specific occurrence as a paralogy reconstitution rather than an overextension, a few other seed families do appear to have retrieved genuinely unrelated sequences during their extension by SHIFT. This therefore represents another potential refinement of our method. In future, incorporating other kinds of biological information in addition to primary sequences could be beneficial to the precision and sensitivity of distant homology searches. Protein 3D structures, for instance, are generally more conserved than sequences, and the recent advances in structural prediction and comparison could be leveraged to improve the retrieval of remote homologues and eliminate anomalous hits that may result in situations of overextension.

1.2 – Environmental diversity of highly conserved gene families

Although SHIFT could conceptually be applied to explore the diversity of any gene or protein family of interest, we developed it with a specific goal in mind: unravelling the environmental genetic variation of highly conserved 'core' genes, especially in uncultured microorganisms. We found that many of those gene families existed in the global ocean with great diversity relative to what is known from cultured species, and that some groups of environmental variants were compatible with putative new deep-branching lineages. Detailed investigations into three specific families showed the different kinds of biological insights that can be expected from this added environmental diversity. In certain families, such as what we observed in polymerase clamp loaders, divergent sequences can be found throughout the gene's phylogeny, suggesting that genetic variation within those families is largely underrepresented in isolate genomes alone. However, divergent environmental homologues may also come from more specific components of the microbiome, for instance certain size fractions (e.g. divergent recombinases from ultra-small organisms) or certain taxa (e.g. structural variants of SMC proteins in Actinobacteria). These results therefore highlight the multifaceted nature of microbial dark matter, which could contribute biological novelty to gene families in a variety of ways, both quantitatively and qualitatively.

From a methodological standpoint, this study demonstrates a new approach to explore the unknown fraction of the gene universe that is particularly well suited for unravelling the environmental diversity of targeted gene families. Previous research groups have applied comparable

remote homology techniques to characterise the dark protein space, albeit with somewhat different purposes, such as the functional annotation of metagenomic ORFans [Lobb et al. 2015] or the identification of novel protein domains [Bitard-Feildel and Callebaut 2017]. On a broader level, the issue of resolving gene function in the microbial dark matter has concentrated numerous research efforts, based on a diversity of approaches including sequence clustering [Brum et al. 2016, Vanni et al. 2022, Pavlopoulos et al. 2023], structural comparison [Durairaj et al. 2023], and deep learning [Bileschi et al. 2022]. Thanks to this plurality of complementary approaches, this collective undertaking is generating new insights into the coding potential of unknown microbial genes, and sketching out the underlying organisation of the overall gene space [Vanni et al. 2022]. In the early years of microbial dark matter research, the majority of the spotlight was occupied by the discovery of novel major lineages (CPR bacteria, DPANN archaea, Asgard archaea, etc.) that was enabled by the reconstruction of extended phylogenetic trees that included MAGs as well as reference genomes [Hug et al. 2016, Castelle and Banfield 2018]. Since then, more room has been made for function-oriented analyses, concomitantly with the realisation that many uncharacterised sequences are not covered by MAGs, such that alternative methods may be necessary to resolve their evolutionary and taxonomic origin. In this context, our multi-marker approach could be relevant to the formulation of new evolutionary hypotheses that could guide future explorations of metagenomes: in the same way that Wyman et al. [Wyman et al. 2018] proposed a “most-wanted list” of conserved but unannotated protein families, so could our protocol suggest a selection of “most-wanted lineages” that encode divergent variants in one or several core gene families. Being able to upscale this method to larger sets of both query families and target datasets, and possibly automating at least some part of the subsequent SSN analysis, would be especially interesting for this last objective, in order to provide a “most-wanted list” that would be relevant for a larger research community.

Although our results using SHIFT to mine the Tara Oceans metagenome are already interesting, a few additional analyses could still be completed in order to bring the study to its full potential. In particular, our investigations into some divergent clusters of specific gene families could be enhanced by digging deeper into their biological context. We were especially interested in finding new gene variants that could suggest the existence of novel basal branches in the tree of life, i.e. currently undiscovered major lineages [Wu et al. 2011]. Some of the divergent groups of environmental homologues that we identified are compatible with such lineages, and the fact that they feature in highly conserved gene families is encouraging in this regard. However, as things presently stand, these divergent clusters were found on a single-marker basis, so that these new gene variants exist in relative isolation to each other and to the rest of the gene space. This limitation makes it difficult to confidently assess their significance outside the evolutionary history of their family. Linking these

variants with other gene products may therefore enlighten us about the genomes, and perhaps even the organisms, bearing these divergent genes.

In order to overcome this limitation, a first step would be to search for them in collections of MAGs and single-cell genomes, especially assembled from Tara Oceans or other marine sequence datasets. Not only could this allow us to measure their phylogenetic distribution in the tree of life, but it could also yield significant information on functional partners of these variants, possibly revealing compensatory adaptations to maintain regular function despite their divergence or, on the contrary, co-evolution patterns in other related genes that may lead to functional innovations in their hosts. The marine virosphere, in particular, is rife with bacteriophages and archaeal viruses that encode a variety of DNA processing genes, and thus some of the gene variants reported may derive from phage genomes, or possibly prophage insertions in genomes of cellular organisms. On the other hand, since only a fraction of metagenome-predicted genes are covered by MAGs, some lower-level analyses may prove equally fruitful to understand the broader context of these divergent clusters of core gene homologues. For instance, given the particularly broad range of locations and environments represented in the OM-RGC dataset, biogeographical annotations are available for the environmental homologues we retrieved, and could thus provide an adequate resolution for co-occurrence analyses. Pronounced levels of co-occurrence between some of our variants of interest, or with other marine sequence clusters, may lead to further insights into their ecological and functional role, especially when genome-level information is unavailable. Finally, in a broader sense, other microbiome data could be explored, especially from different environments than the global ocean, to understand whether these variants are exclusive to marine life or also occur in other habitats on Earth.

Some of the variants identified by our analysis also have particular structural features that raise the question of a possible functional divergence, such as the hinge-less SMC-like proteins found in Actinobacteria. Further research would be required to confirm whether these variants indeed occupy a different function than their more ‘canonical’ homologues, and this could happen in a number of ways. Firstly, breakthroughs in the *in silico* prediction of protein structure, with AlphaFold2 at the forefront, have recently been followed by advances in predicting the structure and interaction of entire complexes of biomolecules (including proteins, RNAs, DNA segments, ligands, etc.) with AlphaFold3 [Abramson et al. 2024]. In the case of our hinge-less SMC proteins, for instance, this could be leveraged by testing computationally the assembly of an SMC complex with either one hinge-less and one regular SMC, or with two hinge-less SMC, as well as the usual accessory proteins. The interaction of these hinge-less or semi-hinge-less complexes with a DNA molecule could even be simulated, to predict how, if at all, these could interface with DNA. However, even these *in silico*

predictions using the most recent deep-learning algorithms of structural biology would eventually require testing *in vitro* and/or *in vivo*. Lab-grown strains of Actinobacteria could for instance be injected with a plasmid encoding a hinge-less SMC variant. This would then allow us to test (i) whether the hinge-less variant is at all expressed in the bacteria, using transcriptomic readings, and (ii) whether this variant is viable for the bacterial population when the 'regular' SMC gene is knocked out. These hypotheses for experimental testing are of course easier to formulate than to actually carry out, but they could still one day be implemented if a given environmental variant is of particular interest for a specific purpose that we (from our bioinformatic standpoint) may not suspect.

2. Gene fusion, gene fission, and the evolution of complex multicellularity

The second chapter of my thesis is dedicated to the study of gene remodelling processes, in particular gene fusion and fission. My work on that subject has been built upon knowledge previously developed at the Lopez & Bapteste lab, in particular by former PhD students Pierre-Alain Jachiet, who transposed the non-transitive gene homology framework to the level of gene families with MosaicFinder [Jachiet et al. 2013], and Jananan Pathmanathan, who generalised in CompositeSearch the detection of composite gene families by implementing a more flexible variation of the non-transitivity model [Pathmanathan et al. 2018]. I applied this latter method to study gene remodelling in two eukaryotic lineages that both evolved towards a complex multicellular lifestyle. I also made methodological contributions to this field, chiefly by developing a polarisation method that allows the classification of remodelling events detected by CompositeSearch into events of gene fusion and gene fission.

2.1 – Comparing the role of gene remodelling in the evolution of animals and brown algae

During my doctoral work, I have been involved in two collaborations that have allowed me to analyse gene remodelling dynamics in two different branches of Eukaryotes: brown algae, within the Phaeoexplorer consortium, and animals, with Mary O'Connell's group at the University of Nottingham. Although they sit in different places in the eukaryotic tree of life, these two lineages have in common the fact that they independently evolved complex multicellularity (CMC) from unicellular ancestors.

The emergence of CMC in Eukaryotes is extensively studied across the five main multicellular groups (animals, land plants, brown algae, and some lineages of fungi and red algae) in which it emerged, seemingly at least 16 different times in total [Sebé-Pedrós, Degnan, and Ruiz-Trillo 2017].

Among these five groups, animals and fungi (collectively, opisthokonts) are particularly well studied. However, the lens through which CMC is investigated is often a functional one: identifying the genetic toolkit that is associated with multicellularity enables a better understanding of the physiological changes that preceded the transition to CMC and ensued from it. In fungi, for instance, multicellular groups are distributed throughout the group's phylogeny, and transversal studies are performed across these groups to understand whether fungal CMC stems from a single origin (followed by a consequent number of losses) or emerged convergently in multiple clades [Nagy, Kovács, and Krizsán 2018]. In this context, comparing the genetic toolset that is involved in CMC in each multicellular group could help elucidate the origin/s of this phenotype. In 2022, Ruiz-Trillo and colleagues [Ocaña-Pallarès et al. 2022] found little support for either hypothesis regarding fungal CMC. However, their broader analysis of genomic trajectories in Opisthokonta provides, in addition to functional analyses, a mechanismic perspective into the onset of CMC in animals and fungi. Their results reveal divergent dynamics of genomic changes in the evolutionary stages before, during and after the emergence of metazoans and of fungi. In particular, they highlight marked gene gains at the root of Metazoa in functions associated with multicellularity (transcription, signal transduction, extracellular structures), and a significant contribution of gene fusions in their genomes, whereas fungi favour gene acquisition by horizontal transfer. In line with this work, our investigations of gene remodelling in animals and in brown algae allow us to draw a number of comparisons between the two, specifically about the ways in which gene fusion and fission events contributed, if at all, to their respective transition to CMC. In particular, because the single origin of multicellularity in animals and in brown algae is well established compared to CMC in fungi, we can compare with better clarity the role of gene remodelling before, during and after the onset of CMC.

Perhaps the most comparable characteristic of remodelled genes in brown algae and animals is their distinct bias towards similar subsets of functional categories. In animals as well as brown algae, gene products of fusion and fission events contribute significantly to information processing pathways (especially translation and signal transduction), as well as functions relevant to cellular and intercellular structures. Both of these functional classes are relevant to CMC, in particular via the complexification of regulatory networks, extracellular matrices and cell-cell communication, which play a role in the diversification of tissue types and the progression of the life cycle. This first comparison thus suggests that in both animals and brown algae, combinatorial gene processes have been mobilised preferentially for the development of phenotypic changes that are associated with multicellularity.

Another observation concerns the frequency of these events across those lineages, in absolute numbers as well as in the fusion-fission ratio. In brown algae, we observed that more than 12% of all gene families were putative composites, compared to only 5% in animal genomes. In both lineages, gene fusions were more frequent than fissions, but to different extents: in animals, the number of fusions was triple that of fissions, broadly conforming with estimations from the literature, whereas brown algal genomes only experienced 40% more fusions than fissions. The genomes of brown algae are therefore much richer in remodelled genes overall, and are particularly enriched in gene families created by fission events, a pattern that has previously been observed in fungi [Leonard and Richards 2012]. Metazoans, on the other hand, seem to rely mostly on gene fusions when it comes to combinatorial evolution, in line with previous estimations of relative rates of fusion and fission across the tree of life [Kummerfeld and Teichmann 2005]. The increased intron content in genomes of brown algae compared to other stramenopiles could be a possible explanation for the unusually high rate of gene fission observed in this lineage, by offering more 'splitting sites' within genes without affecting exon sequences.

This apparent difference in affinity for certain types of gene remodelling events is further attested by the relative retention rates of fused and split gene families in each lineage. In brown algae, the genetic products of fusion and fission events are retained in extant genomes at comparable rates, which exceed the retention of non-remodelled gene families. In animals, however, gene fusion products are preferentially retained compared to non-remodelled genes (predominantly in vertebrate genomes, which concentrate the bulk of gene remodelling in animals; see below), but split genes created by fissions are lost significantly more. The preferred mechanisms of genome evolution thus vary across lineages, with strong biases in favour of gene fusions in animal genomes, as opposed to a more balanced contribution of fusion and fission in the case of brown algae.

Lastly, the chronology of gene remodelling also follows different patterns in the two lineages studied here. Most of the remodelled genes that are present in brown algal genomes date back to the initial emergence and the early evolution of brown algae, with much rarer fusion and fission events in the subsequent stages of Phaeophyceae diversification. Given the preferential retention of remodelled gene families in extant algal genomes, this suggests that genetic fusion and fission may have given rise to a number of new 'core' functions in brown algae that participated in the emergence of this lineage and persist in many algal species to this day. On the other hand, gene gains due to remodelling events are concentrated at specific points in the animal tree of life, predominantly much later than the initial evolution of animals, a result in accordance with previous findings [Ocaña-Pallarès et al. 2022]. The extent of gene remodelling also varies significantly from lineage to lineage, with a

marked tendency of different clades of animals evolving similar remodelled genes convergently. The combinatorics of genes and gene parts in Metazoans therefore appear highly dynamic – with fusions in particular contributing to a number of important phenotypic transitions – and generally seem to have played a more significant part in the diversification of animals than during their early evolution.

In summary, gene remodelling processes have had significant effects on the evolution of both animals and brown algae, in particular in relation to functions associated with complex multicellular phenotypes. Although these functions have been affected by gene remodelling in animals and brown algae at different stages of their respective evolution, in both cases they appear to have participated in physiological and morphological innovations at key points of these multicellular lineages. However, it is interesting to note that these contributions have been made via different genomic trajectories. On the one hand, brown algae have acquired many remodelled genes relevant to CMC at the onset of their lineage, both by means of gene fusion and gene fission, and these genes have been largely retained in extant genomes. On the other hand, in the evolution of Metazoa, gene fissions have played a seemingly negligible role, whereas bursts of gene fusions have induced significant gene gains at specific key points of the animal tree of life. Of particular note is the repeated evolution of similar gene fusions at several points in the tree, amounting to 41% of all composite gene families. In an arborescent, gradualist conception of evolution, convergence in gene sequence can occur under particular selective constraints but is nonetheless rare, and non-adaptive convergence in particular is virtually impossible. In a combinatorial framework that takes genetic rearrangements into account, however, the reinvention of gene forms (e.g. domain architectures) becomes possible via punctuated events of gene fusion or fission. Indeed, at least 25% of multi-domain proteins[18] in Eukaryotes have emerged convergently [Zmasek and Godzik 2012], and several specific fusions with multiple origins have also been identified in animals [Cosby et al. 2021]. This repeatability of successful genetic innovations suggests a highly modular organisation of the gene space, and may raise challenges for many evolutionary and bioinformatic approaches that rely heavily on assumptions of orthology in homologous families.

2.2 – Polarising gene remodelling events using Dollo parsimony

In order to improve the descriptive power of remodelling analyses based on CompositeSearch, I developed a post-treatment method that allows composite gene families to be further classified into gene fusions and gene fissions. It relies on the simple idea that if extant composite and component gene families can be traced back to their ancestral node of origin in the tree of life, then we can

[18] And therefore 20% of all eukaryotic proteins, since 80% of them are multi-domain proteins.

compare the relative positions of these origin nodes to polarise the remodelling event: if the composite is older than its components, then it most likely has undergone a fission event giving rise to the later components, and conversely if the components predate the composite, then that composite must have been created by fusion of the components. This heuristic has the advantage of being easy to implement, as well as computationally efficient, and was able to classify most of the composite families in the two datasets we applied it to. Still, it has a number of conceptual or practical shortcomings that could be addressed to make it applicable to a broader range of studies and improve the reliability of its outputs.

First, this method relies heavily on the phylogeny of species present in the dataset to infer the phylogenetic origin of each gene family. In other words, it must make the assumption that the lineage being studied evolves in a strong 'tree-like' manner, i.e. that the effects of introgressive processes on its evolution are negligible. This is acceptable for most eukaryotic lineages, and in particular multicellular ones, but it may be more questionable when trying to study gene remodelling in prokaryotes or viruses, among which the dominance of horizontal gene transfer brings the evolution model closer to a network than a phylogenetic tree. In these 'non-tree-like' lineages, understanding the dynamics of gene fusion and fission would require alternative methods that also take lateral gene flow into account. If permitted by future methodological developments, studying the coordinated effects of horizontal transfer and gene remodelling in Bacteria and Archaea could lead to some fascinating insights into their evolution. During my doctoral studies, I actually attempted something similar by trying to identify chimeric fusions in the organelles of photosynthetic eukaryotes that would unite genetic material from both mitochondrial and chloroplastic genomes. However, the methodological hurdle of reconciling the phylogenies of eukaryotes, mitochondria and chloroplasts (which have been acquired in several lineages by endosymbiosis of another photosynthetic eukaryote, and therefore have a phylogeny that is incongruent with that of their hosts) prevented us from producing conclusive results within the time constraints of my doctoral studies.

The resolution of our polarisation method also benefits from well-balanced phylogenetic trees, stemming from a relatively uniform sampling of the diversity of species within the lineage studied. This allows for more accurate estimations of the points of emergence for gene families, as well as less biased comparisons between different groups of the species tree. For instance, in the animal tree of life that we used in our study, only three branches correspond to non-bilaterians, as opposed to sixty bilaterians, meaning that little insight can be confidently gained about combinatorial evolution in those basal animal groups. This does not invalidate the results that are identified in other parts of the species phylogeny, but it should be borne in mind when comparing the dynamics of gene remodelling

between different clades, e.g. that have the same taxonomic rank or that emerged around the same point in time.

Another aspect to keep in mind regarding our approach is that it can only characterise the remodelling events within the studied lineage and cannot identify ancestral remodelled genes that emerged prior to the last common ancestor of the species set. This explains why, in both studies for which we applied this polarisation, outgroup species (other Stramenopiles in the case of brown algae; choanoflagellate *Monosiga brevicollis* for animals) were included: their presence allows to separate gene families that appeared in the animal, or brown algal, ancestor from those that are more ancient. Ancestrally remodelled genes cannot be detected beyond the root of the species phylogeny, and therefore their fate in more modern branches cannot be studied: are they just as conserved as new remodelled genes in extant genomes, have they become obsolete and therefore largely lost, or perhaps replaced by other, newer ones? These questions concerning gene turnover could be relevant to clarify the broader dynamics of gene remodelling and would be permitted by the inclusion of genomic data from more outgroup species.

Lastly, the Dollo parsimony model that is used to decide the emergence points of gene families is perhaps a little simplistic and is sensitive to the way gene families are defined in the genomic dataset. One particular weakness is that it does not account for the fact that remodelled genes may replace their ancestral forms, rather than exist in tandem with them. For instance, in a species where two adjacent single-copy genes A and B become fused following the disappearance of their separating intergenic region (and therefore A and B are not encoded as separate genes anymore), then the gene families A and B will be considered absent from that species. Since the nodes of origin of gene families are inferred according to their presence/absence data in extant genomes, this can alter the outcome of that inference, and consequently the polarisation of remodelling events. In fact, a majority of remodelling events can be affected by this: indeed, in animal genomes, we found that in 82% of all remodelling events, the contributing genes (i.e. the composite parent in the case of a gene fission, and the component parents for a fusion) were lost and only the remodelled products remained in extant genomes. The method's accuracy in inferring the origination of each gene family could therefore be improved by taking this phenomenon into account, for instance in cases where counting the presence/absence of components and composites together resolves the paraphyly or polyphyly of some families. Furthermore, in some lineages, genetic rearrangement may be highly dynamic, and it is possible that some remodelling events may be subsequently reverted, e.g. a fused gene undergoing fission and returning to a split form. We tried to address some of those cases when we detected composite families that had an older component but also a more recent one – we then ascribed this

pattern to a fusion followed by a subsequent fission. However, other phyletic patterns can arise from cases of consecutive remodelling events, and refining the parsimony model to allow for some reversals to ancestral states may lead to a finer understanding of these combinatorial genetic dynamics.

3. Using similarity networks to map out the genetic space

The most frequent angle from which the evolutionary relationships between genes are established, represented and analysed is the arborescent model of phylogeny. The phylogenetic tree of a gene family, for instance, simultaneously depicts its diversity and its evolutionary history, in a simple representation that proposes an unambiguous reconstitution of the events that led to the family's contemporary state. It also provides a full hierarchy of the proximity between pairs of sequences, which allows us to adjust the granularity of the model by sorting sequences into coherent groups, all while preserving the hierarchical information between these groups. However, some evolutionary processes are best described by other models than purely arborescent ones, typically when they involve other motifs than evolutionary 'forks' where one ancestor gives rise to two or more offspring [Haggerty et al. 2014]. Such processes exist at the scale of genes (e.g. recombination, horizontal gene transfer), of organisms (e.g. endosymbiosis, hybridisation) and of populations (e.g. admixture).

Over the course of this thesis, the main approaches that we have adopted to model gene families and study their evolution have been through the lens of networks of interconnections, and in particular sequence similarity networks. This representation contrasts with phylogenies in its conception of relationships between sequences: whereas phylogenetic trees are strongly hierarchical, networks are much more horizontal and include information on the proximity of any two sequences. This is not a fundamentally better or a worse model than the arborescent one, nor is it meant to replace it. Rather, networks offer a complementary viewpoint to phylogenies: the former view focuses on the overall structure of the gene space at a given instant, without painting a clear picture of the underlying evolutionary trajectory, whereas the latter view proposes the opposite. Indeed, readers of this thesis will have probably noticed the frequent use of phylogenies as a complement to network analyses. When we detected clusters of divergent environmental variants in SSNs of conserved microbial families, we relied on phylogenetic trees to understand their contribution to the diversity of said families; likewise, when we identified putative remodelling events in SSNs from algal and animal genomes, their classification into fusions or fissions was guided by the phylogeny of the species present in the dataset.

In our specific case, the choice of sequence similarity networks as a framework allows us to overcome some limitations of the classic tree-like approach to map out the diversity of the global genetic space. In particular, we described remote homology relationships that can escape detection by canonical homology search methods and impede the reconstruction of multiple-sequence alignments that most phylogenetic trees are based upon. We also characterised processes of combinatorial, non-linear evolution that are also overlooked by these techniques, and that are intrinsically incompatible with the arborescent representation of gene family evolution. Our use of network-based methods to alleviate some shortcomings of more canonical approaches, themselves complemented by some tree-based analyses when necessary, illustrates the benefits of a conceptual and methodological pluralism to understand the diversity of genes, organisms and evolutionary mechanisms in their globality.

Chapter V. Bibliography

Abramson, Josh, Jonas Adler, Jack Dunger, Richard Evans, Tim Green, Alexander Pritzel, Olaf Ronneberger, et al. 2024. 'Accurate Structure Prediction of Biomolecular Interactions with AlphaFold 3'. *Nature* 630 (8016): 493–500. https://doi.org/10.1038/s41586-024-07487-w.

Altschul, Stephen F., Warren Gish, Webb Miller, Eugene W. Myers, and David J. Lipman. 1990. 'Basic Local Alignment Search Tool'. *Journal of Molecular Biology* 215 (3): 403–10. https://doi.org/10.1016/S0022-2836(05)80360-2.

Altschul, Stephen F., Thomas L. Madden, Alejandro A. Schäffer, Jinghui Zhang, Zheng Zhang, Webb Miller, and David J. Lipman. 1997. 'Gapped BLAST and PSI-BLAST: A New Generation of Protein Database Search Programs'. *Nucleic Acids Research* 25 (17): 3389–3402. https://doi.org/10.1093/nar/25.17.3389.

Apic, Gordana, Julian Gough, and Sarah A Teichmann. 2001. 'Domain Combinations in Archaeal, Eubacterial and Eukaryotic Proteomes1'. *Journal of Molecular Biology* 310 (2): 311–25. https://doi.org/10.1006/jmbi.2001.4776.

Archibald, John M. 2008. 'The Eocyte Hypothesis and the Origin of Eukaryotic Cells'. *Proceedings of the National Academy of Sciences* 105 (51): 20049–50. https://doi.org/10.1073/pnas.0811118106.

Aslam, Bilal, Madiha Basit, Muhammad Atif Nisar, Mohsin Khurshid, and Muhammad Hidayat Rasool. 2017. 'Proteomics: Technologies and Their Applications'. *Journal of Chromatographic Science* 55 (2): 182–96. https://doi.org/10.1093/chromsci/bmw167.

Bailey, Susan F, Luz Angela Alonso Morales, and Rees Kassen. 2021. 'Effects of Synonymous Mutations beyond Codon Bias: The Evidence for Adaptive Synonymous Substitutions from Microbial Evolution Experiments'. *Genome Biology and Evolution* 13 (9): evab141. https://doi.org/10.1093/gbe/evab141.

Baker, Brett J., Luis R. Comolli, Gregory J. Dick, Loren J. Hauser, Doug Hyatt, Brian D. Dill, Miriam L. Land, Nathan C. VerBerkmoes, Robert L. Hettich, and Jillian F. Banfield. 2010. 'Enigmatic, Ultrasmall, Uncultivated Archaea'. *Proceedings of the National Academy of Sciences* 107 (19): 8806–11. https://doi.org/10.1073/pnas.0914470107.

Bapteste, Eric, Maureen A. O'Malley, Robert G. Beiko, Marc Ereshefsky, J. Peter Gogarten, Laura Franklin-Hall, François-Joseph Lapointe, et al. 2009. 'Prokaryotic Evolution and the Tree of Life Are Two Different Things'. *Biology Direct* 4 (1): 34. https://doi.org/10.1186/1745-6150-4-34.

Bar-On, Yinon M., Rob Phillips, and Ron Milo. 2018. 'The Biomass Distribution on Earth'. *Proceedings of the National Academy of Sciences* 115 (25): 6506–11. https://doi.org/10.1073/PNAS.1711842115.

Benson, Dennis A., Mark Cavanaugh, Karen Clark, Ilene Karsch-Mizrachi, David J. Lipman, James Ostell, and Eric W. Sayers. 2013. 'GenBank'. *Nucleic Acids Research* 41 (D1): D36–42. https://doi.org/10.1093/nar/gks1195.

Bernard, Guillaume, Jananan S Pathmanathan, Romain Lannes, Philippe Lopez, and Eric Bapteste. 2018. 'Microbial Dark Matter Investigations: How Microbial Studies Transform Biological Knowledge and Empirically Sketch a Logic of Scientific Discovery'. *Genome Biology and Evolution* 10 (3): 707–15. https://doi.org/10.1093/gbe/evy031.

Bileschi, Maxwell L., David Belanger, Drew H. Bryant, Theo Sanderson, Brandon Carter, D. Sculley, Alex Bateman, Mark A. DePristo, and Lucy J. Colwell. 2022. 'Using Deep Learning to Annotate the Protein Universe'. *Nature Biotechnology* 40 (6): 932–37. https://doi.org/10.1038/s41587-021-01179-w.

Bitard-Feildel, Tristan, and Isabelle Callebaut. 2017. 'Exploring the Dark Foldable Proteome by Considering Hydrophobic Amino Acids Topology'. *Scientific Reports* 7 (1): 41425. https://doi.org/10.1038/srep41425.

Blaxter, Mark, Jenna Mann, Tom Chapman, Fran Thomas, Claire Whitton, Robin Floyd, and Eyualem Abebe. 2005. 'Defining Operational Taxonomic Units Using DNA Barcode Data'. *Philosophical Transactions of the Royal Society B: Biological Sciences* 360 (1462): 1935–43. https://doi.org/10.1098/rstb.2005.1725.

Blondel, Vincent D., Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. 2008. 'Fast Unfolding of Communities in Large Networks'. *Journal of Statistical Mechanics: Theory and Experiment* 2008 (10): P10008. https://doi.org/10.1088/1742-5468/2008/10/P10008.

Brito, Ilana Lauren. 2021. 'Examining Horizontal Gene Transfer in Microbial Communities'. *Nature Reviews Microbiology* 19 (7): 442–53. https://doi.org/10.1038/s41579-021-00534-7.

Brown, Christopher T., Laura A. Hug, Brian C. Thomas, Itai Sharon, Cindy J. Castelle, Andrea Singh, Michael J. Wilkins, Kelly C. Wrighton, Kenneth H. Williams, and Jillian F. Banfield. 2015. 'Unusual Biology across a Group Comprising More than 15% of Domain Bacteria'. *Nature* 523 (7559): 208–11. https://doi.org/10.1038/nature14486.

Brum, Jennifer R., J. Cesar Ignacio-Espinoza, Eun-Hae Kim, Gareth Trubl, Robert M. Jones, Simon Roux, Nathan C. VerBerkmoes, Virginia I. Rich, and Matthew B. Sullivan. 2016. 'Illuminating Structural Proteins in Viral "Dark Matter" with Metaproteomics'. *Proceedings of the National Academy of Sciences* 113 (9): 2436–41. https://doi.org/10.1073/pnas.1525139113.

Buljan, Marija, and Alex Bateman. 2009. 'The Evolution of Protein Domain Families'. *Biochemical Society Transactions* 37 (4): 751–55. https://doi.org/10.1042/BST0370751.

Castelle, Cindy J., and Jillian F. Banfield. 2018. 'Major New Microbial Groups Expand Diversity and Alter Our Understanding of the Tree of Life'. *Cell* 172 (6): 1181–97. https://doi.org/10.1016/j.cell.2018.02.016.

Castelle, Cindy J., Christopher T. Brown, Karthik Anantharaman, Alexander J. Probst, Raven H. Huang, and Jillian F. Banfield. 2018. 'Biosynthetic Capacity, Metabolic Variety and Unusual Biology in the CPR and DPANN Radiations'. *Nature Reviews Microbiology* 16 (10): 629–45. https://doi.org/10.1038/s41579-018-0076-2.

Choi, Seok-Wan, Louis Graf, Ji Won Choi, Jihoon Jo, Ga Hun Boo, Hiroshi Kawai, Chang Geun Choi, et al. 2024. 'Ordovician Origin and Subsequent Diversification of the Brown Algae'. *Current Biology* 34 (4): 740-754.e4. https://doi.org/10.1016/j.cub.2023.12.069.

Claverie, Jean-Michel, and Chantal Abergel. 2013. 'Chapter Two - Open Questions About Giant Viruses'. In *Advances in Virus Research*, edited by Karl Maramorosch and Frederick A. Murphy, 85:25–56. Academic Press. https://doi.org/10.1016/B978-0-12-408116-1.00002-1.

Cobbe, Neville, and Margarete M. S. Heck. 2004. 'The Evolution of SMC Proteins: Phylogenetic Analysis and Structural Implications'. *Molecular Biology and Evolution* 21 (2): 332–47. https://doi.org/10.1093/MOLBEV/MSH023.

Colson, Philippe, Xavier de Lamballerie, Ghislain Fournous, and Didier Raoult. 2012. 'Reclassification of Giant Viruses Composing a Fourth Domain of Life in the New Order Megavirales'. *Intervirology* 55 (5): 321–32. https://doi.org/10.1159/000336562.

Cosby, Rachel L., Julius Judd, Ruiling Zhang, Alan Zhong, Nathaniel Garry, Ellen J. Pritham, and Cédric Feschotte. 2021. 'Recurrent Evolution of Vertebrate Transcription Factors by Transposase Capture'. *Science* 371 (6531): eabc6405. https://doi.org/10.1126/science.abc6405.

Coutinho, Tarcisio José Domingos, Glória Regina Franco, and Francisco Pereira Lobo. 2015. 'Homology-Independent Metrics for Comparative Genomics'. *Computational and Structural Biotechnology Journal* 13 (January):352–57. https://doi.org/10.1016/j.csbj.2015.04.005.

Cromar, Graham, Ka-Chun Wong, Noeleen Loughran, Tuan On, Hongyan Song, Xuejian Xiong, Zhaolei Zhang, and John Parkinson. 2014. 'New Tricks for "Old" Domains: How Novel Architectures and Promiscuous Hubs Contributed to the Organization and Evolution of the ECM'. *Genome Biology and Evolution* 6 (10): 2897–2917. https://doi.org/10.1093/gbe/evu228.

Cummins, Elizabeth A., Rebecca J. Hall, Chris Connor, James O. McInerney, and Alan McNally. 2022. 'Distinct Evolutionary Trajectories in the Escherichia Coli Pangenome Occur within Sequence Types'. *Microbial Genomics* 8 (11): 000903. https://doi.org/10.1099/mgen.0.000903.

Davidson, Eric H., and Douglas H. Erwin. 2006. 'Gene Regulatory Networks and the Evolution of Animal Body Plans'. *Science* 311 (5762): 796–800. https://doi.org/10.1126/science.1113832.

Dittami, Simon M., Svenja Heesch, Jeanine L. Olsen, and Jonas Collén. 2017. 'Transitions between Marine and Freshwater Environments Provide New Clues about the Origins of Multicellular Plants and Algae'. *Journal of Phycology* 53 (4): 731–45. https://doi.org/10.1111/jpy.12547.

Dohmen, Elias, Steffen Klasberg, Erich Bornberg-Bauer, Sören Perrey, and Carsten Kemena. 2020. 'The Modular Nature of Protein Evolution: Domain Rearrangement Rates across Eukaryotic Life'. *BMC Evolutionary Biology* 20 (1): 30. https://doi.org/10.1186/s12862-020-1591-0.

Durairaj, Janani, Andrew M. Waterhouse, Toomas Mets, Tetiana Brodiazhenko, Minhal Abdullah, Gabriel Studer, Gerardo Tauriello, et al. 2023. 'Uncovering New Families and Folds in the Natural Protein Universe'. *Nature* 622 (7983): 646–53. https://doi.org/10.1038/s41586-023-06622-3.

Eccles, David. n.d. 'Polynesian Migration Map'. World History Encyclopedia. Accessed 15 October 2024. https://www.worldhistory.org/image/10691/polynesian-migration-map/.

Ekman, Diana, Åsa K. Björklund, Johannes Frey-Skött, and Arne Elofsson. 2005. 'Multi-Domain Proteins in the Three Kingdoms of Life: Orphan Domains and Other Unassigned Regions'. *Journal of Molecular Biology* 348 (1): 231–43. https://doi.org/10.1016/j.jmb.2005.02.007.

Enright, Anton J., Ioannis Iliopoulos, Nikos C. Kyrpides, and Christos A. Ouzounis. 1999. 'Protein Interaction Maps for Complete Genomes Based on Gene Fusion Events'. *Nature* 402 (6757): 86–90. https://doi.org/10.1038/47056.

Felsenstein, Joseph. 1974. 'THE EVOLUTIONARY ADVANTAGE OF RECOMBINATION'. *Genetics* 78 (2): 737–56. https://doi.org/10.1093/genetics/78.2.737.

Fitch, Walter M. 1970. 'Distinguishing Homologous from Analogous Proteins'. *Systematic Biology* 19 (2): 99–113. https://doi.org/10.2307/2412448.

Forslund, Sofia K., Mateusz Kaduk, and Erik L. L. Sonnhammer. 2019. 'Evolution of Protein Domain Architectures'. In *Evolutionary Genomics: Statistical and Computational Methods*, edited by Maria Anisimova, 469–504. New York, NY: Springer. https://doi.org/10.1007/978-1-4939-9074-0_15.

Fu, Limin, Beifang Niu, Zhengwei Zhu, Sitao Wu, and Weizhong Li. 2012. 'CD-HIT: Accelerated for Clustering the next-Generation Sequencing Data'. *Bioinformatics* 28 (23): 3150–52. https://doi.org/10.1093/bioinformatics/bts565.

Gabaldón, Toni, and Eugene V. Koonin. 2013. 'Functional and Evolutionary Implications of Gene Orthology'. *Nature Reviews Genetics* 14 (5): 360–66. https://doi.org/10.1038/nrg3456.

Giovannoni, Stephen J., H. James Tripp, Scott Givan, Mircea Podar, Kevin L. Vergin, Damon Baptista, Lisa Bibbs, et al. 2005. 'Genome Streamlining in a Cosmopolitan Oceanic Bacterium'. *Science* 309 (5738): 1242–45. https://doi.org/10.1126/science.1114057.

Gruber, Stephan, Prakash Arumugam, Yuki Katou, Daria Kuglitsch, Wolfgang Helmhart, Katsuhiko Shirahige, and Kim Nasmyth. 2006. 'Evidence That Loading of Cohesin Onto Chromosomes Involves Opening of Its SMC Hinge'. *Cell* 127 (3): 523–37. https://doi.org/10.1016/j.cell.2006.08.048.

Haggerty, Leanne S, Pierre-Alain Jachiet, William P Hanage, David A Fitzpatrick, Philippe Lopez, Mary J O'Connell, Davide Pisani, Mark Wilkinson, Eric Bapteste, and James O McInerney. 2014. 'A Pluralistic Account of Homology: Adapting the Models to the Data'. *Molecular Biology and Evolution* 31 (3): 501–16. https://doi.org/10.1093/molbev/mst228.

Halary, Sébastien, Jessica W. Leigh, Bachar Cheaib, Philippe Lopez, and Eric Bapteste. 2010. 'Network Analyses Structure Genetic Diversity in Independent Genetic Worlds'. *Proceedings of the National Academy of Sciences* 107 (1): 127–32. https://doi.org/10.1073/pnas.0908978107.

Hand, Nicholas J., Reinhard Klein, Anke Laskewitz, and Mechthild Pohlschröder. 2006. 'Archaeal and Bacterial SecD and SecF Homologs Exhibit Striking Structural and Functional Conservation'. *Journal of Bacteriology* 188 (4): 1251–59. https://doi.org/10.1128/JB.188.4.1251-

1259.2006/ASSET/4757EE22-0FB5-450E-A7E8-21C33DE75A8A/ASSETS/GRAPHIC/ZJB0040654590006.JPEG.

Hirano, Tatsuya. 2002. 'The ABCs of SMC Proteins: Two-Armed ATPases for Chromosome Condensation, Cohesion, and Repair'. *Genes & Development* 16 (4): 399–414. https://doi.org/10.1101/GAD.955102.

Hollywood, Katherine, Daniel R. Brison, and Royston Goodacre. 2006. 'Metabolomics: Current Technologies and Future Trends'. *PROTEOMICS* 6 (17): 4716–23. https://doi.org/10.1002/pmic.200600106.

Hug, Laura A. 2018. 'Sizing Up the Uncultured Microbial Majority'. *mSystems* 3 (5): 10.1128/msystems.00185-18. https://doi.org/10.1128/msystems.00185-18.

Hug, Laura A, Brett J Baker, Karthik Anantharaman, Christopher T Brown, Alexander J Probst, Cindy J Castelle, Cristina N Butterfield, et al. 2016. 'A New View of the Tree of Life'. *Nature Microbiology* 1 (5): 1–6. https://doi.org/10.1038/nmicrobiol.2016.48.

Hunt, Terry L., and Carl P. Lipo. 2006. 'Late Colonization of Easter Island'. *Science* 311 (5767): 1603–6. https://doi.org/10.1126/science.1121879.

Imachi, Hiroyuki, Masaru K. Nobu, Nozomi Nakahara, Yuki Morono, Miyuki Ogawara, Yoshihiro Takaki, Yoshinori Takano, et al. 2020. 'Isolation of an Archaeon at the Prokaryote–Eukaryote Interface'. *Nature* 577 (7791): 519–25. https://doi.org/10.1038/s41586-019-1916-6.

Irwin, Nicholas A. T., Alexandros A. Pittis, Thomas A. Richards, and Patrick J. Keeling. 2022. 'Systematic Evaluation of Horizontal Gene Transfer between Eukaryotes and Viruses'. *Nature Microbiology* 7 (2): 327–36. https://doi.org/10.1038/s41564-021-01026-3.

Iyer, Lakshminarayan M., S. Balaji, Eugene V. Koonin, and L. Aravind. 2006. 'Evolutionary Genomics of Nucleo-Cytoplasmic Large DNA Viruses'. *Virus Research*, Comparative Genomics and Evolution of Complex Viruses, 117 (1): 156–84. https://doi.org/10.1016/j.virusres.2006.01.009.

Jachiet, Pierre-Alain. 2014. 'Étude de l'évolution Combinatoire Des Gènes Par l'analyse de Réseaux de Similarité de Séquence'. PhD Thesis, Université Pierre et Marie Curie-Paris VI. https://theses.hal.science/tel-01127379/.

Jachiet, Pierre-Alain, Romain Pogorelcnik, Anne Berry, Philippe Lopez, and Eric Bapteste. 2013. 'MosaicFinder: Identification of Fused Gene Families in Sequence Similarity Networks'. *Bioinformatics* 29 (7): 837–44. https://doi.org/10.1093/bioinformatics/btt049.

Jiao, Jian-Yu, Lan Liu, Zheng-Shuang Hua, Bao-Zhu Fang, En-Min Zhou, Nimaichand Salam, Brian P Hedlund, and Wen-Jun Li. 2021. 'Microbial Dark Matter Coming to Light: Challenges and Opportunities'. *National Science Review* 8 (3): nwaa280. https://doi.org/10.1093/nsr/nwaa280.

Jolley, Keith A., James E. Bray, and Martin C. J. Maiden. 2018. 'Open-Access Bacterial Population Genomics: BIGSdb Software, the PubMLST.Org Website and Their Applications'. *Wellcome Open Research* 3 (September):124. https://doi.org/10.12688/wellcomeopenres.14826.1.

Karlin, S, and S F Altschul. 1990. 'Methods for Assessing the Statistical Significance of Molecular Sequence Features by Using General Scoring Schemes.' *Proceedings of the National Academy of Sciences* 87 (6): 2264–68. https://doi.org/10.1073/pnas.87.6.2264.

Keeling, Patrick J. 2009. 'Chromalveolates and the Evolution of Plastids by Secondary Endosymbiosis'. *Journal of Eukaryotic Microbiology* 56 (1): 1–8. https://doi.org/10.1111/j.1550-7408.2008.00371.x.

Kersting, Anna R., Erich Bornberg-Bauer, Andrew D. Moore, and Sonja Grath. 2012. 'Dynamics and Adaptive Benefits of Protein Domain Emergence and Arrangements during Plant Genome Evolution'. *Genome Biology and Evolution* 4 (3): 316–29. https://doi.org/10.1093/gbe/evs004.

Kjær, Kurt H., Mikkel Winther Pedersen, Bianca De Sanctis, Binia De Cahsan, Thorfinn S. Korneliussen, Christian S. Michelsen, Karina K. Sand, et al. 2022. 'A 2-Million-Year-Old Ecosystem in Greenland Uncovered by Environmental DNA'. *Nature* 612 (7939): 283–91. https://doi.org/10.1038/s41586-022-05453-y.

Kleine, Tatjana, Uwe G. Maier, and Dario Leister. 2009. 'DNA Transfer from Organelles to the Nucleus: The Idiosyncratic Genetics of Endosymbiosis'. *Http://Dx.Doi.Org/10.1146/Annurev.Arplant.043008.092119* 60 (April):115–38. https://doi.org/10.1146/ANNUREV.ARPLANT.043008.092119.

Knoll, Andrew H. 2011. 'The Multiple Origins of Complex Multicellularity'. *Annual Review of Earth and Planetary Sciences* 39 (Volume 39, 2011): 217–39. https://doi.org/10.1146/annurev.earth.031208.100209.

Koch, Robert. 1877. 'Untersuchungen Uber Bakterien V. Die Aetiologie Der Milzbrand-Krankheit, Begrunder Auf Die Entwicklungegeschichte Bacillus Anthracis'. *Beitrage Zur Biologie Der Pflanzen* 2 (2): 277–310.

Koonin, Eugene V. 2012. *The Logic of Chance: The Nature and Origin of Biological Evolution*. Upper Saddle River, N.J: Pearson Education.

Koonin, Eugene V., and Natalya Yutin. 2019. 'Chapter Five - Evolution of the Large Nucleocytoplasmic DNA Viruses of Eukaryotes and Convergent Origins of Viral Gigantism'. In *Advances in Virus Research*, edited by Margaret Kielian, Thomas C. Mettenleiter, and Marilyn J. Roossinck, 103:167–202. Academic Press. https://doi.org/10.1016/bs.aivir.2018.09.002.

Kummerfeld, Sarah K., and Sarah A. Teichmann. 2005. 'Relative Rates of Gene Fusion and Fission in Multi-Domain Proteins'. *Trends in Genetics* 21 (1): 25–30. https://doi.org/10.1016/j.tig.2004.11.007.

Lannes, Romain. 2019. 'Recherche de séquences environnementales inconnues d'intérêt médical/biologique par l'utilisation de grands réseaux de similarité de séquences'. Phdthesis, Sorbonne Université. https://theses.hal.science/tel-02954131.

Leonard, Guy, and Thomas A. Richards. 2012. 'Genome-Scale Comparative Analysis of Gene Fusions, Gene Fissions, and the Fungal Tree of Life'. *Proceedings of the National Academy of Sciences* 109 (52): 21402–7. https://doi.org/10.1073/pnas.1210909110.

Liu, Mingyi, and Andrei Grigoriev. 2004. 'Protein Domains Correlate Strongly with Exons in Multiple Eukaryotic Genomes – Evidence of Exon Shuffling?' *Trends in Genetics* 20 (9): 399–403. https://doi.org/10.1016/j.tig.2004.06.013.

Lloyd, Karen G., Andrew D. Steen, Joshua Ladau, Junqi Yin, and Lonnie Crosby. 2018. 'Phylogenetically Novel Uncultured Microbial Cells Dominate Earth Microbiomes'. *mSystems* 3 (5): 10.1128/msystems.00055-18. https://doi.org/10.1128/msystems.00055-18.

Lobb, Briallen, Daniel A. Kurtz, Gabriel Moreno-Hagelsieb, and Andrew C. Doxey. 2015. 'Remote Homology and the Functions of Metagenomic Dark Matter'. *Frontiers in Genetics* 6 (July). https://doi.org/10.3389/fgene.2015.00234.

LoDuca, S. T., N. Bykova, M. Wu, S. Xiao, and Y. Zhao. 2017. 'Seaweed Morphology and Ecology during the Great Animal Diversification Events of the Early Paleozoic: A Tale of Two Floras'. *Geobiology* 15 (4): 588–616. https://doi.org/10.1111/gbi.12244.

Lopez, Philippe, Sébastien Halary, and Eric Bapteste. 2015. 'Highly Divergent Ancient Gene Families in Metagenomic Samples Are Compatible with Additional Divisions of Life'. *Biology Direct* 10 (1): 64. https://doi.org/10.1186/s13062-015-0092-3.

López-García, Purificación, and David Moreira. 2021. 'Physical Connections: Prokaryotes Parasitizing Their Kin'. *Environmental Microbiology Reports* 13 (1): 54–61. https://doi.org/10.1111/1758-2229.12910.

Lowe, Rohan, Neil Shirley, Mark Bleackley, Stephen Dolan, and Thomas Shafee. 2017. 'Transcriptomics Technologies'. *PLOS Computational Biology* 13 (5): e1005457. https://doi.org/10.1371/journal.pcbi.1005457.

Marcy, Yann, Cleber Ouverney, Elisabeth M. Bik, Tina Lösekann, Natalia Ivanova, Hector Garcia Martin, Ernest Szeto, et al. 2007. 'Dissecting Biological "Dark Matter" with Single-Cell Genetic Analysis of Rare and Uncultivated TM7 Microbes from the Human Mouth'. *Proceedings of the National Academy of Sciences of the United States of America* 104 (29): 11889–94. https://doi.org/10.1073/pnas.0704662104.

Marsh, Joseph A., and Sarah A. Teichmann. 2010. 'How Do Proteins Gain New Domains?' *Genome Biology* 11 (7): 126. https://doi.org/10.1186/gb-2010-11-7-126.

Mazéas, Lisa, Rina Yonamine, Tristan Barbeyron, Bernard Henrissat, Elodie Drula, Nicolas Terrapon, Chikako Nagasato, and Cécile Hervé. 2023. 'Assembly and Synthesis of the Extracellular Matrix in Brown Algae'. *Seminars in Cell & Developmental Biology*, Special Issue: Algal model organisms by Susana Coelho and Olivier de Clerck, 134 (January):112–24. https://doi.org/10.1016/j.semcdb.2022.03.005.

Mistry, Jaina, Sara Chuguransky, Lowri Williams, Matloob Qureshi, Gustavo A Salazar, Erik L L Sonnhammer, Silvio C E Tosatto, et al. 2021. 'Pfam: The Protein Families Database in 2021'. *Nucleic Acids Research* 49 (D1): D412–19. https://doi.org/10.1093/nar/gkaa913.

Mitelman, Felix, Bertil Johansson, and Fredrik Mertens. 2004. 'Fusion Genes and Rearranged Genes as a Linear Function of Chromosome Aberrations in Cancer'. *Nature Genetics* 36 (4): 331–34. https://doi.org/10.1038/ng1335.

Muller, H. J. 1932. 'Some Genetic Aspects of Sex'. *The American Naturalist* 66 (703): 118–38. https://doi.org/10.1086/280418.

Nagy, László G., Gábor M. Kovács, and Krisztina Krizsán. 2018. 'Complex Multicellularity in Fungi: Evolutionary Convergence, Single Origin, or Both?' *Biological Reviews* 93 (4): 1778–94. https://doi.org/10.1111/brv.12418.

Nayfach, Stephen, Simon Roux, Rekha Seshadri, Daniel Udwary, Neha Varghese, Frederik Schulz, Dongying Wu, et al. 2020. 'A Genomic Catalog of Earth's Microbiomes'. *Nature Biotechnology* 39 (4): 499–509. https://doi.org/10.1038/s41587-020-0718-6.

Needleman, Saul B., and Christian D. Wunsch. 1970. 'A General Method Applicable to the Search for Similarities in the Amino Acid Sequence of Two Proteins'. *Journal of Molecular Biology* 48 (3): 443–53. https://doi.org/10.1016/0022-2836(70)90057-4.

Newman, M. E. J., S. H. Strogatz, and D. J. Watts. 2001. 'Random Graphs with Arbitrary Degree Distributions and Their Applications'. *Physical Review E* 64 (2): 026118. https://doi.org/10.1103/PhysRevE.64.026118.

Ocaña-Pallarès, Eduard, Tom A. Williams, David López-Escardó, Alicia S. Arroyo, Jananan S. Pathmanathan, Eric Bapteste, Denis V. Tikhonenkov, Patrick J. Keeling, Gergely J. Szöllősi, and Iñaki Ruiz-Trillo. 2022. 'Divergent Genomic Trajectories Predate the Origin of Animals and Fungi'. *Nature* 609 (7928): 747–53. https://doi.org/10.1038/s41586-022-05110-4.

Ohno, Susumu, Ulrich Wolf, and Niels B. Atkin. 1968. 'Evolution from Fish to Mammals by Gene Duplication'. *Hereditas* 59 (1): 169–87. https://doi.org/10.1111/j.1601-5223.1968.tb02169.x.

Padalko, Anastasiia, Govind Nair, and Filipa L. Sousa. 2024. 'Fusion/Fission Protein Family Identification in Archaea'. *mSystems* 9 (6): e00948-23. https://doi.org/10.1128/msystems.00948-23.

Pasek, Sophie, Jean-Loup Risler, and Pierre Brézellec. 2006. 'Gene Fusion/Fission Is a Major Contributor to Evolution of Multi-Domain Bacterial Proteins'. *Bioinformatics* 22 (12): 1418–23. https://doi.org/10.1093/bioinformatics/btl135.

Pathmanathan, Jananan Sylvestre, Philippe Lopez, François-Joseph Lapointe, and Eric Bapteste. 2018. 'CompositeSearch: A Generalized Network Approach for Composite Gene Families Detection'. *Molecular Biology and Evolution* 35 (1): 252–55. https://doi.org/10.1093/molbev/msx283.

Pavlopoulos, Georgios A., Fotis A. Baltoumas, Sirui Liu, Oguz Selvitopi, Antonio Pedro Camargo, Stephen Nayfach, Ariful Azad, et al. 2023. 'Unraveling the Functional Dark Matter through Global Metagenomics'. *Nature* 622 (7983): 594–602. https://doi.org/10.1038/s41586-023-06583-7.

Paysan-Lafosse, Typhaine, Matthias Blum, Sara Chuguransky, Tiago Grego, Beatriz Lázaro Pinto, Gustavo A Salazar, Maxwell L Bileschi, et al. 2023. 'InterPro in 2022'. *Nucleic Acids Research* 51 (D1): D418–27. https://doi.org/10.1093/nar/gkac993.

Pogliano, Joseph A, and Jon Beckwith. 1994. 'SecD and SecF Facilitate Protein Export in Escherichia Coli'. *The EMBO Journal* 13 (3): 554–61.

Quast, Christian, Elmar Pruesse, Pelin Yilmaz, Jan Gerken, Timmy Schweer, Pablo Yarza, Jörg Peplies, and Frank Oliver Glöckner. 2013. 'The SILVA Ribosomal RNA Gene Database Project: Improved Data Processing and Web-Based Tools'. *Nucleic Acids Research* 41 (D1): D590–96. https://doi.org/10.1093/nar/gks1219.

Remmert, Michael, Andreas Biegert, Andreas Hauser, and Johannes Söding. 2012. 'HHblits: Lightning-Fast Iterative Protein Sequence Searching by HMM-HMM Alignment'. *Nature Methods* 9 (2): 173–75. https://doi.org/10.1038/nmeth.1818.

Rinke, Christian, Patrick Schwientek, Alexander Sczyrba, Natalia N Ivanova, Iain J Anderson, Jan-Fang Cheng, Aaron Darling, et al. 2013. 'Insights into the Phylogeny and Coding Potential of Microbial Dark Matter'. *Nature* 499 (7459): 431–37. https://doi.org/10.1038/nature12352.

Rodrigues-Oliveira, Thiago, Florian Wollweber, Rafael I. Ponce-Toledo, Jingwei Xu, Simon K.-M. R. Rittmann, Andreas Klingl, Martin Pilhofer, and Christa Schleper. 2023. 'Actin Cytoskeleton and Complex Cell Architecture in an Asgard Archaeon'. *Nature* 613 (7943): 332–39. https://doi.org/10.1038/s41586-022-05550-y.

Rogozin, Igor B., Yuri I. Wolf, Vladimir N. Babenko, and Eugene V. Koonin. 2006. 'Dollo Parsimony and the Reconstruction of Genome Evolution'. In *Parsimony, Phylogeny, and Genomics*, edited by Victor A. Albert, 0. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199297306.003.0011.

Rost, Burkhard. 1999. 'Twilight Zone of Protein Sequence Alignments'. *Protein Engineering, Design and Selection* 12 (2): 85–94. https://doi.org/10.1093/protein/12.2.85.

Sacerdot, Christine, Alexandra Louis, Céline Bon, Camille Berthelot, and Hugues Roest Crollius. 2018. 'Chromosome Evolution at the Origin of the Ancestral Vertebrate Genome'. *Genome Biology* 19 (1): 166. https://doi.org/10.1186/s13059-018-1559-1.

Schäffer, Alejandro A., L. Aravind, Thomas L. Madden, Sergei Shavirin, John L. Spouge, Yuri I. Wolf, Eugene V. Koonin, and Stephen F. Altschul. 2001. 'Improving the Accuracy of PSI-BLAST Protein Database Searches with Composition-Based Statistics and Other Refinements'. *Nucleic Acids Research* 29 (14): 2994–3005. https://doi.org/10.1093/nar/29.14.2994.

Schulz, Frederik, Natalya Yutin, Natalia N. Ivanova, Davi R. Ortega, Tae Kwon Lee, Julia Vierheilig, Holger Daims, et al. 2017. 'Giant Viruses with an Expanded Complement of Translation System Components'. *Science* 356 (6333): 82–85. https://doi.org/10.1126/science.aal4657.

Sebé-Pedrós, Arnau, Bernard M. Degnan, and Iñaki Ruiz-Trillo. 2017. 'The Origin of Metazoa: A Unicellular Perspective'. *Nature Reviews Genetics* 18 (8): 498–512. https://doi.org/10.1038/nrg.2017.21.

Shakhnovich, Boris E., and Eugene V. Koonin. 2006. 'Origins and Impact of Constraints in Evolution of Gene Families'. *Genome Research* 16 (12): 1529–36. https://doi.org/10.1101/gr.5346206.

Simakov, Oleg, Ferdinand Marlétaz, Jia-Xing Yue, Brendan O'Connell, Jerry Jenkins, Alexander Brandt, Robert Calef, et al. 2020. 'Deeply Conserved Synteny Resolves Early Events in Vertebrate Evolution'. *Nature Ecology & Evolution* 4 (6): 820–30. https://doi.org/10.1038/s41559-020-1156-z.

Smith, T. F., and M. S. Waterman. 1981. 'Identification of Common Molecular Subsequences'. *Journal of Molecular Biology* 147 (1): 195–97. https://doi.org/10.1016/0022-2836(81)90087-5.

Söding, Johannes. 2005. 'Protein Homology Detection by HMM–HMM Comparison'. *Bioinformatics* 21 (7): 951–60. https://doi.org/10.1093/bioinformatics/bti125.

Soucy, Shannon M., Jinling Huang, and Johann Peter Gogarten. 2015. 'Horizontal Gene Transfer: Building the Web of Life'. *Nature Reviews Genetics* 16 (8): 472–82. https://doi.org/10.1038/nrg3962.

Spang, Anja, Jimmy H Saw, Steffen L Jørgensen, Katarzyna Zaremba-Niedzwiedzka, Joran Martijn, Anders E Lind, Roel van Eijk, Christa Schleper, Lionel Guy, and Thijs J G Ettema. 2015. 'Complex Archaea That Bridge the Gap between Prokaryotes and Eukaryotes'. *Nature* 521 (7551): 173–79. https://doi.org/10.1038/nature14447.

Staley, James T, and Allan Konopka. 1985. 'Measurement of in Situ Activities of Nonphotosynthetic Microorganisms in Aquatic and Terrestrial Habitats'. *Annual Review of Microbiology* 39 (1): 321–46.

Stricker, Stefan H., Anna Köferle, and Stephan Beck. 2017. 'From Profiles to Function in Epigenomics'. *Nature Reviews Genetics* 18 (1): 51–66. https://doi.org/10.1038/nrg.2016.138.

Sunagawa, Shinichi, Luis Pedro Coelho, Samuel Chaffron, Jens Roat Kultima, Karine Labadie, Guillem Salazar, Bardya Djahanschiri, et al. 2015. 'Structure and Function of the Global Ocean Microbiome'. *Science* 348 (6237): 1261359. https://doi.org/10.1126/science.1261359.

Sung, Patrick, and Hannah Klein. 2006. 'Mechanism of Homologous Recombination: Mediators and Helicases Take on Regulatory Functions'. *Nature Reviews Molecular Cell Biology* 7 (10): 739–50. https://doi.org/10.1038/nrm2008.

The UniProt Consortium. 2023. 'UniProt: The Universal Protein Knowledgebase in 2023'. *Nucleic Acids Research* 51 (D1): D523–31. https://doi.org/10.1093/nar/gkac1052.

Tokuriki, Nobuhiko, and Dan S Tawfik. 2009. 'Stability Effects of Mutations and Protein Evolvability'. *Current Opinion in Structural Biology*, Carbohydradtes and glycoconjugates / Biophysical methods, 19 (5): 596–604. https://doi.org/10.1016/j.sbi.2009.08.003.

Vanni, Chiara, Matthew S. Schechter, Silvia G. Acinas, Albert Barberán, Pier Luigi Buttigieg, Emilio O. Casamayor, Tom O. Delmont, et al. 2022. 'Unifying the Known and Unknown Microbial Coding Sequence Space'. *eLife* 11 (March). https://doi.org/10.7554/ELIFE.67667.

Vos, Michiel. 2009. 'Why Do Bacteria Engage in Homologous Recombination?' *Trends in Microbiology* 17 (6): 226–32. https://doi.org/10.1016/j.tim.2009.03.001.

Watson, Andrew K, Romain Lannes, Jananan S Pathmanathan, Raphaël Méheust, Slim Karkar, Philippe Colson, Eduardo Corel, Philippe Lopez, and Eric Bapteste. 2019. 'The Methodology Behind Network Thinking: Graphs to Analyze Microbial Complexity and Evolution'. In *Evolutionary Genomics: Statistical and Computational Methods*, edited by Maria Anisimova, 271–308. Methods in Molecular Biology. New York, NY: Springer. https://doi.org/10.1007/978-1-4939-9074-0_9.

Wensel, Caroline R., Jennifer L. Pluznick, Steven L. Salzberg, and Cynthia L. Sears. 2022. 'Next-Generation Sequencing: Insights to Advance Clinical Investigations of the Microbiome'. *The Journal of Clinical Investigation* 132 (7). https://doi.org/10.1172/JCI154944.

Whitman, W B, D C Coleman, and W J Wiebe. 1998. 'Prokaryotes: The Unseen Majority'. *Proceedings of the National Academy of Sciences* 95 (12): 6578–83. https://doi.org/10.1073/pnas.95.12.6578.

Willerslev, Eske, Anders J. Hansen, Regin Rønn, Tina B. Brand, Ian Barnes, Carsten Wiuf, David Gilichinsky, David Mitchell, and Alan Cooper. 2004. 'Long-Term Persistence of Bacterial DNA'. *Current Biology* 14 (1): R9–10. https://doi.org/10.1016/j.cub.2003.12.012.

Woese, Carl R., and George E. Fox. 1977. 'Phylogenetic Structure of the Prokaryotic Domain: The Primary Kingdoms'. *Proceedings of the National Academy of Sciences* 74 (11): 5088–90. https://doi.org/10.1073/pnas.74.11.5088.

Wolf, Yuri I., Pavel S. Novichkov, Georgy P. Karev, Eugene V. Koonin, and David J. Lipman. 2009. 'The Universal Distribution of Evolutionary Rates of Genes and Distinct Characteristics of Eukaryotic Genes of Different Apparent Ages'. *Proceedings of the National Academy of Sciences* 106 (18): 7273–80. https://doi.org/10.1073/pnas.0901808106.

Wu, Dongying, Martin Wu, Aaron Halpern, Douglas B. Rusch, Shibu Yooseph, Marvin Frazier, J. Craig Venter, and Jonathan A. Eisen. 2011. 'Stalking the Fourth Domain in Metagenomic Data: Searching for, Discovering, and Interpreting Novel, Deep Branches in Marker Gene Phylogenetic Trees'. *PLOS ONE* 6 (3): e18011. https://doi.org/10.1371/journal.pone.0018011.

Wyman, Stacia K., Aram Avila-Herrera, Stephen Nayfach, and Katherine S. Pollard. 2018. 'A Most Wanted List of Conserved Microbial Protein Families with No Known Domains'. *PLOS ONE* 13 (10): e0205749. https://doi.org/10.1371/journal.pone.0205749.

Yanai, Itai, Adnan Derti, and Charles DeLisi. 2001. 'Genes Linked by Fusion Events Are Generally of the Same Functional Category: A Systematic Analysis of 30 Microbial Genomes'. *Proceedings of the National Academy of Sciences* 98 (14): 7940–45. https://doi.org/10.1073/pnas.141236298.

Yang, Bo, Yong Wang, and Pei-Yuan Qian. 2016. 'Sensitivity and Correlation of Hypervariable Regions in 16S rRNA Genes in Phylogenetic Analysis'. *BMC Bioinformatics* 17 (1): 135. https://doi.org/10.1186/s12859-016-0992-y.

Zickler, Denise, and Nancy Kleckner. 2015. 'Recombination, Pairing, and Synapsis of Homologs during Meiosis'. *Cold Spring Harbor Perspectives in Biology* 7 (6): a016626. https://doi.org/10.1101/cshperspect.a016626.

Zmasek, Christian M., and Adam Godzik. 2012. 'This Déjà Vu Feeling—Analysis of Multidomain Protein Evolution in Eukaryotic Genomes'. *PLOS Computational Biology* 8 (11): e1002701. https://doi.org/10.1371/journal.pcbi.1002701.

Zou, Quan, Gang Lin, Xingpeng Jiang, Xiangrong Liu, and Xiangxiang Zeng. 2020. 'Sequence Clustering in Bioinformatics: An Empirical Study'. *Briefings in Bioinformatics* 21 (1): 1–10. https://doi.org/10.1093/bib/bby090.

Chapter VI. Appendix

Draft article – SHIFT: Sequence Homology Iterative Finding Tool for remote homology detection

SHIFT: Sequence Homology Iterative Finding Tool for remote homology detection

E. Corel, R. Lannes, D. Sussfeld, P. Lopez, E. Bapteste

28 juin 2024

Introduction

Detecting remote homology between related sequences is a challenging bioinformatic task. Established sequence similarity detection tools (such as the classical BLAST, but also more recent tools like DIAMOND, or MMSEQ2) display a good behaviour above a percentage of identity around 25% to 30% (the so-called "twilight zone"), below which the signal is overwhelmed by spurious similarities, in such a way that it becomes impossible to tell weak *bona fide* similarities from others that contain a small more highly conserved region.

Avoiding this problem can be tackled by repeatedly applying the similarity detection tool to the newly detected homologous sequences. In this way, the signal is propagated (as along a transitive closure), and a certain amount of similarity can be expected to hold.

The main existing tool for such a task is the software PSI-BLAST [Altschul et al., 1997], which iteratively updates a position-site-specific matrix (PSSM), and uses it to detect sequences having a weaker similarity with the initial sequence. However, the performance (especially the time requires) of PSI-BLAST is a limit to its use for very large datasets, typically for the search of divergent homologs in metagenomic data.

In this paper, we propose an iterative sequence similarity detection tool that runs around 10 to 50 times faster than PSI-BLAST, while retaining a comparable level of sensitivity.

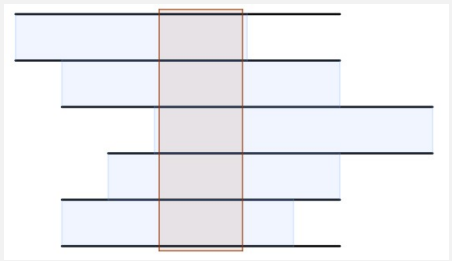
1 Material and Methods

The SHIFT software suite presented here is based on an iterative sequence similarity detection procedure, that uses fast and reliable programs to detect pairwise similarities, and repeatedly applies it to a set of extended sequences.

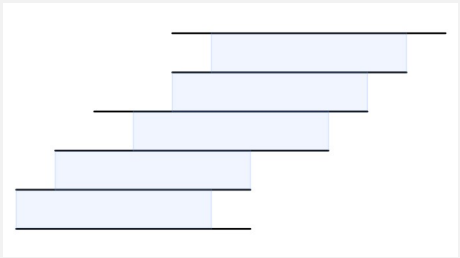
Detecting remote similarities by iteration can be marred by two kinds of problems : if we restrict the search to similar regions, the resulting similarity can rapidly "shrink" along with the iterations, and ends up in a strong but very narrow signal (a pervasive domain, like a Zn-finger for instance). On the contrary, if we extend the search and add up the previous regions found, the detected similarities can grow laterally in a "staircase-like" fashion, resulting in a complete lack of homology between the start and end sequences (Figure 2).

In designing SHIFT, we have tried to avoid both pitfalls, by including only similarities spanning a sufficient reciprocal cover (typically above 80%), for all retained sequences *i.e.* not only the direct similarities, but also the indirectly inferred ones.

FIGURE 1 – Coverage criterion : an alignment (displaying over 30% of identity) A-B (in blue) is retained if it moreover covers at least 80% of both sequences.



(a) "Finger" effect : strong but narrow signal detected.



(b) "Staircase" effect : no similarity between reference and final target sequences.

FIGURE 2 – Schematic representation of the possible pitfalls of the iterative alignment scheme.

1.1 Algorithm

1.1.1 General principle

Consider a reference input set $\mathcal{Q}$ of finite *query* protein or DNA sequences. The dataset of search sequences $\mathfrak{S}$ (typically consisting of a very large number of sequences), is queried using an existing established tool (in our case, we implemented BLAST, DIAMOND and MMSEQ2). Both sets of sequences are updated at each round of the similarity detection algorithm.

Current input sequences are screened against the current search dataset. Sequences in the search dataset are retained if three simultaneous criteria are met with respect to some previously found sequence.

S : a sequence similarity of over 30% is detected between both sequences,

C : the reported sequence similarity covers at least 80% of each sequence,

T : the coverage condition can be traced back to the reference set (*transitivity condition*).

The retained sequences $\tilde{\mathfrak{S}}$ are added to the query set and removed from the search set. In this way, no similarities are searched for between sequences retained at a given round of the algorithm.

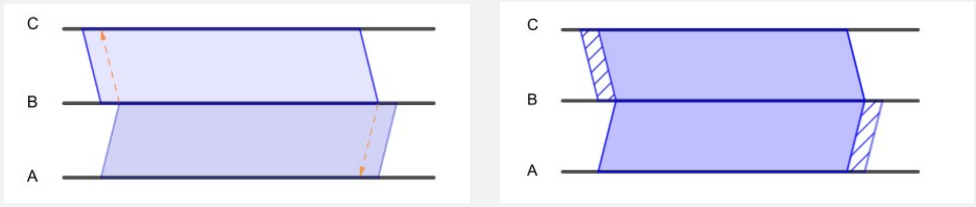
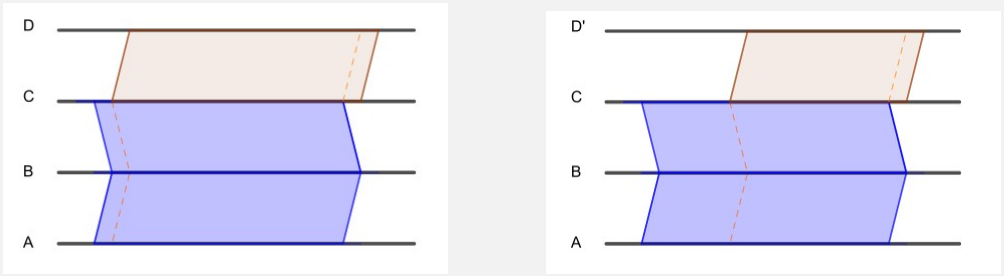


FIGURE 3 – Coverage criterion 2 : the intersection over B of the alignments A-B and B-C is projected back on A and C. The restriction gives an implied indirect alignment A-C. If coverage is sufficient for all three alignments, alignment is retained as a path A-B-C.



(a) C-D passes the transitivity criterion : sequence D is retained, and the path A-B-C is extended to D.

(b) Transitive coverage is not sufficient : D' is discarded.

FIGURE 4 – Coverage criterion 3 : Iterative step – a new alignment is added.

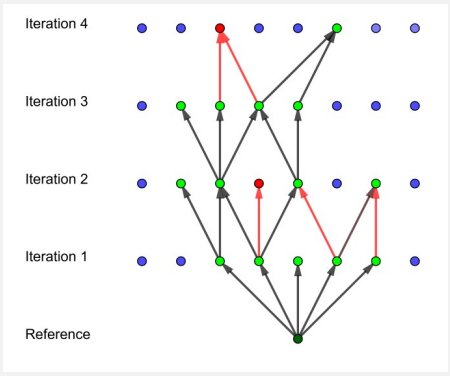
The search is repeated from the new query set $\mathcal{Q} \cup \tilde{\mathfrak{S}}$ to the new search set $\mathfrak{S} \backslash \tilde{\mathfrak{S}}$, until no more new sequences are found.

The first round is an ordinary BLAST-like search : we filter the output with criteria **C** and **S**, keep the selected sequences as our new query set $\mathcal{Q}$, and remove them from the search space $\mathfrak{S}$.

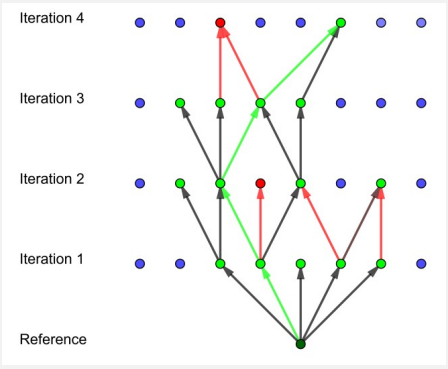
In the next rounds, in order to avoid both pitfalls described in the introduction, we ask that the similarity extends over at least the chosen cover percentage for all sequences. Alignments in the next rounds are therefore kept if moreover all the *implied* alignments satisfy the coverage criterion (criterion **T**). In this way we expect to find similarities that go below the threshold of 30%, while corresponding to actual homologies.

1.1.2 Description of the algorithm

We can model the algorithm with a *similarity* graph $G = (V, E)$, with $V = \mathcal{Q} \cup \mathfrak{S}$, and where an edge is drawn between two sequences s and s' if the criteria **C** and **S** are directly fulfilled. Our criterion can be formulated as the existence of a path $s_0, s_1, \ldots, s_n$ from a *reference sequence* s_0 to s_n, such that



(a) Red arrows denote matches that do not pass the transitive criterion for any path reaching their source vertex.



(b) A green terminal vertex is accepted if there exists an admissible path (in green). There can be several admissible paths ending at the same vertex.

FIGURE 5 – Iterations in the similarity graph : starting from a reference sequence (in dark green), with the target set consisting of blue vertices, arrows denote a match satisfying criteria **S** and **C**. Green vertices are retained by the algorithm, while red ones are found but not retained. Green terminal vertices correspond to accepted distant homologs that do not return significant matches with the remainder of the search space.

— (s_i, s_{i+1}) is an edge

— the implied alignment of any pair of sequences s_i and s_j in the path satisfies criterion **C**.

In a more inductive fashion, assume that s is an accepted sequence. A further match from s to s' will lead to accepting sequence s' if (s, s') extends at least one path ending in s (see Figure 5).

Segments and matches. The notions used here are pretty straightforward when only ungapped alignments are considered. However, these definitions can become rather involved when using gapped alignments. We will therefore start with the ungapped case, while being at the same time slightly more formal than necessary to allow for a precise description of the general case.

A *segment* $S = (s, b, e)$ in a sequence s, of length $|s|$, is the collection of consecutive positions from b to e in sequence s, where $1 \leqslant b \leqslant e \leqslant |s|$. The *length* $\ell(S)$ of a segment $S = (s, b, e)$ is naturally defined as $\ell(S) = e - b + 1$. Two segments $S = (s, b, e)$ and $\tilde{S} = (s, \tilde{b}, \tilde{e})$ of the same sequence *intersect* as

$$S \cap \tilde{S} = \begin{cases} (s, b^*, e^*) \text{ if } b^* \overset{\text{def.}}{=} \max(b, \tilde{b}) \leqslant \min(e, \tilde{e}) \overset{\text{def.}}{=} e^* \\ \varnothing \text{ otherwise.} \end{cases}$$

A *match* is a pairwise alignment between two sequences. More precisely, it is a pair $M = (S, T, \mu)$ of segments $S = (s, b, e)$ and $T = (t, \beta, \varepsilon)$, together with a totally ordered subset μ of pairs of *distinct* matching positions in s and t, starting with the pair (b, β) and ending with (e, ε). Sequence s is the *source* $\sigma(M)$ and t the *target* $\tau(M)$ of the match M. If the pairwise alignment is ungapped, this set of matching positions will be omitted, since it is simply the set of (pairs of) consecutive positions.

In general however, the segments only have the same length when gaps are included. The actual position of the gaps is determined by the pairwise alignment software, but is essentially irrelevant for the present discussion (Figure 7).

Definition 1. Let μ be a totally ordered set of pairs of distinct positions in sequences s and t

$$\mu = \{((s, p_1), (t, q_1)), \ldots, ((s, p_\ell), (t, q_\ell))\} \text{ with } p_1 < \cdots < p_\ell \text{ and } q_1 < \cdots < q_\ell.$$

1. The sets of *matching positions* on s and t are defined as

$$\mu(s) = \{(s, p_1), \ldots, (s, p_\ell)\} \text{ and } \mu(t) = \{(t, q_1), \ldots, (t, q_\ell)\}.$$

2. The *match induced by* μ is defined as

$$\Xi(\mu) = (S, T, \mu) \text{ where } S = (s, p_1, p_\ell) \text{ and } T = (t, q_1, q_\ell).$$

We say that $S = \overline{\mu}(s)$ and $T = \overline{\mu}(t)$ are the *segments induced by* μ on s and t.

Consistently with the accepted usage, the complete segments S and T are deemed to be aligned, although indels themselves are by definition not aligned with anything, However, flanking indels should not be included.

We need to be a little careful when restricting gapped alignments to a given segment. Distinguishing between *matching* and *non-matching* positions, we say that the segment $I = (s, i, j)$ in sequence s *restricts a match* M to $M \cap I = (S^*, T^*, \mu^*)$, where the pair of segments $S^* = (s, b^*, e^*)$ and $T^* = (t, \beta^*, \varepsilon^*)$ are defined as follows (see Figure 6) :

- β^* is the position in t that matches the leftmost matching position b^* in s *to the right* of b and i,
- ε^* is the position in t that matches the rightmost matching position e^* in s *to the left* of e and j.

The set of matching positions μ^* is defined as the matching positions that are included in the restriction. The resulting match $M \cap I$ does not start or end with gaps, as required. The same definition holds for the restriction of M to a segment J from sequence t.

	6				7	8	9	10	11	12				13	14	15	16	17	18
t	A	-	-	-	A	A	M	S	G	R	-	-	-	T	H	R	A	D	H
s	S	S	K	G	-	-	N	H	G	E	Y	V	G	R	-	-	V	D	H
	26	27	28	29			30	31	32	33	34	35	36	37			38	39	40

FIGURE 6 – Restriction to the interval $I = (s, 27, 38)$ of the match $M = (S, T, \mu)$ where $S = (s, 26, 40)$ and $T = (t, 6, 18)$. The leftmost matching position on s to the right of $b = 27$ is $b^* = 30$ and so we get $\beta^* = 9$. Similarly, we have $e^* = 38$ and thus $\varepsilon^* = 16$. The restricted match $M \cap I$ is in blue, as flanking indels have been excluded.

Admissible paths and stacks of alignments. A path in the similarity graph (Figure 5) corresponds to a sequence of consecutive alignments (or *stack*) $\mathcal{M} = (M_0, M_1, \ldots, M_n)$ such that

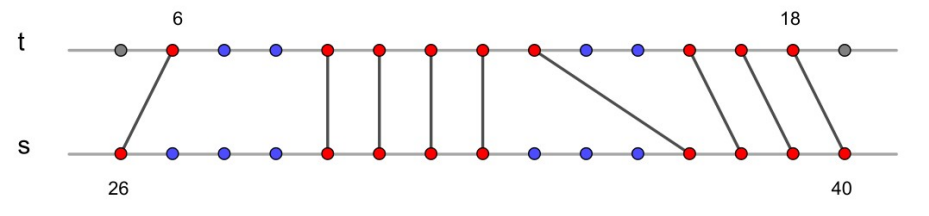


FIGURE 7 – Abstraction of the pairwise alignment from Figure 6. Dots represent positions in sequences s, t, and edges connecting them, the matching set μ. Gaps are implicitly represented by missing edges. Red dots correspond to matching positions in the sequences. Blue dots correspond to unaligned positions that are included in the aligned segments $S = \overline{\mu}(s)$ and $T = \overline{\mu}(t)$, and gray dots are unaligned positions outside the segments.

$\sigma(M_{i+1}) = \tau(M_i)$, *i.e.* the target of any match is equal to the source of the next (without passing twice on the same sequence). We first investigate when a fixed stack corresponds to an admissible path in the similarity graph.

Putting $s_i = \sigma(M_i)$ and $s_{n+1} = \tau(M_n)$, the *support* $\mathfrak{S}(\mathcal{M})$ of the stack $\mathcal{M}$ is the ordered collection of sequences $(s_0, \dots, s_{n+1})$ aligned by $\mathcal{M}$. It is convenient to introduce another graph to model the stack.

Definition 2. Let $\mathcal{M} = (M_0, \dots, M_n)$ be a stack of alignments, where $M_i = (S_i, T_i, \mu_i)$. The set of *sites* in the support $\mathfrak{S}(\mathcal{M})$ of $\mathcal{M}$ is defined as

$$\Sigma(\mathcal{M}) = \{(s,p) \in \mathfrak{S}(\mathcal{M}) \times \mathbb{N} \mid 1 \leqslant p \leqslant |s|\} \text{ partially ordered by } (s,p) \leqslant (s',p') \iff \begin{cases} s = s' \\ p \leqslant p'. \end{cases}$$

The *alignment graph* $G_{\mathcal{M}}$ of the stack $\mathcal{M}$ is $G_{\mathcal{M}} = \left(\Sigma(\mathcal{M}), \bigcup_{i=0}^{n} \mu_i\right)$.

In short, we define an edge for every pair of matching sites in μ_i, and consider the graph $G_{\mathcal{M}}$ defined by these edges. Stacks have a particularly simple structure.

Lemma 1. Let $\mathcal{M}$ be a stack, and let $G_{\mathcal{M}}$ be its alignment graph. Let $\Gamma_{\mathcal{M}}$ be the set of paths in $G_{\mathcal{M}}$. We say that such a path is *complete* if it extends from the first to the last sequence of the stack.

1. For every pair of sites σ, σ' in $\Sigma(\mathcal{M})$, there exists at most one path $\gamma \in \Gamma_{\mathcal{M}}$ relating them in $G_{\mathcal{M}}$.
2. The ordering $\leqslant$ in $\Sigma(\mathcal{M})$ induces a partial ordering on the set of connected components of $G_{\mathcal{M}}$ defined by

$$\gamma \leqslant \gamma' \iff \left(\sigma \leqslant \sigma' \text{ for any pair of comparable sites } \sigma \in \gamma \text{ and } \sigma' \in \gamma'\right).$$

   The ordering $\leqslant$ is total on the set of complete paths of $G_{\mathcal{M}}$.

Although we prove this statement, it should be clear from Figure 8.

Démonstration. With the notations of definition 2, by construction, all edges in $G_{\mathcal{M}}$ have the form $((s_i, p), (s_{i+1}, p'))$ with $0 \leqslant i \leqslant n$, and every vertex has degree at most 2. If there are two paths ending at $\sigma' = (s_{i+1}, p')$ the last edge of each path connects σ' with the same site (s_i, p), hence the last edges are equal. By induction, both paths are equal.

Additionally, if $e = ((s_i, p), (s_{i+1}, p'))$ and $e' = ((s_i, q), (s_{i+1}, q'))$, then (p, q) and (p', q') are in the same order. Therefore, we can prove the second statement by induction on n. We identify connected components in $G_{\mathcal{M}}$ with maximal paths in $\Gamma_{\mathcal{M}}$. The case $n = 0$ is the definition of a pairwise alignment $M = (S, T, \mu)$, for the set of paths is then equal to the set of matching positions μ, which is totally ordered. Assume that the result has been proved for all sequences of n alignments, and let $\mathcal{M} = (M_0, \ldots, M_n, M_{n+1})$. Let $\tilde{\mathcal{M}} = (M_0, \ldots, M_n)$ and define the *restriction* $\tilde{\gamma}$ of $\gamma \in \Gamma_{\mathcal{M}}$ to $\tilde{\mathcal{M}}$ in the obvious way. The path γ is either equal to its restriction $\tilde{\gamma}$, or is the extension of $\tilde{\gamma}$ by an edge (σ_n, σ_{n+1}). Let $\gamma, \gamma' \subset \Gamma_{\mathcal{M}}$, such that $\gamma \leqslant \gamma'$ and $\gamma' \leqslant \gamma$. By definition of the transitive closure, there exist paths γ_i and δ_j such that

$$\gamma \leqslant \gamma_1 \leqslant \cdots \leqslant \gamma_s \leqslant \gamma' \text{ and } \gamma' \leqslant \delta_1 \leqslant \cdots \leqslant \delta_t \leqslant \gamma$$

such that two consecutive paths have non-empty support intersection. The restrictions of these relations to $\tilde{M}$ and the induction hypothesis imply that $\tilde{\gamma} = \tilde{\gamma}'$. Therefore, either $\tilde{\gamma}$ and $\tilde{\gamma'}$ are maximal, and there is nothing to prove, or they are both extended by an edge. Assume that (σ_n, σ_{n+1}) and $(\sigma'_n, \sigma'_{n+1})$ be the extensions of $\tilde{\gamma}$ and $\tilde{\gamma}'$. The set of edges μ_n being totally ordered, we have then $\sigma_n = \sigma'_n$, and therefore $\gamma = \gamma'$. □

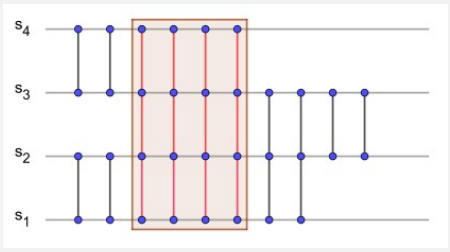
Lemma 1 means that a stack $\mathcal{M}$ defines a combinatorial (or topological) multiple sequence alignment [Corel et al., 2010] of the sequences $s_0, \ldots, s_{n+1}$: every connected component is a column (completed with gaps if needed), and every topological sorting of these connected components give rise to an MSA in its familiar matrix form. Therefore, we refer to complete paths in $\Gamma_{\mathcal{M}}$ as *anchor points* for $\mathcal{M}$.

Definition 3. Let $\mathcal{M}$ be a stack of alignments, and let $G_{\mathcal{M}}$ be the alignment graph of $\mathcal{M}$.

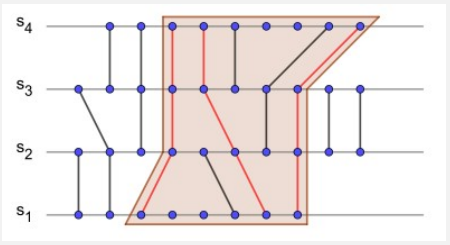
1. Two sites $\sigma = (s, p)$ and $\sigma' = (t, q)$ in $\Sigma(\mathcal{M})$ are *implicitly aligned by* $\mathcal{M}$ if there exists a path relating them in $G_{\mathcal{M}}$.
2. For two sequences $s, t \in \mathfrak{S}(\mathcal{M})$, the *alignment of s and t implied by* $\mathcal{M}$ is the match $\Xi(\mu_{s,t})$ where $\mu_{s,t}$ is the set of pairs of sites in s, t implicitly aligned by $\mathcal{M}$.
3. Let $s \in \mathfrak{S}(\mathcal{M})$. Let $S_t = \overline{\mu}_{s,t}(s)$ be the segment induced on s by $\Xi(\mu_{s,t})$. We define the *core segment on s* as $K_{\mathcal{M}}(s) = \bigcap_{t \in \mathfrak{S}(\mathcal{M})} S_t$

Lemma 2. Let $\mathcal{M}$ be a stack. For every match M in $\mathcal{M}$, between the sequences s and t, the restrictions $M \cap K_{\mathcal{M}}(s)$ and $M \cap K_{\mathcal{M}}(t)$ coincide.

Definition 4. Let $\mathcal{M} = (M_0, \ldots, M_n)$ be a stack with support $\mathfrak{S}(\mathcal{M}) = (s_0, \ldots, s_{n+1})$. The *core* of the stack $\mathcal{M}$ is the stack of alignments $K(\mathcal{M})$ obtained by restricting all matches in $\mathcal{M}$ to the core segments $K_{\mathcal{M}}(s_k)$ for $0 \leqslant k \leqslant n + 1$.



(a) Ungapped case : restrictions of implied alignments have all the same length.



(b) Gapped case : core is equivalently defined as the convex hull of leftmost and rightmost complete paths.

FIGURE 8 – Representation of the alignment graph of a stack. The core of the stack (in shaded brown) is defined by the restriction of all implied alignments. Blue dots are aligned positions in sequences $s_1, \ldots, s_4$. Red edges indicate complete paths.

The core of a stack can be empty. The key property of the core is that, for every sequence in the support, all the implied alignments coincide on the same segment. In particular, the core $K(\mathcal{M})$ can be written as

$$K(\mathcal{M}) = (M_0, \ldots, M_n) \text{ where } M_i = (S_i, S_{i+1}, \mu_i).$$

This notion is simple enough when considering only ungapped alignments (see Figure 8, left). However, its definition becomes straightforward in the gapped case only with the introduction of the alignment graph (as shown in Figure 8, right). The main result of this section is the following.

Proposition 1. The core $K(\mathcal{M})$ of a stack of alignments $\mathcal{M}$ is equal to the convex hull of all complete paths in the graph $G_{\mathcal{M}}$.

Representing the stack $\mathcal{M}$ as an MSA, the core represents the restriction of the alignment to the region limited by the rightmost and leftmost anchor points of $\mathcal{M}$. We are now ready to define when a stack corresponds to an admissible path.

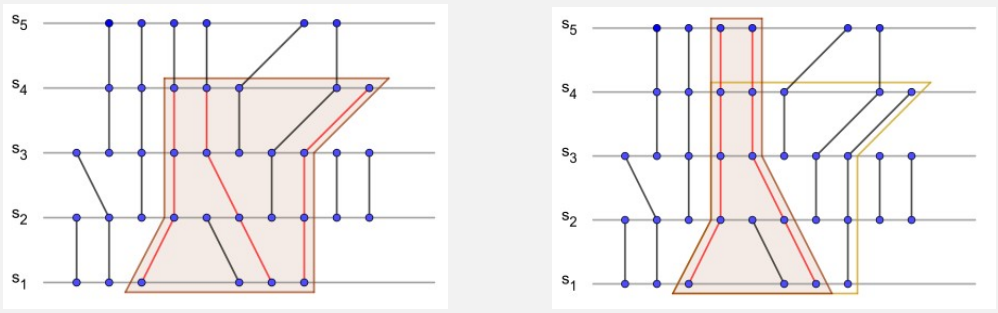
Definition 5. Let $\mathcal{M}$ be a stack of alignments with support $\mathfrak{S}(\mathcal{M}) = (s_0, \ldots, s_{n+1})$, and non-empty core $K(\mathcal{M}) = (M_0, \ldots, M_n)$, where $M_i = (S_i, S_{i+1}, \mu_i)$. The *coverage* of the stack is defined as the coverage of its core

$$\kappa(\mathcal{M}) = \min_{0 \leqslant i \leqslant n+1} \left(\frac{\ell(S_i)}{|s_i|} \right).$$

A stack is *admissible at threshold* τ if $\kappa(\mathcal{M}) \geqslant \tau$.

A stack of alignments corresponds to an admissible path if its core has a coverage of at least 0.8. In the ungapped case, all segments of the core have the same length, say ℓ, and the criterion becomes simply

$$\mathcal{M} \text{ admissible} \iff \ell \geqslant 0.8 \max_{0 \leqslant i \leqslant n+1} |s_i| \,. \tag{1}$$



(a) Edges from μ are added to the alignment graph.

(b) Complete paths from s_1 to s_5 define the core of the extended stack.

FIGURE 9 – Core of the extension of a stack : a match (S, T, μ) between s_4 and s_5 is added to the stack of alignments. The resulting core (in shaded brown) extends over more sequences but covers less positions per sequence. Unaligned positions are not shown for clarity.

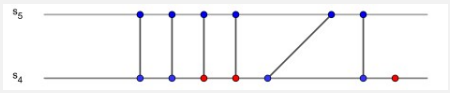
Extensions of stacks – Transitive criterion. With the result stated in proposition 1, it becomes obvious how to define the extension of a stack, and its core. Let $\mathcal{M}$ be a stack of alignments ending with sequence s, with core $K(\mathcal{M})$ and alignment graph $G_{\mathcal{M}}$. The core $\tilde{K}$ of the stack $\tilde{\mathcal{M}}$ obtained by adding the match $M = (S, T, \mu)$ on sequences s, t is obtained by extending the paths in $G_{\mathcal{M}}$ by the edges given by μ, and taking the convex hull of the extended anchor points (Figure 9). Note that this "topological" setting circumvents the problem of the apparent alignment of positions that one faces usually when adding a row to an MSA in matrix form.

A glance at Figure 9 might suggest that our method is actually prone to the "finger" effect, since the core of an extended stack can only shrink in coverage. The short answer to this remark is that, from sequence s obtained at a given round, we can extend *any* of the admissible paths that end in s (Figure 5).

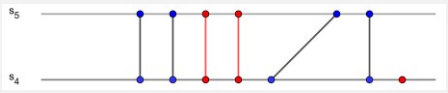
It seems that we have therefore to keep in memory all these paths, *i.e* all stacks of alignments from the reference set to all currently found sequences, which potentially leads to a combinatorial explosion, and huge memory requirements. In this section, we show how to ensure the existence of an admissible path extension, say from s to t, without having to store all admissible paths from the reference set to s. Once again, the answer is easy to state using the alignment graph formalism. Indeed, we can sum up a stack $\mathcal{M}$ ending on a sequence s by the set of positions in s that are terminal vertices of complete paths in $G_{\mathcal{M}}$. The computation of an extension becomes then essentially trivial, since it consists simply in checking which edges of the extension are incident to terminal vertices (Figure 10).

In the ungapped case, the computation of the coverage of an extended stack is also very easy. Indeed, reducing the segment on the last sequence by one also reduces the span of the whole core by one. As a consequence, we have the following criterion.

Lemma 3. Let $\mathcal{M}$ be a stack of *ungapped* pairwise alignments ending on sequence s. The core $K(\mathcal{M})$ of the stack can be summed up as the pair $\rho(\mathcal{M}) = (K_{\mathcal{M}}(s), m)$ where $m = \max_{t \in \mathfrak{S}(\mathcal{M})} |t|$. Let $M = (S, T)$ be a match from s to $t \notin \mathfrak{S}(\mathcal{M})$, and let $K_{\mathcal{M}}(s) \cap M = (S^*, T^*)$. The match M



(a) Red vertices in s_4 denote terminal vertices of complete paths ending on s_4.



(b) Added edges from s_4 to s_5 allow to define terminal vertices of the extended stack.

FIGURE 10 – Stack extension as its trace on the last sequence, shown on the example of Figure 9.

extends $\mathcal{M}$ to an admissible stack $\tilde{\mathcal{M}}$ if and only if $\ell(T^*) \geqslant 0.8 \max_{s \in \mathfrak{S}(\mathcal{M}) \cup \{t\}} |s|$. We have then

$$\rho(\tilde{\mathcal{M}}) = (T^*, \tilde{m}) \text{ where } \tilde{m} = \max(m, |t|).$$

In the gapped case, however, a minimal reduction of the coverage in the last sequence can imply a large one somewhere along the core. In principle, we should therefore keep track of the implied core coverage for all *pairs* of terminal vertices, resulting in a quadratic additional memory requirement.

Definition 6. Let $\mathcal{M}$ be a stack ending with sequence s. Let $A_{\mathcal{M}}(s)$ be the set of $\mathcal{M}$-*anchor positions*, *i.e* of endings of anchor points for $\mathcal{M}$. For any $\sigma, \tau \in A_{\mathcal{M}}(s)$, define the *restriction* $\mathcal{M}_{[\![\sigma,\tau]\!]}$ of $\mathcal{M}$ to the interval $[\![\sigma, \tau]\!]$ as the convex hull of the anchor points ending in σ and τ. We define the set of $\mathcal{M}$-*admissible extensions of* $\sigma \in A_{\mathcal{M}}(s)$ as

$$\mathcal{E}_{\mathcal{M}}(\sigma) = \left\{\tau \in A_{\mathcal{M}}(s) \mid \kappa\left(\mathcal{M}_{[\![\sigma,\tau]\!]}\right) \geqslant 0.8\right\}.$$

The set of *admissible extensions of* $\sigma \in \Sigma(s)$ is defined as $\mathcal{E}(\sigma) = \bigcup_{\mathcal{M} \mid \sigma \in A_{\mathcal{M}}(s)} \mathcal{E}_{\mathcal{M}}(\sigma)$.

In other terms, for a given sequence s, we label as admissible extensions of an anchor position $\sigma \in \Sigma(s)$ all sites τ in s such that
— there exists a stack $\mathcal{M}$ ending in s for which τ is an anchor position
— the cover of the restriction of $\mathcal{M}$ to the interval $[\![\sigma, \tau]\!]$ is at least 0.8.
We say that a match M is *admissible* if it extends at least one stack $\mathcal{M}$ as an admissible stack.

Lemma 4. Let $M = (S, T, \mu)$ be a match from sequences s to t. Let σ be the leftmost anchor position of s aligned by M, *i.e* there exists a stack $\mathcal{M}$ ending on s such that $\sigma = \min \mu(s) \cap A_{\mathcal{M}}(s)$. Let $\tau \in \Sigma(t)$ be the position in t aligned by μ, *i.e* such that $(\sigma, \tau) \in \mu$. The match M is *admissible* if

1. the segment $\mu(s)$ aligned by M contains at least one admissible extension of σ,
2. there exists $\sigma' \in \mu(s) \cap \mathcal{E}(\sigma)$ such that the position $\tau' \in \Sigma(t)$ aligned by μ satisfies $\tau' - \tau + 1 \geqslant 0.8\,|t|$.

The set of such positions is the updated set of admissible extensions $\mathcal{E}(\tau)$ of $\tau \in \Sigma(t)$.

In the gapped case, we sum up all admissible stacks ending in s as the association $(\sigma, \mathcal{E}(\sigma))$ of an anchor position for some admissible stack to the set of all admissible extensions for all admissible stacks, the previous result showing how to update these definitions when accepting a new alignment.

Implementation details. The theoretical quadratic complexity involved by the definition of an admissible extension can be reduced by further considerations. A position $\sigma = (s, p)$ is said to be *initial* when $p \leqslant 0.2\,|s| + 1$ and *final* when $p \geqslant 0.8\,|s|$. An anchor position has admissible extensions only when it is initial, and the admissible extensions are final positions. Therefore, the quadratic term can tempered by at least a 0.04 multiplicative coefficient.

On the other hand, the higher the number of stacks, the denser the sets of admissible extensions. Since a set of consecutive positions can be stored as a segment (and not as a set of distinct positions), we can expect that a simpler data structure can sometimes store the sets of admissible extensions $\mathcal{E}(\sigma)$.

A trade-off is then likely to happen between the sparseness of extensions (resulting in a smaller quadratic coefficient), and the simplification of the description of the admissible extensions.

However, in most of the cases, the balance is unfavourable for the time performance of the exact algorithm. We therefore investigate the following heuristic, which extends to the gapped case the criterion for the ungapped one : namely, we accept a match $M = (S, T, \mu)$ as an extension of a stack $\mathcal{M}$ only when it is possible to predict exactly the coverage reduction of the core $K_{\mathcal{M}}$ due to the intersection with M.

Références

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

[Corel et al., 2010] Corel, E., Pitschi, F., and Morgenstern, B. (2010). A min-cut algorithm for the consistency problem in multiple sequence alignment. *Bioinformatics*, 26(8) :1015–1021.

Chapter VII. Résumé français

Note préalable : sauf indication contraire, les références faites à des figures dans ce résumé se rapportent aux figures du corps de texte principal. Le lecteur est donc invité à se référer à la Table des figures, présente en début de manuscrit, pour retrouver facilement les figures mentionnées ci-après.

Introduction

La démarche scientifique, en particulier en ce qui concerne les sciences de la nature, peut être considérée comme l'union deux grands « archétypes » de pratiques. D'un côté, l'approche expérimentale consiste à élaborer et réaliser des expériences contrôlées, dans le but de valider (ou de réfuter) certaines hypothèses, ou bien de mesurer certaines grandeurs. Citons à titre d'exemple un essai clinique, visant à démontrer l'efficacité supposée d'un traitement contre une pathologie donnée. A l'inverse, la démarche historique a pour objectif d'inférer des évènements passés afin d'expliquer un état actuel du système étudié. Ainsi, en cosmologie, l'observation du fond diffus cosmologique et l'abondance des éléments légers ont permis d'établir le Big Bang comme origine la plus probable de l'univers. De même, en biologie de l'évolution, notre objectif général est d'inférer des évènements et relations évolutives entre la diversité des organismes contemporains, notamment en s'appuyant sur l'observation et la comparaison de caractères communs, homologues, entre différentes lignées. Avant d'explorer plus en profondeur cette notion d'homologie, notons tout de même que les deux archétypes présentés ci-dessus sont loin d'être incompatibles, et que de nombreuses disciplines (y compris la biologie de l'évolution) se basent conjointement sur des connaissances expérimentales et historiques pour étayer leurs cadres d'étude.

Si le terme « homologie », dans sa signification biologique, ne remonte qu'au XIXème siècle, la notion en elle-même est en revanche plus ancienne. Le naturaliste Pierre Belon met ainsi en évidence dès le milieu du XVIème siècle des similarités structurelles entre les squelettes d'humains et d'oiseaux, obéissant au même plan d'organisation. C'est l'anatomiste anglais Richard Owen qui, en 1843, utilise pour la première fois le terme d'homologie, décrivant l'existence de « mêmes organes dans des animaux différents » selon des similarités de position, de composition et de développement. La notion d'homologie s'oppose à celle d'analogie, qui désigne des caractères semblables (de par leur forme ou leur fonction) ne satisfaisant pas ces trois critères. Dans *L'Origine des espèces*, publié pour la première fois en 1859, Charles Darwin apporte une coloration évolutive au concept d'homologie, expliquant que les similarités entre caractères homologues découlent d'une ascendance à un ancêtre commun

chez qui le caractère est apparu. Si, à l'époque de Darwin, cette définition de l'homologie s'adressait avant tout à des comparaisons morphologiques, elle peut tout à fait être étendue pour qualifier d'autres traits héréditaires ayant une origine commune. Des gènes descendant d'un même ancêtre peuvent ainsi être dits homologues, et la reconstruction de familles d'homologie (c'est-à-dire de groupes de gènes descendant tous du même gène ancestral) est l'un des principes d'organisation majeurs de l'espace de séquences génétiques.

La dimension évolutive apportée à la notion d'homologie par Darwin a deux conséquences importantes pour l'étude des relations d'homologie entre gènes. La première est qu'elle établit l'homologie comme une relation binaire (ou plutôt, en termes de logique formelle, booléenne) : soit deux gènes sont homologues, s'ils partagent un ancêtre commun, soit ils ne le sont pas, dans le cas contraire. Cela implique en particulier que l'espace des séquences génétiques est partitionné en familles d'homologie au sein desquelles tous les gènes sont homologues entre eux, et à aucun autre gène dans d'autres familles. La seconde est que l'homologie devient une relation non plus empirique mais historique (au sens défini plus tôt), qui ne peut donc qu'être inférée sur la base d'observations des gènes contemporains[19].

L'établissement d'un lien d'homologie entre deux gènes passe le plus souvent par la comparaison de leurs séquences, en construisant un alignement : de la même manière que l'on peut faire correspondre les os du bras d'un humain et ceux de l'aile d'un oiseau pour illustrer leur similarité morphologique, on fait correspondre les positions de chaque séquence avec celles de l'autre pour mettre en lumière leur similarité (Figure 4). Une similarité « excessive » entre séquences – sous-entendu, par rapport à la similarité que l'on s'attend à trouver entre deux séquences choisies au hasard – sert alors de base pour inférer une relation d'homologie entre les gènes considérés. Cet excès de similarité se mesure par différentes métriques, notamment la E-value associée à l'alignement, qui quantifie le nombre d'alignements d'une qualité égale ou supérieure qui apparaîtraient entre deux séquences aléatoires de cette taille, ainsi que le pourcentage d'identité entre les positions alignées et la couverture de l'alignement sur chacune des séquences (Figure 6). Dans le cadre standard de l'homologie, qui repose sur des comparaisons directes entre séquences génétiques, on ne considère donc que les paires de séquences qui s'alignent entre elles, avec une similarité prononcée, et ce

[19] En réalité, pas exclusivement : le séquençage de fragments d'ADN ancien (aDNA), préservés par des processus de congélation ou de momification naturelle ou artificielle, permet également d'obtenir des informations évolutives. Cependant, la dégradation progressive de l'ADN complique fortement l'exploitation d'échantillons au-delà de quelques millions d'années, faisant de l'aDNA un outil principalement adapté à l'étude d'une histoire évolutive relativement récente.

(puisque les gènes sont soit homologues, soit non) le long d'une région recouvrant la majeure partie de leur longueur.

Si ce cadre opérationnel permet de révéler et de qualifier un très grand nombre de relations évolutives, il présente toutefois quelques angles morts par rapport à l'ensemble des liens d'ascendance pouvant exister au sein de l'espace génétique dans son ensemble. Au cours de ma thèse, je me suis intéressé spécifiquement à deux de ces liens d'ascendance qui échappent au modèle canonique de l'homologie, et j'ai développé de nouvelles méthodes d'analyse basées sur les réseaux de similarité de séquences pour étudier ces relations (Figure 16).

Le premier de ces angles morts concerne l'existence de liens d'homologie plus distante que celle usuellement décrite par les alignements de séquences. En effet, deux gènes peuvent tout à fait être homologues sans pour autant présenter suffisamment de similarité pour qu'un alignement soit construit entre eux. Cela s'explique, au moins partiellement, par des limitations techniques que rencontrent les algorithmes d'alignement lorsqu'il s'agit de comparer des séquences présentant un faible niveau de similarité (typiquement, dans le cas de BLAST, sous la barre des 30% d'identité entre séquences protéiques). En raison de cet effet, appelé « twilight zone » de l'alignement de séquences, les liens d'homologie qui résultent en des similarités faibles ne sont pas détectés, et sont donc rarement considérés. Cette homologie distante peut notamment se produire lorsqu'une lignée spécifique d'une famille de gènes diverge d'une manière accélérée, en accumulant un nombre accru de mutations qui érode progressivement sa similarité avec d'autres lignées jusqu'à passer sous le seuil de détectabilité des aligneurs de séquences.

Le second problème du cadre standard de l'homologie réside, plus fondamentalement, dans la manière dont on définit celle-ci. En effet, nous avons jusque là parlé exclusivement de familles de gènes discrètes, déconnectées les unes des autres ; cependant, des mécanismes combinatoires entrent aussi en jeu dans l'évolution des gènes. Dans le cas d'une fusion de gènes, par exemple, l'union d'un gène A et d'un gène B pour former le gène AB produit des motifs de similarité incompatibles avec cette vision discrète des familles de gènes : AB est similaire à A sur une partie de sa séquence seulement, et similaire à B sur une autre. Il n'apparaît donc pas tout à fait correct de dire qu'AB est homologue à A et/ou à B, mais il n'est pas non plus satisfaisant de dire qu'ils ne sont pas homologues du tout. De tels processus d'évolution des gènes sont donc incompatibles avec la définition de l'homologie comme une relation strictement binaire, et font apparaître en filigrane l'idée que tous les mécanismes de l'évolution n'adhèrent pas nécessairement au modèle de l'arborescence pour les illustrer.

Pour étudier les relations entre gènes dans leur ensemble, d'autres modèles que ceux basés sur les arbres phylogénétiques peuvent donc parfois être préférables (ou au moins complémentaires). Pendant ma thèse en particulier, les méthodes que j'ai développées et appliquées s'appuient sur les réseaux de similarité de séquences (SSN en anglais). Dans un tel réseau, chaque séquence est représentée par un nœud, et deux séquences sont connectées par une arête lorsqu'elles présentent une similarité excédant un seuil prédéfini. L'avantage de ces réseaux est que certaines relations évolutives particulières produisent dans le réseau des patrons d'interconnexion distinctifs, qui peuvent alors être détectés et analysés computationnellement. En raison de cette propriété, les SSN ont été utilisés pour étudier une grande variété de facettes de l'évolution, notamment certaines ne s'inscrivant pas dans le cadre opérationnel classique de l'homologie : évolution des protéines multi-domaines, remodelage de gènes suite à une endosymbiose, réseaux de partage de gènes entre différents biomes… Dans la lignée de ces travaux, ma thèse propose donc d'étudier l'homologie distante et le remodelage de gènes en développant de nouvelles analyses basées sur ces représentations en réseaux.

Homologues distants de familles de protéines très conservées

Les limitations pratiques rencontrées par les algorithmes d'alignement pour établir des similarités entre séquences homologues mais peu semblables font de la détection d'homologues distants un problème récurrent en biologie. Ces liens d'homologie distante sont pourtant particulièrement intéressants, car ils peuvent révéler des variants génétiques divergeant fortement de la diversité connue au sein d'une famille de gènes. Ces variants peuvent par ailleurs être de différentes natures : certains peuvent représenter des lignées génétiques récentes ayant accumulé rapidement un grand nombre de mutations, par exemple dans le cas de la néo-fonctionnalisation de gènes dupliqués ; d'autres, à l'inverse, peuvent correspondre à des lignées génétiques plus basales, dont la faible similarité de séquence s'explique par une divergence ancestrale par rapport aux lignées connues dans cette famille de gènes. La détection et l'analyse de liens d'homologie distante est donc capitale pour améliorer notre compréhension des manières dont les familles de gènes évoluent, et nous permettre de saisir l'étendue réelle de leur diversité dans le monde vivant.

Au cours de ma thèse, j'ai participé au développement d'une méthode de détection d'homologues distants, nommée SHIFT. Le principe de base de cette méthode consiste à effectuer des recherches itératives d'homologie dans une large base de séquences cible, dans le but d'accumuler des variants de plus en plus divergents autour de séquences de référence (Figure 20A). L'ambition est

alors de pouvoir retrouver, de proche en proche, des homologues trop distants des références pour être atteints par une recherche directe. A partir des séquences choisies comme références pour la famille de gènes étudiée, un premier tour de BLAST permet d'identifier leurs homologues « directs » dans la base de données cible. Ces nouvelles séquences peuvent alors être utilisées comme queries pour une seconde recherche BLAST, afin d'identifier leurs propres homologues dans la base de données, qui sont donc des homologues « de second degré » des références. En itérant ce principe, utilisant à chaque tour pour queries les séquences retrouvées à l'itération précédente, on peut alors agréger des variants de plus en plus distants des séquences références (Figure 20B). Cependant, en appliquant cette procédure de manière « naïve » sans contrôler la nature des séquences retrouvées à chaque itération, il est possible d'aboutir à une surextension du champ de recherche, en rapatriant des séquences n'ayant en réalité pas d'origine commune avec les références initiales (Figure 13). Il est donc vital d'implémenter un contrôle pour s'assurer, entre deux tours successifs de BLAST, que les séquences nouvellement retrouvées semblent bien correspondre à des homologues distants des références. Plus précisément, dans SHIFT, les séquences nouvellement identifiées à chaque tour d'alignement ne sont retenues que si leurs régions alignées peuvent être rétro-propagées jusqu'à une séquence référence, de manière à ce que la région commune recouvre au moins 80% de toutes les séquences le long de la chaîne d'alignement (Figure 20C). Avec ce critère en place, les recherches itératives ont beaucoup moins de chances de retenir des séquences n'ayant pas de rapport évolutif avec les références. Les itérations de SHIFT se poursuivent jusqu'à ce qu'aucune nouvelle séquence ne soit trouvée, après quoi un réseau de similarité de séquences « étendu », regroupant à la fois les séquences de référence et leurs homologues plus ou moins proches, est produit par un dernier alignement BLAST tout-contre-tout.

L'aspect méthodologique de SHIFT fait l'objet d'un article actuellement en cours de rédaction. Pendant ma thèse, j'ai également appliqué ce protocole afin de conduire, dans un contexte de génomique environnementale, une recherche d'homologues profonds de familles de gènes très conservées, en particulier dans un métagénome océanique riche en organismes non cultivés. Avant de présenter les résultats de ces travaux, je propose de restituer brièvement le contexte de la génomique environnementale et de l'exploration de la matière noire microbienne.

Depuis maintenant quelques décennies, il est communément admis que la grande majorité des microorganismes présents sur Terre sont incompatibles avec les approches actuelles de cultivation en laboratoire. Ce fait est particulièrement marqué dans les environnements « naturels », c'est-à-dire non associés aux microbiomes humains ou autres milieux anthropiques, où l'écrasante majorité des organismes appartiennent à des genres, des ordres ou même des phylums sans souche représentative

cultivée (Figure 15). Par conséquent, la plupart des métagénomes environnementaux sont dominés par des séquences sans origine phylogénétique et/ou fonction biologique connues, et cette fraction du monde microbien est parfois appelée « matière noire microbienne ».

Cette importante diversité environnementale a fait l'objet de nombreuses études au cours des quinze dernières années, lesquelles ont notamment révélé de nouvelles lignées majeures dans l'arbre du vivant. On peut ainsi citer les bactéries CPR, un large superphylum ayant la particularité de présenter des cellules et génomes bien plus petits que la norme des autres bactéries ; les DPANN, similaires aux CPR de par la taille de leurs génomes et cellules du côté des archées ; ou encore les Asgard, un groupe diversifié d'archées apparaissant comme les plus proches parents des eucaryotes connus à ce jour. Au-delà même du monde cellulaire, les investigations de génomique environnementale ont également révélé une diversité insoupçonnée de la virosphère, en particulier chez les virus à ARN. Cependant, au cours des dernières années, le rythme des nouvelles découvertes majeures dans l'arbre du vivant semble avoir fortement ralenti. Ce reflux mène certains biologistes à conjecturer que l'ensemble des grands groupes phylogénétiques existant sur Terre ont désormais été découverts, de sorte que la diversité restante au sein de la matière noire microbienne représenterait surtout de nouvelles lignées moins basales, ainsi que des groupes inconnus sur le plan fonctionnel.

Dans ce contexte, la recherche d'homologie distante peut être pertinente afin d'explorer plus en profondeur les fractions inconnues de métagénomes environnementaux. En particulier, un nombre réduit de familles de gènes s'avère fortement conservées sur le plan évolutif, au sens où elles sont présentes dans tous les Domaines du vivant, rarement perdues ou transférées horizontalement entre génomes, et présentant des divergences de séquence relativement faibles au vu de leur âge considérable. Ces familles « core » peuvent donc être considérées, d'une certaine manière, comme essentielles à la vie cellulaire. Trouver des groupes de variants homologues distants dans ces familles est donc d'un intérêt biologique notable, car ces variants peuvent alors indiquer des lignées ou des fonctions divergentes dans des processus considérés comme clés pour les organismes vivants. C'est dans ce but précis que j'ai conduit, en utilisant SHIFT, une recherche d'homologues distants pour un certain nombre de ces familles « core », spécifiquement au sein du métagénome océanique OM-RGC assemblé par l'expédition Tara Océans. Ces travaux ont fait l'objet d'un autre article de recherche, actuellement en cours de révision pour le journal *Environmental Microbiome*.

J'ai constitué un ensemble de 53 familles de gènes fortement conservées, et j'ai utilisé SHIFT pour chacune d'elles afin de trouver leurs homologues (proches et distants) dans le métagénome OM-RGC. La convergence de SHIFT a été atteinte après en moyenne sept itérations, multipliant par plus de six la quantité de séquences dans cet ensemble de familles de gènes par rapport aux seules

séquences de référence. Les homologues environnementaux détectés par SHIFT ont ensuite été alignés par BLAST contre la base données *nr* du NCBI, permettant ainsi de quantifier le pourcentage d'identité entre chacune de ces séquences et son plus proche homologue dans l'ensemble de la diversité génétique cultivée. Nous avons trouvé que seules 6.7% des séquences rapportées par SHIFT étaient 90% similaires ou plus à leur plus proche parent connu, tandis que 20.5% des homologues environnementaux divergent plus de l'ensemble de la diversité connue que la divergence observée en moyenne entre séquences bactériennes et archées dans nos familles initiales.

L'objectif étant d'identifier des variants d'intérêt dans la diversité environnementale de ces familles très conservées, il est naturellement plus intéressant et plus significatif de trouver des groupes cohérents de séquences très divergentes, plutôt que des séquences individuelles isolées du reste. J'ai donc effectué un partitionnement du réseau de similarité de séquences de chacune des familles en plusieurs communautés (ou clusters) de séquences fortement connectées entre elles, en appliquant pour cela l'algorithme de Louvain. Cela nous a permis d'identifier des clusters riches en séquences environnementales très divergentes de la diversité connue, qui peuvent donc représenter des lignées divergentes dans leurs familles de gènes respectives. Je me suis alors intéressé plus en profondeur à ces clusters divergents dans trois familles spécifiques, dont les résultats sont restitués dans le preprint de l'article figurant dans le Chapitre II de ma thèse. Dans ce résumé, je présenterai mes résultats pour deux de ces trois familles, représentant des variants génétiques de différente nature.

Le premier exemple concerne un variant génétique appartenant à une lignée bien établie, mais qui présente toutefois des particularités intéressantes du point de vue fonctionnel et structural. Il appartient à la famille de protéines appelée SMC, qui assure les dynamiques de repliement et de dépliement des chromosomes au cours des différentes étapes du cycle cellulaire. Une protéine SMC est constituée de deux longs brins hélicoïdaux enroulés ensemble, présentant à une extrémité un domaine globulaire ATPase, et à l'autre un domaine en demi-cercle appelé *hinge* (voir Figure 5 du preprint dans le Chapitre II). Deux protéines SMC s'associent ensemble par le biais de leur domaine *hinge*, et recrutent à leurs extrémités ATPase des protéines accessoires pour former un complexe SMC en forme d'anneau, capable d'encercler une molécule d'ADN double-brin pour réguler son organisation spatiale. Plus spécifiquement (c'est important pour la suite), cet attachement autour du chromosome s'effectue par l'ouverture transiente de l'interface entre les domaines *hinge* des deux protéines SMC.

Dans cette famille SMC, SHIFT a permis l'identification d'un cluster de séquences divergentes de la diversité connue. Ce cluster a la particularité d'être très abondant dans le métagénome OM-RGC,

près de sept fois plus que le reste des séquences SMC retrouvées par SHIFT. Au sein de la phylogénie des protéines SMC, ce variant océanique s'inscrit au sein des séquences d'Actinobactéries (voir Figure 4 du preprint dans le Chapitre II). Surtout, les séquences de ce variant présentent la particularité de ne pas posséder de domaine *hinge*, et les structures protéiques que nous avons inférées confirment cette observation. En d'autres termes, nous avons identifié dans l'océan un variant « hinge-less » des protéines SMC, ayant perdu le domaine *a priori* indispensable pour l'attachement du complexe SMC à l'ADN préalablement à toute régulation de son organisation. Il semble peu probable que ces variants SMC « hinge-less » assurent la même fonction que leurs homologues usuels en l'absence de leur interface avec l'ADN. Pour autant, la forte abondance de ces séquences dans l'environnement pousse à croire qu'elles réalisent bien une certaine fonction dans leurs hôtes, qu'il sera important d'élucider pour mieux comprendre la signification biologique de ce nouveau variant.

A l'inverse, le second exemple de famille conservée présentant des variants environnementaux divergents concerne plutôt l'identification de potentielles nouvelles lignées basales dans l'évolution de cette famille. Il s'agit cette fois-ci de la protéine recombinase A, qui assure plusieurs fonctions dans la réparation de dommages subis par l'ADN (notamment des cassures double-brin) et permet la tenue de recombinaisons homologues chez les procaryotes. Dans cette famille, nous avons identifié quatre groupes d'homologues environnementaux divergents (voir Figure 6 du preprint dans le Chapitre II). En particulier, deux de ces clusters étaient fortement enrichis en séquences venant de fractions de taille de l'ordre du nanomètre, typique des virus ou encore des bactéries CPR et des archées DPANN. Dans la phylogénie des séquences de cette famille de gènes, l'un de ces clusters semblait correspondre à des séquences d'origine bactérienne, tandis que l'autre formait un clade entre les bactéries et les archées, un placement qui pourrait être compatible avec des lignées microbiennes très divergentes dans l'arbre du vivant. En particulier, malgré leur présence dans des fractions de taille « ultra-petites », ces variants océaniques divergents ne correspondaient pas aux séquences de recombinase A connues pour les CPR ou les DPANN. Ils pourraient par conséquence appartenir à des bactériophages, qui encodent parfois de tels gènes essentiels pour leurs hôtes, ou encore à de nouvelles lignées cellulaires inconnues, ayant potentiellement des diamètres cellulaires particulièrement faibles.

Homologie partielle et remodelage de gènes dans deux lignées multicellulaires

Dans le cadre conceptuel standard de l'homologie, les gènes sont principalement considérés comme des unités atomiques (indivisibles) d'évolution, divergeant selon une variété de processus qui peuvent être représentés par des arbres phylogénétiques (Figure 7). Cependant, les gènes évoluent

également par des processus combinatoires, qui mettent en jeu des réarrangements de *parties* de gènes, tels que les fusions et fissions génétiques. Ces mécanismes font alors apparaître des motifs de similarité partielle entre des gènes qui ne partagent qu'une portion de leur séquence. Loin d'être un phénomène de marge, le remodelage de gènes est amplement reconnu et documenté, principalement dans le cadre opérationnel des domaines protéiques, voulus comme des sous-unités conservées de séquence, de structure et de fonction. La majorité des protéines, que ce soit chez les procaryotes ou les eucaryotes, comportent par ailleurs plusieurs domaines, soulignant l'importance de ces processus combinatoires dans l'évolution des gènes (Figure 23).

En raison de la versatilité de la définition des domaines protéiques, la plupart des investigations concernant le remodelage de gènes ont été conduites par le prisme des assemblages de domaines. Pour autant, les domaines ne décrivent pas l'intégralité de l'espace des séquences génétiques : environ 20% des protéines connues ne contiennent pas de domaine répertorié, et les domaines ne couvrent qu'un peu plus de 50% des résidus dans l'ensemble du protéome connu (Figure 24). Par conséquent, étudier l'évolution combinatoire par ce prisme uniquement ne peut offrir qu'une vision partielle de l'étendue de ces processus, qui peut être complétée en définissant des « briques de base » des réarrangements génétiques par d'autres moyens. L'une de ces approches, mise en place dans mon laboratoire avant ma thèse et intitulée CompositeSearch, permet ainsi de détecter des événements de remodelage génétique sur la base seule des similarités partielles entre séquences. Cette méthode s'affranchit donc de la définition des domaines protéiques, mais a l'inconvénient que les fusions de gènes comme les fissions peuvent résulter en un même patron de similarités partielles, ne permettant donc pas de les distinguer *a priori* (Figure 25). On parle donc plutôt de gènes composites et composants au sein de ces patrons, afin de ne pas induire de biais terminologique envers l'une ou l'autre de la fusion ou de la fission de gènes. En outre, une analyse complémentaire doit être entreprise afin de « polariser » ces événements de remodelage détectés, c'est-à-dire restaurer l'information de fusion ou fission associée à chaque événement.

Pendant ma thèse, j'ai développé une telle méthode de polarisation comme analyse subséquente à CompositeSearch. Cette méthode s'appuie sur le signal phylogénétique afin de déterminer, entre une famille composite et ses familles composantes associées, laquelle des formes (associée, dans les gènes composites ; dissociée, dans les gènes composants) préexistait par rapport à l'autre. Les données de présence/absence des familles composite et composantes dans chaque génome de la lignée étudiée sont utilisées pour inférer, par parcimonie Dollo, leurs points d'origine dans la phylogénie de la lignée (Figure 26A). Ces origines sont ensuite comparées entre elles pour déterminer la polarisation. Des familles composantes ayant émergé avant l'apparition du gène

composite indiquent ainsi une fusion de gènes ayant donné lieu à ce composite (Figure 26B) ; à l'inverse, une forme composite pré-datant les formes composantes suggère davantage un événement de fission (Figure 26C). De nombreux cas intermédiaires peuvent émerger en dehors de ces cas de figure « idéaux », par exemple si l'origine de la forme composite pré-date une composante mais est ultérieure à une autre. Ces cas de figure correspondent à des scénarios plus complexes qu'une simple fusion ou fission, par exemple une fusion de gènes suivie d'une perte de l'une des composantes expliquant son émergence inférée comme plus tardive.

J'ai appliqué durant ma thèse cette méthode dans le cadre de deux études distinctes. La première de ces études est le fruit d'un large consortium réuni autour d'un projet de séquençage et d'analyse de 60 nouveaux génomes d'algues brunes, dont les résultats ont été récemment publiés dans la revue *Cell*. Au sein de ce projet, j'ai réalisé une analyse consistant à détecter et polariser les gènes remodelés dans ces génomes, ainsi qu'à étudier leurs fonctions et leur stabilité au cours de l'évolution de la lignée des algues brunes. La seconde étude s'est concentrée plus spécifiquement sur les événements de remodelage dans les génomes animaux, à partir de 63 génomes déjà publiés. Dans cette recherche, j'ai appliqué mon analyse de polarisation à des composites déjà existants, et j'ai pu à nouveau étudier la stabilité de ces gènes remodelés dans les génomes animaux. Les résultats de cette recherche sont à l'heure actuelle disponibles à l'état de preprint sur bioRxiv.

Un point commun entre ces analyses est le fait qu'elles se concentrent toutes deux sur des lignées ayant acquis indépendamment un phénotype multicellulaire complexe. Plus généralement, la multicellularité se retrouve dans cinq grands groupes d'eucaryotes : en plus des animaux et des algues brunes, on peut citer les plantes et algues vertes, les algues rouges, ainsi que les champignons. L'émergence de ce phénotype requiert un grand nombre d'adaptations physiologiques, notamment des systèmes de transport d'oxygène et de nutriments, ou encore un programme de développement pour passer de la cellule-œuf au stade adulte. L'apparition répétée de cette multicellularité est donc loin d'être anodine, et il est particulièrement intéressant de comprendre les « recettes » génomiques permettant la transition de l'unicellularité à la multicellularité. Dans le spécifique cadre du remodelage de gènes, des recherches ont déjà établi une association entre complexité des architectures de domaines et complexité phénotypique des organismes. Cette association est d'ailleurs largement documentée chez les animaux, qui présentent des combinaisons de domaines protéiques particulièrement dynamiques.

Dans les algues brunes, j'ai pu identifier une forte contribution des phénomènes de remodelage à l'évolution des génomes. En effet, 6.7% de l'ensemble des familles de gènes dans cette lignée ont été produits par un événement de fusion, et près de 5% par une fission de gènes. La plupart de ces

gènes remodelés ont émergé tôt dans l'évolution des algues brunes, notamment dans les branches menant à l'ancêtre de cette lignée ainsi que dans ses premières diversifications (Figure 27). J'ai ensuite évalué le taux de rétention de ces gènes remodelés dans les génomes contemporains, c'est-à-dire la proportion de gènes effectivement présents dans le génome d'une espèce parmi tous ceux qui ont émergé dans la lignée menant à cette espèce. J'ai alors pu observer que chez les algues brunes et leurs plus proches parents, les gènes issus de fusions et de fissions étaient préférentiellement conservés par rapport aux gènes non-remodelés, ce qui n'était pas le cas dans d'autres espèces de Straménopiles plus éloignées (Figure 27). Cela suggère que les produits du remodelage génétique peuvent occuper des fonctions importantes pour les algues brunes, qui ont émergé tôt dans l'histoire de la lignée et ont été conservées ensuite au cours de la diversification de cette lignée. Par ailleurs, les remodelages de gènes ont été bien plus fréquents pour certaines catégories fonctionnelles spécifiques, qui ont la particularité d'être fréquemment associées à la multicellularité (Figure 28). Ainsi, des contributions de fusions et fissions de gènes aux processus liés au métabolisme des glucides et à la synthèse de la paroi cellulaire ont pu participer au développement de la paroi et de la matrice extra-cellulaire des algues brunes, basées sur les alginates et qui assurent la cohésion intercellulaire dans ces algues. De même, le remodelage de gènes a particulièrement affecté les catégories fonctionnelles de la transcription et de la transduction du signal, qui chez les multicellulaires sont liés à une complexification des voies de signalisation et de communication intercellulaire.

Chez les animaux, nous avons observé une dynamique différente dans les processus d'évolution combinatoire des familles de gènes. Environ 5% des familles de gènes étaient composites, dont trois quarts de fusions de gènes et un quart de fissions. Plutôt qu'une contribution progressive et continue, les événements de remodelage se concentrent à certains nœuds spécifiques de la phylogénie animale, notamment à l'émergence des bilatères et des Euteleostomi (Figure 29). Le remodelage génétique chez les animaux est également caractérisé par une forte dynamicité et réversibilité : dans la majorité des cas de fusion, le gène fusionné est ultérieurement fissionné à nouveau dans au moins une des espèces hôtes. En analysant les taux de rétention des gènes remodelés dans les génomes contemporains, comme expliqué précédemment pour les algues brunes, j'ai par ailleurs pu observer que chez les vertébrés spécifiquement, et de manière d'autant plus prononcée chez les Euteleostomi, les gènes issus de fusions sont largement plus conservés que les gènes non-remodelés, tandis qu'à l'inverse les gènes issus de fissions sont significativement plus perdus (Figure 30). Ce motif suggère ainsi un biais significatif envers les fusions de gènes dans les génomes animaux, qui malgré leur réversibilité restent bien plus conservés que les produits de fissions. En outre, ces fusions participent significativement à certaines catégories de fonctions qui peuvent être associées à la grande diversité des phénotypes animaux. Des contributions substantielles aux fonctions de transcription, de

transduction du signal et de modifications post-traductionnelles pourraient ainsi avoir favorisé l'émergence de voies de régulation complexes, qui chez les animaux sont particulièrement importantes au développement des organismes. De même, une participation marquée des fusions de gènes dans les fonctions liées aux structures extra-cellulaires pourrait être en lien avec la grande diversité de tissus et d'organes présente à travers la diversité du « règne » animal.

Discussion & Perspectives

Au cours de ma thèse, mon objectif a été de remédier à deux écueils principaux du cadre opérationnel standard de l'homologie entre gènes, dans le but de prendre en compte un spectre plus large de relations d'ascendance évolutive. En particulier, je me suis intéressé d'une part à la notion d'homologie distante, qui passe en quelque sorte sous le radar des analyses basées sur des alignements standards entre séquences, et d'autre part à des processus d'évolution combinatoire, incompatibles avec une définition binaire, « tout ou rien » de l'homologie. J'ai pour cela développé et appliqué différentes méthodes bio-informatiques basées sur la construction et l'analyse de réseaux de similarité de séquences.

Ma recherche de variants océaniques distants dans des familles de gènes « core » très conservées a permis de révéler une grande diversité de variants génétiques divergents dans l'océan global. En particulier, j'ai pu identifier des groupes divergents de différente nature, avec notamment d'une part des variants structuraux et fonctionnels dans des lignées phylogénétiques bien établies, ainsi que d'autre part des groupes plus profonds, potentiellement compatibles avec de nouvelles lignées basales dans l'arbre du vivant. Les résultats produits dans le cadre de cette étude sur trois familles spécifiques ne sont en réalité qu'une fraction de la diversité que nous avons réellement trouvés, car 25 des 53 familles considérées au total comportaient au moins un cluster d'homologues très divergents. Caractériser un plus grand nombre de ces variants permettrait donc de révéler dans de plus amples détails la diversité génétique présente en dehors des limites de la cultivation microbienne en laboratoire. Cette approche et ces résultats pourraient potentiellement guider la formulation de nouvelles hypothèses en génomique environnementale, par exemple en dressant une sorte de liste des lignées inconnues les « plus recherchées », qui pourraient porter plusieurs de ces variants environnementaux. En outre, tirer profit des avancées récentes en prédiction et comparaison de structures protéiques permettrait d'améliorer les capacités d'identification d'homologues distants, les structures de protéines étant généralement bien plus conservées évolutivement que les séquences primaires.

Dans un second axe de recherche, j'ai pu mettre en évidence au cours de ma thèse une contribution significative des processus de remodelage génétique dans l'évolution des algues brunes et des animaux, en particulier dans des catégories fonctionnelles généralement associées à l'évolution de phénotypes multicellulaires complexes. Cette contribution des processus de remodelage s'est pour autant faite suivant des tendances distinctes. Du point de vue mécanistique d'abord, puisque les algues brunes ont tiré profit d'un équilibre relatif entre fusions et fissions de gènes, tandis qu'un biais marqué en faveur des fusions a pu être observé chez les animaux. Du point de vue chronologique également, les algues brunes ayant surtout acquis leurs gènes remodelés tôt au cours de l'évolution de leur lignée, alors qu'à l'inverse le remodelage de gènes chez les animaux semble plutôt avoir eu lieu par pics, à certains points spécifiques de la lignée métazoaire. Au vu de ces différences, il serait d'autant plus intéressant d'étendre ces études aux autres lignées ayant acquis une forme de multicellularité complexe : cette multicellularité ayant émergé indépendamment au moins 16 fois au cours de l'évolution des eucaryotes, étudier la contribution du remodelage de gènes à chacune d'entre elles permettrait d'éclairer une compréhension d'ensemble des « recettes » génomiques pouvant être mises en œuvre pour acquérir ce phénotype complexe.

L'approche que j'ai adoptée durant ma thèse, qui consiste à étudier l'évolution des familles de gènes par le spectre principal des réseaux de similarité, vise à apporter une vision plus holistique des liens évolutifs pouvant exister entre les gènes. Cette démarche vise non pas à supplanter, mais à étendre et complémenter les approches basées sur la vision arborescente des phénomènes évolutifs, précisément dans le but de capturer les relations évolutives qui ne peuvent être décrites dans ces approches. En cela, ma thèse souligne les bénéfices d'un pluralisme de concepts et de méthodes en biologie de l'évolution, afin de comprendre la diversité des gènes, des organismes et des mécanismes évolutifs dans leur globalité.

Résumé : L'augmentation toujours plus importante de données génomiques et métagénomiques appelle de nouveaux développements méthodologiques et bio-informatiques, afin de caractériser avec davantage de précision les phénomènes évolutifs dans leur ensemble. En particulier, certaines des méthodes usuelles pour étudier l'évolution des (familles de) gènes s'avèrent inadaptées lorsque la parenté entre séquences n'est que partiellement supportée. Ainsi, la définition et la reconstruction de familles de gènes se heurtent à l'obstacle de l'homologie distante, qui passe sous le seuil de détection des alignements de séquences. De même, les mécanismes d'évolution combinatoire, tels que les fusions et fissions de gènes, remettent en cause les représentations purement arborescentes de l'évolution des familles de gènes. L'application de méthodes complémentaires basées sur les réseaux de similarité de séquences permet de contourner certaines de ces lacunes, en proposant une représentation holistique des similarités entre gènes. La détection et l'analyse d'homologues très divergents de familles de gènes fortement conservées dans des jeux de données environnementaux est notamment facilitée par la recherche itérative d'homologie fondée sur les réseaux. Cette fouille itérative de métagénomes révèle une immense diversité de variants environnementaux dans ces familles, qui divergent de la diversité connue tant par leur séquence que par la structure des protéines qu'ils encodent, et elle permet de suggérer des pistes pour guider de futures explorations de la matière noire microbienne. En outre, en prenant en compte des liens d'homologie partielle entre séquences génétiques, les réseaux de similarité de séquences permettent une identification systématique des évènements de fusion et de fission de gènes. Il devient ainsi possible d'évaluer l'impact de ces processus au cours de l'évolution de lignées biologiques d'intérêt, permettant de comparer le rôle qu'ils ont joué lors de l'émergence de phénotypes multicellulaires complexes dans plusieurs telles lignées. Plus généralement, ces approches basées sur les réseaux illustrent l'intérêt de prendre en compte une pluralité de modèles pour étudier une plus grande variété de processus évolutifs.

Abstract: The ever-increasing accumulation of genomic and metagenomic data calls for new methodological developments in bioinformatics, in order to characterise evolutionary phenomena as a whole with better accuracy. In particular, some of the canonical methods to study the evolution of genes and gene families may be ill-suited when the relatedness of sequences is only partially supported. For instance, the definition and reconstruction of gene families face the hurdle of remote homology, which falls beneath the detection thresholds of sequence alignments. Likewise, combinatorial mechanisms of evolution, such as gene fusion and gene fission, challenge the purely tree-based representations of gene family evolution. The use of complementary methods based on sequence similarity networks allows us to circumvent some of these shortcomings, by offering a more holistic representation of similarities between genes. The detection and analysis of highly divergent homologues of strongly conserved families in environmental sequence datasets, in particular, is facilitated by iterative homology search protocols based on networks. This iterative mining of metagenomes reveals an immense diversity of environmental variants in these families, diverging from the known diversity in primary sequence as well as in the tertiary structure of the proteins they encode. It is thus able to suggest possible directions of future explorations into microbial dark matter. Furthermore, by factoring in relationships of partial homology between gene sequences, sequence similarity networks allow for a systematic identification of gene fusion and fission events. It thus becomes possible to assess the effects of these processes on the evolution of biological lineages of interest, enabling us for instance to compare the role that they played in the emergence of complex multicellular phenotypes between several such lineages. More generally, these network-based approaches illustrate the benefits of taking a plurality of models into account, in order to study a broader range of evolutionary processes.